On gradual regime switching models: A generalisation of Hamilton's method of time-series analysis.

Bodger, Owen Galdan

On Gradual Regime Switching Models: A Generalisation of Hamilton's Method of Time-Series Analysis

by

Owen Galdan Bodger,
BSc (Hons) University of Wales Swansea

Thesis

submitted to the University of Wales
in candidature for the degree of

PHILOSOPHIÆ DOCTOR

European Business Management School
University of Wales, Swansea
Swansea SA2 8PP
United Kingdom

December 2005

Acknowledgements

That I have made it this far is testament to the support I have received during my research.

Without question the most important contribution has been made by my supervisor, Assad Jalali. In addition to the plentiful support you provided, your patience and tact were greatly appreciated.

The only other person I shall name individually is Jacky. Your selflessness and tolerance have been beyond all expectation. I only hope that I am as understanding when our roles are reversed.

Some credit must also go to my colleagues for providing the kind of dysfunctional social environment in which I could thrive. I may even have enjoyed their company, but don't tell them that.

I am also grateful to the EPSRC for investing in me. Well, gambling actually.

OWEN BODGER

University of Wales
December 2005

Summary

The class of Markov Switching time series models, introduced by Professor James Hamilton, is nearly twenty years old. Despite this, relatively little work has been done on allowing gradual transitions between the regimes of the model. Almost all of the published work relates to modelling a transition between two regression lines rather than incorporating it into a time series model. We decided to approach the problem from two directions.

First, we wanted to look at Filtered Telegraph signals (Filtered Markov processes) and consider their suitability for time series analysis. Secondly, we wished to extend the existing Regime Switching models to allow a gradual transition between regimes.

In our work on the Filtered Markov process we present a method for obtaining moments for a signal with any number of regimes, rather than the usual two. This enables us to find the stationary, transient and conditional moments of the signal. We include an expression for the covariance of two observations from a signal, obtained using the conditional moments.

While considering how to fit a Filtered Markov process we identify several new methods that can be used for estimating the parameters of a sample from the Beta distribution where the observations have been contaminated by noise. We also include extensive tables of the percentiles of the estimators for each of the methods.

We also present a new algorithm that utilises the Filtered Markov process to generate random Beta variates.

Finally we take a more practical approach, introducing some simple models that, while useful in their own right, could also be used to bridge the gap between the two-regime Markov switching model and the Filtered Markov process. These Ladder models are then applied to several data sets to explore the problems faced by gradual switching models and collect evidence of their suitability.

To Enid Webb

Perhaps we aren't such a rotten lot after all.

Contents

List of Figures

List of Tables

Notation

General

- θ Sum of α and β (or a and b where $\tau = 1$).
- I An Interval.
- dt Short time interval.
- π A stationary distribution.
- $\mathbf{\Theta}$ The set of structural parameters of a model.
- $\mathbf{\Pi}$ The initial conditions.
- $\mathbf{\Sigma}$ The combination of parameters and missing data $[\mathbf{\Theta}, \mathbf{s}_N]$.
- i, j The Regime index (usually i before, j after).
- u, v The Level index (usually u before, v after).

Filtered Markov Process

- α, β The Switching Intensities when in regimes 0 and 1.
- c, d The Scale and Shift parameters.
- l_i Mean level of the signal associated with regime i.
- τ The Time-Scaling factor
- ρ equal to $e^{-\frac{1}{\tau}}$.
- σ Standard deviation of Gaussian noise.
- $S(t)$ The Markov process that determines the regime of the process.
- $X(t)$ The values taken by a noise-free process.
- $Y(t)$ The values taken by a process contaminated by noise (usually Gaussian).
- $y(t)$ Observations from a process contaminated by noise (usually Gaussian).
- $\epsilon(t)$ A Gaussian error term.
- t Index of time.
- T The maximum time in the series.

Discrete Models

- a, b The Switching probabilities for regimes 0 and 1.
- l_i Mean level of the signal associated with regime i.
- p, q The Probabilities of movement when in regimes 0 and 1.
- σ Standard deviation of Gaussian noise.
- $S(n)$ The Markov chain that determines the regime of the discrete chain.

$X(n)$ · The values taken by a noise-free chain.

$Y(n)$ The values taken by a chain contaminated by noise (usually Gaussian).

$y(n)$ Observations of a chain contaminated by noise (usually Gaussian).

$\mathbf{y}_n$ The observation history $\{y(1), y(2), ..., y(n)\}$.

$\mathbf{s}_n$ The regime history $\{s(1), s(2), ..., s(n)\}$.

$L(n)$ The index of the level of the Markov chain (S_n, L_n).

n Index of discrete time.

N The Maximum time in the series.

$N_{(L)}$ The number of levels in either regime for a symmetric model.

$N_{(U)}, N_{(D)}$ The number of levels in the upper and lower regime for an asymmetric model.

Sample Moments

μ, μ' Mean of a clean and noisy sample.

υ, υ' Variance of a clean and noisy sample.

γ, γ' Skewness of a clean and noisy sample.

κ, κ' Kurtosis of a clean and noisy sample.

η, η' Fifth Central Moment of a clean and noisy sample.

ψ, ψ' Sixth Central Moment of a clean and noisy sample.

K_i The i^{th} Cumulant.

ζ A function of clean to noisy variance ratio.

Estimators and Distributions

F A cumulative distribution function (CDF).

E Empirical CDF.

f A probability density function.

D Test Statistic of the KS test.

H HyperGeometric Function.

R Estimator of ρ.

$Beta[a, b]$ The Beta Distribution.

$N[\mu, \sigma^2]$ Normal Distribution.

$G[a]$ Gamma distribution.

Models

$L[N_{(U)}, N_{(D)}]$ A Ladder Model.

$S[N_{(U)}, N_{(D)}]$ A Slide Model.

$UL[N_{(U)}, N_{(D)}, c]$ An Unravelled Ladder model.

$US[N_{(U)}, N_{(D)}, c]$ An Unravelled Slide model.

$LF[N_{(U)}, N_{(D)}]$ A Line Fit model.

$LFAR[N_{(U)}, N_{(D)}]$ A Line Fit model with autoregressive noise.

Chapter 1

Introduction

We describe data as a time series when it consists of a set of observations which are ordered in time. The study of time series is an ancient one, but the theoretical models we use are constantly evolving to allow greater levels of sophistication. In this chapter we shall look at one particular area of development, namely the development of time-series models with non-linear dynamics.

1.1 The Road to Non-Linearity

The examination of time series observations is hardly a new science. The record of sunspot data may be traced as far back as 28BC. It is fitting then, that the era of linear time series modelling began with a study of sunspots. The key feature of time series models is that they should capture not only the information contained in the values taken by the series, but also the information contained in the order in which those values occur. We do not expect this relationship between successive values to completely describe the process and will usually also include an irregular or unpredictable element, often termed 'noise'. These first models can be described as both linear and Gaussian.

- Linear, in this context, means that there is a straight-line relationship between a value and its predecessor/s.
- Gaussian refers to the use of the Normal distribution to model the noise that obscures the true values of the series when we try to observe them.

Both of these two assumptions are of great assistance in allowing simple, mathematically tractable models to be developed. On the downside they are also restrictions on the type of behaviour that can be modelled.

These first linear Gaussian models, proposed by Yule (1927) and Slutsky (1937), were to dominate the development of time series model building for many decades to come. This class of models, known collectively as autoregressive models, combine a deterministic term (dependent upon previous values) with a 'noise' term (independent Gaussian variates) to

form the distribution of the forecast variable. The term 'autoregressive' refers to the way in which the deterministic term is formed, calculated from a linear combination of values of the observed variable at earlier points in time. Example 1 shows the simplest form this model can take, the first-order Autoregressive model. This title is usually abbreviated to AR(1). The precise definition of an Autoregressive model such as this also requires that the model be stationary. For this to be the case there are restrictions on the values that can be taken by the constant terms. For instance, in Example 1 these restrictions require that $|\phi| < 1$. This class of models proved extremely successful and much work has been done developing it to greater generality.

Example 1 (An AR(1) Model)

$$Y(n) = \phi Y(n-1) + \epsilon(n)$$

Where $Y(n)$ is the values of the process,at time n, ϕ is a constant coefficient and $\{\epsilon(n)\}$ is a strict white noise process.

One way in which these models have been generalised is to extend the autoregressive behaviour to the noise term. Allowing both the deterministic and noise terms in the model to depend on previous values gives us the more general Auto Regressive Moving Average (ARMA) model, a simple version of which is shown in Example 2. For the process to be stationary the values taken by the constant coefficients are under the same restrictions to those found in the simpler Autoregressive model.

Example 2 (An ARMA(1,1) Model)

$$Y(n) - \phi Y(n-1) = \epsilon(n) - \psi\epsilon(n-1) \tag{1.1}$$

Where $Y(n)$ is the values of the process,at time n, ϕ and ψ are constant coefficients and $\{\epsilon(n)\}$ is a strict white noise process.

Largely due to the success of these models time series analysis has become a highly developed subject, and there are now well-established methods for fitting a wide range of models to time series data. These methods are well documented in literature devoted to time series, including those by Box & Jenkins (1970), Brillinger (1975), Koopmans (1995), Priestley (1981) and more recently Brockwell & Davis (1991) and Brockwell & Davies (2002).

To recognise the importance of these models is not to say that they do not have their limitations. At their heart are two key assumptions, and the strength of model in a given context is directly related to the strength of its assumptions. The two key assumptions of which we speak in this case are that the series is 'stationary' and that it is 'linear'.

- Stationarity: Roughly speaking, stationarity refers to the property of a series that the distribution of future values remains constant through time. To put it more technically, the probability distributions of the process are time-invariant. In practice there are different levels of stationarity (weak or second order) but we are only concerned with the general idea.

- Linearity: Where we refer to linearity we are requiring that the data has only a linear association with previous values. These linear relationships may extend beyond the previous value in the series, possibly even encompassing seasonal behaviour.

An alternative version of the ARMA model was proposed to deal with non-stationary series. If a time series appears to be non-stationary then we cannot use the traditional ARMA model. However, we may find that the first difference of the time series is itself stationary. If so then we could model this, the growth rate of the original series, using an ARMA model. This contruction is known as an Auto-Regressive Integrated Moving Average model (ARIMA). If we had a data series $X(n)$, which is was non-stationary, and we defined the differenced series $Z(n)$

$$Z(n) = X(n) - X(n-1)$$

Then fitting an ARIMA model to $X(n)$ would be equivalent to applying an ARMA model, as specified in (1.1), to $Z(n)$. We shall not be using ARIMA models in this research but the concept of working with a differenced series is important. It is a common theme in economic time series data where the growth rate, rather than the level, of a series is often modelled.

The assumptions of both linearity and stationarity are simplifications which may not hold in some circumstances. This would undermine the validity of a model that required those properties to be present. Specific applications may raise other queries that undermine the use of AR models. The area of application with which we are primarily concerned is the representation of econometric data, and any appraisal of the strengths and weaknesses of time series models will be done with reference to this context. Two examples of conditions under which linear models may be deemed unsuitable for use with economic time series are those showing evidence of time-irreversibility and those with varying volatility (Heteroskedasticity).

- Time-Irreversibility: In a time reversible model the statistical properties of a process are the same as those of the same process observed in reverse. Many authors find evidence of time irreversibility in observed time series. For instance, in business cycle fluctuations it is often documented that there is a tendency for downswings to be faster than upswings and Ramsey and Rothman (1996) find evidence of time irreversibility in many other macroeconomic variables. The Gaussian linear models we have been considering up to this point are generally considered unsuitable for modelling data

exhibiting time irreversibility. Since economic time series are 'true' time series in that they are non-anticipating of future values, there is no theoretical reason why such processes should be reversible. There is no reason then that we should restrict ourselves to Gaussian linear models.

- Heteroskedasticity: Although not unique amongst time series, those in economics often display periods during which they are more volatile and hence less predictable. We call series that display constant volatility Homoskedastic and those with varying volatility Heteroskedastic. The class of linear models are unsuitable for modelling this kind of phenomena.

There are many documented cases where Gaussian linear models are unsuitable. Often ways are found to transform the data so that these models can be used. The ARIMA models are an example of this. Despite modifications such as this, there remained the requirement for time series models capable of embracing these complexities rather than evading them.

1.2 Non-Linear Time Series Models

Faced with these challenges some revision was necessary and, given the power and simplicity of the existing models, it seemed natural to adapt the ARMA models rather than rejecting them completely. Two obvious ways in which the limitations of the previous linear models could be overcome were to drop the requirement that noise be Gaussian or make an attempt to incorporate some non-linear dynamics into the behaviour of the model. But these modifications were not the only new approaches, most of which were developed to address a specific perceived failing of the existing models. Much work has been done in this area and many non-linear time series models have been proposed in recent years. Between them they are capable of describing a large variety of non-linear structures. A thorough study of the field is available from Tong (1990). We shall give a brief summary of some of the models that are of current interest. Though we could not hope to cover all the models on offer, most of the key innovations have been included here.

1.2.1 Piecewise Linear Models

An early attempt to update the ARMA models was the class of Piecewise-Linear models. These retained many of the strengths of the traditional ARMA models while adding one key development. They allowed the definition of a set of linear models rather than a single model. The whole set is then included in the definition of the process, which is able to switch from one (linear model) to another on meeting certain criteria. For example a simple threshold autoregressive model would switch from the behaviour of one linear model to the other when the current value of the process crossed a predetermined threshold. This general approach is very simple to understand for anyone with experience of ARMA models but is capable of complex behaviour that had previously been impossible to model. As a

result many variations of this model have been developed, some quite specialised. In a Self Exciting Threshold AutoRegressive (SETAR) model a partition of the real line is defined. Each interval is associated with a particular linear model. When the observed series $Y(n)$ moves into a particular interval the movement of the process will be governed by the linear model associated with that interval. It is common for there to be a delay between a change in interval and a change in model. In the example given below it is the observed level of the series d steps earlier that controls the choice of models.

Example 3 (A SETAR(2;2,2) Model)

$$Y(n) = \begin{cases} \phi_{0,0} + \phi_{0,1}Y(n-1) + \phi_{0,2}Y(n-2) + \epsilon_{0,n} & If\ Y(n-d) \in R_0 \\ \phi_{1,0} + \phi_{1,1}Y(n-1) + \phi_{1,2}Y(n-2) + \epsilon_{1,n} & If\ Y(n-d) \in R_1 \end{cases}$$

Where $\phi_{i,j}$ and d are constant coefficients and R_0 and R_1 are intervals forming a partition of the real line. Both $\epsilon_{0,n}$ and $\epsilon_{1,n}$ are strict, white-noise processes.

Much of the innovation in the development of these models concerns methods for determining which of a set of models should be followed at a point in time. The ideas used in the SETAR models were extended even further in Open Loop Threshold Autoregressive models (TARSO). In these models the series no longer triggers its own transition from one regime (linear model) to another. In the TARSO models we have two processes running concurrently, an observable input and an observable output. The method of triggering a change is much the same as with the SETAR models with the regime inhabited by the observable output dependent on the value of the observable input series.

Example 4 (A TARSO(2;1,1,1,1) Model)

$$X(n) = \begin{cases} \phi_{0,0} + \phi_{0,X}X(n-1) + \phi_{0,Y}Y(n) + \epsilon_{0,n} & If\ Y(n-d) \in R_0 \\ \phi_{1,0} + \phi_{1,X}X(n-1) + \phi_{1,Y}Y(n) + \epsilon_{1,n} & If\ Y(n-d) \in R_1 \end{cases}$$

Where d and the ϕ are constant coefficients and R_0 and R_1 are intervals forming a partition of the real line. Both $\epsilon_{0,n}$ and $\epsilon_{1,n}$ are strict, white-noise processes.

In the TARSC model the two processes interact, with the observable output of one acting as the input to the other (i.e. $(X(n), Y(n))$ and $(Y(n), X(n))$.are both TARSO). Each of these models proposes more complex methods for determining the current regime of the process. One exception to this trend breaks the causal link between the level of the process and the regime it inhabited, and that is the Markov Switching model. In this variation the changes of regimes are determined not by the level of the process, but by an unobserved state variable typically modelled as a Markov chain.

Example 5 (A Two-Regime Markov Chain Driven Model)

$$Y(n) \begin{cases} = \phi_1 Y(n-1) + \epsilon(n) & if\ S(n) = 0 \\ = \phi_2 Y(n-1) + \epsilon(n) & if\ S(n) = 1 \end{cases}$$

Where $S(n)$ is a two-state Markov chain taking values 0 and 1. $Y(n)$ are the observations while $\{\epsilon(n)\}$ is a strict white noise process.

All these processes, however, share a common drawback. They are all dependent upon the idea of a structural break. By this we mean clear divisions, either in time or in level, between quite different behaviour patterns. In some circumstances, engineering perhaps, where the movement of a series may obey clearly understood laws we may know the position and nature of these structural breaks. Their application to economic applications may be less justified as we may not be able to say with certainty that any structural breaks exist. Before we could even start to fit models we would first have to estimate the number and nature of these breaks. There is a great danger here of increasing the number of levels until we had overfitted the model. Another problem with these methods is that we are still left with many of the weaknesses of linear models (see Leamer (1978) and Campbell *et al.* (1997) for a consideration of this issue).

1.2.2 Smooth Threshold Models

Some attempts have been made to free the threshold models from the problem of structural breaks by having one flexible model. This flexibility can be built in by allowing variation of the model itself over time. One way of doing this is to make the parameters of the model dependent on the earlier values (history) of the process. Amplitude Dependent Exponential Autoregressive models (EXPAR) introduce an exponential multiplier for the deterministic term that is dependent on the magnitude of previous observations. It allows the process to self regulate, bringing extreme values back towards a central level. In fact there are no pure constant terms in the expression at all, which has profound consequences for the behaviour of the model. Tong (1990) took the view that the resultant process may be "too restrictive for many applications other than pure vibrations".

Example 6 (An EXPAR(2) Model)

$$Y(n) = [a_1 + b_1.\exp(-cY^2(n-1))].Y(n-1) + [a_2 + b_2.\exp(-cY^2(n-1))].Y(n-2) + \epsilon(n)$$

Where a_i, b_i and c are constant coefficients
and $\{\epsilon(n)\}$ is a strict white noise process.

An attempt to avoid the restrictions of a rigid regime structure incorporates a random element to the deterministic term in addition to the noise. These Random coefficient autoregressive (RCA) models allow for unpredictable changes to the definition of the model itself. This will still produce a symmetric process though and also adds more parameters to be estimated, namely the properties of the random term.

Example 7 (An RCA(1) Model)

$$Y(n) = [a + B(n)].Y(n-1) + \epsilon(n)$$

Where $B(n)$ is a random variable independent of $\{\epsilon(n)\}$.

Another interesting method takes the idea of random multipliers and adapts it to the noise term rather than the deterministic term. This model, known as the Product Autoregressive (PAR) model, is noticeably different to the other alternatives. This is quite distinct from the previous models we have covered in that it has dispensed with additive noise entirely. The stated form of this model is a little too simple to allow much flexibility but the central feature of it will reappear later in this chapter in the (more significant) ARCH models.

Example 8 (A PAR Model)

$$Y(n) = \eta(n)Y^{\alpha}(n-1)$$

Where $\eta(n)$ is a sequence of independent, identically distributed positive random variables and α is a constant term.

1.2.3 Non Linear AutoRegressive Models.

Another attempt to generalise the existing models was to drop the requirement for linearity, without otherwise changing the approach. The autoregressive structure of the deterministic term in the model was simply extended to allow non-linear dependence on previous values of the process. This method has one obvious and one less obvious drawback. First it is clear that this will do nothing to address the Gaussian nature of the process, and secondly for higher orders of polynomial dependence on earlier values the process we have a problem of stability. Without some censoring of extreme values of the process it will only remain stable (or bounded) for certain polynomial relationships. This problem of stability can be addressed by limiting ourselves to the Bilinear (polynomial of order 2) case. The general case is termed the Bilinear Autoregressive Moving Average (BARMA) model. This has been used effectively to capture non-linearities with the data but does little to provide a flexible framework for modelling periodic changes in variance. The idea is introduced and explored in Granger & Anderson (1978)

Example 9 (A BARMA(1,1,1,1) Model)

$$Y(n) = \phi Y(n-1) + \psi\epsilon(n) + \lambda Y(n-1)\epsilon(n-1)$$

Where ϕ, ψ and λ are constant coefficients and $\{\epsilon(n)\}$ is the strict white noise process.

1.2.4 Deterministic Systems

An extremely novel contribution to the problem of non-linear series comes from the field of Chaos Theory. Relatively simple deterministic systems of equations have been shown to display extremely complex behaviour. It seems unlikely that this new approach will prove as useful as it is appealing due to the nature of econometric processes. In economic time series we do not generally expect to find the clear and rigid structural relationships between factors that you may find in Physics, for instance. This leaves us with no real reason to favour one kind of non-linearity over another which does not help us in model fitting. Another concern is that should we embark upon fitting a suitable general model we would have no way to conduct controlled experiments to allow us to estimate parameters. Due to the extreme sensitivity of these models to initial conditions any error in the estimate of the parameters could have an enormous effect on the behaviour of the final model. As a result it seems that this approach may remain a tantalising curiosity in econometric analysis rather than a useful tool.

1.2.5 Conditional Heteroskedasticity

Many of these different models have proved extremely useful in various different areas of time series, but none have really attempted to directly model the volatility observed in econometric data. The first really successful attempt to do this was the introduction of the ARCH model by Engle (1982). The two key features of the Autoregressive Conditional Heteroskedasticity (ARCH) model are the multiplicative error term and the scaling factor dependent on previous values of the process. By combining these two features, both seen before in other models, it is possible to meet most of the requirements we have stated.

Example 10 (An ARCH(2) Model)

$$\begin{aligned} If \quad Y(n) &= \epsilon(n)\sqrt{V(n)} \\ and \quad V(n) &= \phi_0 + \phi_1 Y^2(n-1) + \phi_2 Y^2(n-2) \\ with \quad \phi_0 &> 0 \text{ and } \phi_1, \phi_2 \geqslant 0 \end{aligned}$$

Where ϕ_i are constant coefficients, and $\{\epsilon(n)\}$ is white noise.

When $Y(n)$ and $V(n)$ are defined as in Example 10 then $Y(n)$ is said to follow an ARCH model. A further generalisation can be made to this to obtain the Generalised Autoregressive Heteroskedastic (GARCH) model, dropping the requirement for normality of the noise term. This leaves us with a model that is non-symmetric, non-Gaussian and can display heteroskedasticity. This is not to say that ARCH is the end of the story. Though the new model was quite successful in describing volatility dynamics it has shown certain weaknesses. As with many non-linear innovations the actual gain in terms of model fitting is relatively modest and this can overshadow the structural developments. The wide range

of variations on ARCH reflects the enthusiasm for the idea of modelling heteroskedasticity but it seems that the nature of the volatility modelled by ARCH is more suited to financial data than to many other areas.

1.3 Summary

As we can see, there has been a proliferation of ideas for developing time series models capable of displaying non-linear dynamics. Many of these models have achieved prominence in specific fields, but none have shown signs of achieving the same level of dominance the ARMA (and ARIMA) models had. This seems unlikely to change, regardless of future developments, given the breadth of the field of non-linear models. It is a mistake to think of non-linearity as being an extension to linearity. Rather, non-linearity should be seen as the whole, with linearity being one minor subset. It is possible models will be developed that try to span this field incorporating many different non-linear terms, but these could become unwieldy. Is seems more likely that methods will be developed for identifying the dominant dynamics and a suitable choice made for modelling them.

Chapter 2

Hamilton's New Approach

In his 1989 paper "A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle", Professor James Hamilton introduced a innovative method for modelling time series. We look briefly at the form of the model and the Bayesian filters he used to fit the model will be detailed. A consideration will be made of the influence of the paper and both the debate and development prompted by it.

2.1 Markov Switching Models

Despite the stunning proliferation of non-linear time series models in recent years none have seemed likely to span a range of applications nearly so wide as that achieved by the linear models. These non-linear models often seem restricted to specific areas of use. Each model incorporates a different type of non-linearity, and does so using a different mechanism. The strength of each model where its features are appropriate becomes its weakness when they are not. Even when limiting ourselves to the field of economic time series the field is too wide for one model to dominate. Diebold (1998) is of the opinion that "...many of the non-linearities relevant in fields such as finance simply don't seem to be important in macroeconomics..." and that volatility models, such as ARCH, were "...much more important in high-frequency financial data.". He goes on to stress the relevance of regime switching of the Threshold models to macroeconomic forecasting.

The concept of regime-switching (as characterised by the Piecewise Linear models) has the capacity to fit neatly with the well established idea of business cycles. As far back as the thirties Keynes (1936) suggested that an economy fluctuated between longer periods of growth and short-lived but violent contractions. Where more than one regime appears to exist, each displaying (self-) consistent behaviour that differs from that of the other regime/s then a regime switching model may be appropriate. Under these conditions one simple model would struggle to capture the behaviour of the series, but switching between a small number of simple models may be sufficient. One of the criticisms of Piecewise Linear models is the requirement for structural breaks. These are the points where the process

will switch from one model to another. Unlike previous threshold models the Markov switching model introduces switching between regimes that is not triggered by any quality or property of the series itself but by a hidden Markov chain independent of the observed series. The analysis undertaken by Hamilton (1989) looked at the first difference of the log of US GNP between 1952 and 1984. This can be seen as a natural extension of Neftci's (Neftci (1984)) analysis of U.S. unemployment data. Where Neftci deduced the current state of the hidden process using the growth rate of unemployment, Hamilton postulates that the current state of the Markov chain is independent of the data, and "...only one of many influences governing the dynamic process...". The use of the hidden Markov chain in this way has the advantage that it "...does not suffer from some of the statistical biases that models of structural breaks do; the regime shifts are "identified" by the interaction between the data and the Markov chain, not by a priori inspection of the data." (Campbell *et al.* (1997)).

These models, despite their simplicity, are a flexible and effective tool for use with any data we suspect may display irregular cyclic behaviour. They retain much of the simplicity of the linear models and indeed linear models could be seen as a special case of switching models.

2.1.1 Hamilton's Model

Early attempts to model the behaviour of economies faced the problem that many of the series were non-stationary. In order to progress economists focused on the growth rate of a series rather than their level. The assumption that underpinned much of the previous work on GNP figures is that the first difference of the log of GNP follows a linear stationary process. The 'New Approach' that Hamilton introduced was "...of specifying that first differences of the observed series follow a non-linear process rather than a linear stationary process.". Further he proposed that the process could be effectively explained using a piecewise linear model with switching between regimes controlled by a Markov chain independent of the visible series itself. He chose to use a chain with two regimes, expecting that these would represent the behaviour of the economy in periods of 'slow' and 'fast' growth. As it turned out these two regimes appeared to correspond to 'growth' and 'recession' states of the economy, fitting neatly into recognised measurements of previous business cycles. So closely did these states correspond to the published NBER estimates of the peaks and troughs of the business cycle that the research went on to propose that this kind of model would be useful as an "alternative objective algorithm" for determining and measuring their movements.

Now we shall specify the model in slightly more formal terms. For the original time series data we shall use the notation $\{z(n)\}$. This original series used is the level of real GNP measured at an annual rate in 1982 dollars. The series contains quarterly figures running from 1952: II until 1984:IV.

We transform this to obtain a new series $\{y(n)\}$ by taking 100 times the difference in

the log of $\{z(n)\}$.

$$y(n) = 100(\ln(z(n)) - \ln(z(n-1)))$$

It is proposed then, that this transformed variable can be modelled as a combination of a growth term $g(n)$ and a noise term $e(n)$. The growth term corresponds to the growth rate of the economy and is dependent upon the regime of a two-state Markov Chain $\{S(n)\}$. He found that his results may suggest investigating a higher order Markov process.

$$y(n) = g(n) + e(n)$$

The Markov Chain can take the values 0 and 1 and the growth term is derived from this using:

$$g(n) = (l_1 - l_0).s(n) + l_0 \tag{2.1}$$

where l_0 and l_1 are the two possible rates of growth (in regimes 0 and 1 respectively) and $s(n)$ indicates the regime occupied by the unobserved Markov Chain at time n.

The transition between states is governed by a first order Markov process given by:

$$\begin{aligned}
\Pr[S(n) &= 1|S(n-1) = 1] = a \\
\Pr[S(n) &= 0|S(n-1) = 1] = 1 - a \\
\Pr[S(n) &= 0|S(n-1) = 0] = b \\
\Pr[S(n) &= 1|S(n-1) = 0] = 1 - b
\end{aligned}$$

He chose to model the residual error terms using an AR process with constant coefficients ϕ_i and "arbitrarily" set the number of lags to 4 although it seems likely that this choice was influenced by the fact the figures were quarterly. So we have the chain

$$y(n) = (l_1 - l_0).s(n) + l_0 + e(n)$$

$$e(n) = \phi_1 e(n-1) + \phi_2 e(n-2) + ... + \phi_r e(n-r) + \epsilon(n)$$

On to this autoregression for $e(n)$ we add a Gaussian noise term $\{\epsilon(n)\}$. The process of building this model was influenced by an attempt to develop the model which allowed the computationally simplest maximum likelihood estimation. In addition to the model itself Hamilton proposes algorithms for use with this class of models.

2.1.2 Impact of Hamilton's Model

Since its publication, Hamilton's "seminal paper" (Kim *et al.* (2005)) has drawn a lot of attention. Applications of his model are numerous. For example Garcia (1998) applied his model to U.S. interest rates and inflation, Engle & Hamilton (1990) applied it to exchange rates and Cecchetti *et al.* (1990) applied it to the stock market. In particular,

for the measurement of macroeconomic fluctuations, the Markov switching model has become increasingly popular since Hamilton's application of the technique to measure the U.S. business cycle. There have been a number of recent investigations of business cycles, in individual countries, with Markov switching models, including those of Lam (1990), Albert & Chib (1993), Goodwin (1993), Acemoglu & Scott (1994), Diebold & Rudebusch (1996), Diebold *et al.* (1994), Hamilton & Susmel (1994), McCulloch & Tsay (1994), Ghysels (1994), Kim (1994), Kähler & Marnet (1994), Krolzig & Lütkepohl (1995), Sensier (1996), Clements & Krolzig (1998), Boldin (1996), Krolzig (1997). There have also been studies in the analysis of international business cycles using similar models by Phillips (1991) and Filardo & Gordon (1994). These papers, inspired by Hamilton, can be broadly grouped into three categories.

- The first category comprises those who have applied the model, perhaps with some changes, in a different context. For instance, Goodwin simply takes the model and applies it to eight market economies. He finds only a little improvement over linear models and evidence that the model does not capture all the non-linearities for some economies. He also agrees that the model is valuable for dating business cycles. Acemoglu and Scott, after concluding that "strong non-linearities" exist in the UK labour market, successfully utilise the Markov switching model. The credibility of the model seems to have taken it beyond being an academic curiosity. We can find the ideas being applied in earnest by Buckle *et al.* (2002) in a New Zealand Treasury working paper to model growth and volatility regimes in the economy. They use several different models with three or four growth rates and multiple volatility regimes. They had little trouble fitting the models using maximum likelihood and the EM algorithm, and were again impressed by the usefulness of the model in tracking regime changes.

- The second category of papers concentrated on the theoretical and practical weaknesses of Hamilton's paper and offered alternative approaches. Hansen (1992) focuses on weaknesses, that Hamilton was aware of, with hypothesis testing of the model. He offers an alternative method for testing the fit of the model and suggests that a simple switching model (where the states arrive independently over time) would be more appropriate than a Markov switching model. He is unable to reject the hypothesis that the good fit of Hamilton's original model is "simply due to sampling error" but finds his simple switching model, with more state dependence in the parameter values (a change Hamilton was to make later) fits far better than the original, linear, AR(4) model. Boldin (1996) is even more scathing when appraising the robustness of Hamilton's original model. He find numerous local maxima when trying to obtain a Maximum Likelihood Estimate (MLE), a grave concern when using hill climbing optimisation techniques. Curiously he only reported similar results to Hamilton when

using exactly the same sample period and starting values. He also felt, as did Goodwin, that a three regime model for GNP growth was "more robust and plausible".

- The third, and final, category built upon Hamilton's foundations to develop the idea further. Lam (1990) generalised the model to the case in which the autoregressive component need not contain a unit root and suggested an algorithm for fitting it. His paper is not conclusive on the issue of the relative merits of the various stages of model, finding evidence to support each. Kim (1994) extends the Markov switching model to general state-space representation and adds the work previously done by Lam. He also claims a degree of success in improving efficiency in generating maximum likelihood estimations.

As we can see from this selection, there is both praise and criticism for parts of Hamilton's paper. A true indication of the impact of the paper can be found in the fact that so much has been written dissecting (or building upon) his work.

2.1.3 Possible Developments

We have heard already of many changes that have been, or could be, made to Hamilton's model. Some attempt to generalise the model, leading perhaps to new classes of models developing, while others question specific assumptions or restrictions.

Many of the papers using this model concluded (or simply decided) that two-regimes were not enough. Three have been suggested when modelling European economies, up to four were used by Buckle, Haugh and Thomson but many more could be used.

Hamilton settled on dependence of his autoregressive component on up to four past values although this does not appear to have been a structural decision. Given the subjective nature of this decision and the developments in generalising the model, this, and many other questions, can be seen as more relevant to the application of the model rather than the development of the model itself. The various possibilities left to explore are likely to involve changing principles that will fundamentally change the structure, leading to a new model rather than a refined or alternative version of the old one.

2.2 Fitting the Markov Switching Model

This kind of estimation problem is sometimes known as a problem of Incomplete Data. We have a set of observations $\mathbf{y}_N = \{y(1), y(2), ..., y(N)\}$ that we propose have been generated by a model with structural parameters $\boldsymbol{\Theta}$. We wish to make inference about $\boldsymbol{\Theta}$ using $\mathbf{y}_N$. The problem arises from the fact that the model does not obtain $\mathbf{y}_N$ directly from $\boldsymbol{\Theta}$. Instead it uses $\boldsymbol{\Theta}$ to generate $\mathbf{s}_N = \{s(1), s(2), ..., s(N)\}$ and obtains $\mathbf{y}_N$ from a combination of $\boldsymbol{\Theta}$ and $\mathbf{s}_N$. It is our inability to observe the values of $\mathbf{s}_N$ (the missing data) that gives this type of problem its name.

There are two different ways of approaching model fitting by Maximum Likelihood in a case like this. The first is to attept to identify the regime history $\mathbf{s}_N$ that allows us to maximise likelihood. We shall use this approach later and will refer to it as the Sequence Maximum Likelihood (SML). Sclove(1983) used this approach obtaining

$$f(\mathbf{y}_N, \mathbf{s}_N | \mathbf{\Theta}, \mathbf{\Pi})$$

Where $\mathbf{\Pi}$ are the initial conditions.

This was then maximised with respect to $\mathbf{\Theta}$, $\mathbf{\Pi}$ and $\mathbf{s}_N$. His result was therefore based on an imputed historical sequence for $S(n)$. One of the notable weaknesses of this method is that imputing the history of $S(n)$ greatly increases the number of variables to be estimated. This can lead to difficulties in maximising the likelihood function.

The second approach, and that favoured by Hamilton, is to maximise only with respect to $\mathbf{\Theta}$ to obtain

$$f(\mathbf{y}_N | \mathbf{\Theta}, \mathbf{\Pi})$$

This approach is based on ideas developed by Cosslett & Lee (1985) and gave rise the Bayesian Smoothing algorithms used by Hamilton. As no attempt is made to recontruct the original state sequence only one pass through the data is required to obtain inference hence guaranteeing a bounded time to generate a likelihood figure. There is a potential weakness in this method in that it treats every interval between observations in isolation from the rest of the series deriving the likelihood of the series from the series of likelihoods. The algorithm is detailed below, first in simple terms, and then later in more detail.

2.2.1 The Basic Filter

There are two versions of the algorithm presented by Hamilton. The first is the Basic Filter, a simple and effective method for obtaining a likelihood value for each observation $y(n)$ of the observed series. The full sample smoother extends this principle to allow inference to be made about the value of the missing data $s(n)$ using the whole sample rather than just the recent history. When using the basic filter a repetition of 5 simple steps for each observation will give us the Log-Likelihood for a set of parameter values.

We can summarise the algorithm in words.
Let us say we have been observing the series up to time n and have an estimate of the distribution of $S(n)$ based on the previous observations. This enables us to forecast the distribution of $S(n+1)$. We can then modify this distribution using the likelihood of observing $y(n)$ in either regime. This, modified distribution, allows us to forecast $S(n+2)$ and so on. We obtain the likelihood of each observation $y(n+1)$ from the modified distribution of $S(n+1)$ as this is constucted from the likelihood in each regime, weighted by the inferred distribution between them.

A more formal description of the method is given below. The notation we shall use is as follows

$S(n)$ is the state of the Markov Chain at time n.

$Y(n)$ is the observed value at time n.

ϕ_i are the coefficients of the AR process.

σ is the standard deviation of the Gaussian noise.

r is the maximum lag of the AR process.

$\mathbf{y}_n$ is the regime history $\{y(n), y(n-1), ...y(1), y(0), ..., y(-r+1)\}$

Note that $[y(0), ..., y(-r+1)]$ are the first few values that provide $y(1)$ with a history required as input for the r^{th} lag autoregression.

The Basic Filter (for time n) accepts as input the joint conditional probability given all previously observed data.

$$\Pr[S(n-1) = s(n-1), S(n-2) = s(n-2), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1}]$$

and has two outputs, concerning time n.

The joint conditional probability (for time n) using all data up to, and including, time n.

$$\Pr[S(n) = s(n), S(n-1) = s(n-1), ..., S(n-r+1) = s(n-r+1)|\mathbf{y}_n]$$

and the conditional likelihood of the observation $y(n)$

$$f(y(n)|\mathbf{y}_{n-1})$$

Given the input at a point in time n we apply the following 5 steps to find the output at time n which then acts as the input at time $n+1$.

- **Step 1: Forecast current distribution from past experience**
  By the Markov property of $S(n)$

$$\begin{aligned}\Pr[S(n) = s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1}] \\ = \Pr[S(n) = s(n)|S(n-1) = s(n-1)] \\ \times \Pr[S(n-1) = s(n-1), S(n-2) = s(n-2), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1}]\end{aligned}$$

- **Step 2: Modify the forecast distribution using current value**
  By incorporating the likelihood of each regime history having given rise to the observation

$$\begin{aligned} f(y(n), S(n) &= s(n), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1}) \\ &= f(y(n)|S(n) = s(n), ..., S(n-r) = s(n-r), \mathbf{y}_{n-1}) \\ \times \Pr[S(n) &= s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1}] \end{aligned}$$

  Where we know that, if $g(n)$ is defined as in 2.1

$$\begin{aligned} f(y(n)|S(n) &= s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r), \mathbf{y}_{n-1}) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}((y(n) - g(n)) - \sum_{k=1}^{r}[\phi_k(y(n-k) - g(n-k))])^2\right] \end{aligned}$$

- **Step 3: Evaluate Likelihood of current observation**
  Summing over all probabilities

$$\begin{aligned} f(y(n)|\mathbf{y}_{n-1}) &= \sum_{s(n)=0}^{1} \sum_{s(n-1)=0}^{1} ... \sum_{s(n-r)=0}^{1} f(y(n), S(n) = s(n), S(n-1) = s(n-1), ... \\ ..., S(n-r) &= s(n-r)|\mathbf{y}_{n-1}) \end{aligned}$$

- **Step 4: Obtain current distribution from probabilities**
  As we know that some value must be taken by $S(n)$ we tranform from absolute probabilities to a distribution

$$\begin{aligned} \Pr[S(n) &= s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r)|\mathbf{y}_n] \\ &= \frac{f(y(n), S(n) = s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r)|\mathbf{y}_{n-1})}{f(y(n)|\mathbf{y}_{n-1})} \end{aligned}$$

- **Step 5: Adjust dimension of current distribution back to r**
  By summing over the two regimes of $S(n-r)$

$$\begin{aligned}\Pr[S(n) &= s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r+1)|\mathbf{y}_n] \\ &= \sum_{s(n-r)=0}^{1} \Pr[S(n) = s(n), S(n-1) = s(n-1), ..., S(n-r) = s(n-r)|\mathbf{y}_n]\end{aligned}$$

The algorithm must be started with some initial distribution. This could be postulated or estimated. The simplest approach is to use the stationary distribution of the process, given the proposed parameter values. This is the method used by Hamilton. Another approach that has been suggested is to run the filter backwards and take the final distribution (at time $n = 0$), as this will allow the full set of observations to 'suggest' the most fitting initial distribution.

2.2.2 Full Sample Smoothing

The basic filter proposed by Hamilton is effective for evaluating the likelihood of an observed series, but has limitations when it comes to imputing the history of $S(n)$. The full sample smoother is intended to "draw a more reliable inference about the lagged value of the state using currently available information" (Hamilton(1989)). The basic filter did make inference about the current regime $s(n)$, but used only observations predating this i.e. $\{y(m) : m \leqslant n\}$. The full sample smoother is intended to rectify this.

Summarising the algorithm in words.
From the results of the basic filter we have a set of inferred distributions. Each element in these distributions represents one unique value of the Markov chain $[S(n), S(n-1), ..., S(n-r+1)]$. *We take the first element and find the likelihood of the future observations of the series if the Markov chain occupied this element at time n. We then do likewise for each of the other possible regimes and then modify the distribution between elements using the likelihood of the future observations. The resulting distributions are our full sample inference of* $\{s(n)\}$.

The input to the Full Sample Smoother is the output of the Basic Filter (see Section 2.2.1). We require the likelihood values of the observations and the regime distributions at each point. These are the starting points for the new algorithm.

$$\begin{aligned}&f(y(n)|\mathbf{y}_{n-1}) \text{ and} \\ \Pr[S(n) &= s(n), S(n-1) = s(n-1), ..., S(n-r+1) = s(n-r+1)|\mathbf{y}_n]\end{aligned}$$

The output from the Full Sample Smoother will be the regime distributions modified to include information from future values of the series, $\{y(n+1), y(n+2), ..., y(N)\}$

$$\Pr[S(n) = s(n), S(n-1) = s(n-1), ..., S(n-r+1) = s(n-r+1)|\mathbf{y}_N]$$

To obtain the Full Sample Inference at a time n we apply the following steps:

- **Step 1: Take one possible regime history**
  One such history may be

$$[S(k), S(k-1), ..., S(k-r+1)] = [\hat{s}(k), \hat{s}(k-1), ..., \hat{s}(k-r+1)]$$

  Start with the distribution

$$\Pr[S(k) = s(k), S(k-1) = s(k-1), ..., S(k-r) = s(k-r+1)] = \quad (2.2)$$
$$\begin{cases} 1 & If\ S(k) = \hat{s}(k)\ \&\ ...\ \&\ S(k-r+1) = \hat{s}(k-r+1) \\ 0 & Otherwise \end{cases}$$

- **Step 2: Run the Basic Filter from this point**
  Start the basic filter from this point, $n = k$, and run it forward using (2.2) as an initial distribution. This will produce the following likelihood values for each future observation $y(n)$

$$f(y(n)|S(k) = \hat{s}(k), S(k-1) = \hat{s}(k-1), ..., S(k-r+1) = \hat{s}(k-r+1), \mathbf{y}_{n-1})$$

- **Step 3: Find the smoothed probability**
  Use the future likelihood values to modify the current distribution to give the probability for the condition $[S(k), S(k-1), ..., S(k-r+1)] = [\hat{s}(k), \hat{s}(k-1), ..., \hat{s}(k-r+1)]$. The modified probabilities are given by

$$\begin{aligned} \Pr[S(k) &= \hat{s}(k), ..., S(k-r+1) = \hat{s}(k-r+1)|\mathbf{y}_N] \\ &= \Pr[S(k) = \hat{s}(k), ..., S(k-r+1) = \hat{s}(k-r+1)|\mathbf{y}_k] \\ &\quad \times \frac{f(y(k+1)|S(k) = \hat{s}(k), ..., S(k-r+1) = \hat{s}(k-r+1), \mathbf{y}_k)}{f(y(k+1)|\mathbf{y}_k)} \\ &\quad \times \frac{f(y(k+2)|S(k)_k = \hat{s}(k), ..., S(k-r+1) = \hat{s}(k-r+1), \mathbf{y}_{k+1})}{f(y(k+2)|\mathbf{y}_{k+1})} \\ &\quad \times ... \times \frac{f(y_N|S(k) = \hat{s}(k), ..., S(k-r+1) = \hat{s}(k-r+1), \mathbf{y}_{N-1})}{f(y_N|\mathbf{y}_{N-1})} \end{aligned}$$

- **Step 4: Repeat Steps 1-3**
  Perform the steps 1-3 for each possible set of regimes $[\hat{s}(k), \hat{s}(k-1), ..., \hat{s}(k-r+1)]$

- **Step 5: Rescale to obtain the Full Sample Inference**
  These values are the smoothed probabilities conditional on the values of $[s(k), s(k-1), ..., s(k-r+1)]$ at time $n = k$. Standardising so they sum to 1 will give the inferred distribution between states and the output of the Full Sample Smoother

This process is quite time intensive to use and complex to encode in its current form. In order to produce a set of smoothed probabilities we require many passes through the data set. If we have a data set $\{y(n) : 1 \leqslant n \leqslant N\}$ then we require:

1 pass of N steps to complete the Basic Filter

1 pass of $N-k$ steps for the k^{th} of N observations

$$\text{Total number of steps required} = N + \sum_{n=1}^{N-1}(N-n) = N + \frac{N(N+1)}{2} = o(N^2)$$

There is a quicker way.

2.2.3 The Matrix Approach

When these Filters are implemented they are required to record large amounts of information as they operate in the form of distributions and likelihoods.

At the cost of even greater storage requirements it is possible to reduce the time required by the Full Sampler Smoothers. The key is to replace the mountain of iterations with some matrix algebra. This allows us to greatly reduce the number of steps required by removing much of the repetition. When performing the Bayesian steps given by Hamilton the probabilities are adjusted at every step to normalise them. As long as the cell values do not become too small (and we can correct them if they do) we can dispense with this step and represent the whole Filter in matrix form, normalising at the end of the process, and still arrive at the same figures. We shall consider the simplest case, with no autoregression, to keep the matrices manageable.

First we need to define some matrices:
The Forward Regime Transition matrix is given by F.

$$Where\ F_{ij} = \Pr[S(n+1) = j | S(n) = i]$$

And the Backward Regime Transition matrix by B.

$$Where\ B_{ij} = \Pr[S(n-1) = j | S(n) = i]$$

As we are dealing with the relatively straightforward case of a finite number of states in a irreducible matrix we will always be able to find a Backward Transformation matrix.

If we then define a series of (diagonal) Likelihood Matrices $L(n)$.

$$\begin{aligned} L_{ik}(n) &= f(y(n)|S(n)=k) \; if \; i=k \\ and \; L_{ik}(n) &= 0 \; otherwise \end{aligned}$$

Then we have:

$$\begin{bmatrix} f(y(n+1)|s(n)=0) \\ f(y(n+1)|s(n)=1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times F \times L(n+1) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$\begin{bmatrix} f(y(n-1)|s(n)=0) \\ f(y(n-1)|s(n)=1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times B \times L(n-1) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Then, in general

$$\begin{aligned} & \begin{bmatrix} f(y(n+1),...,y(N)|s(n)=0) \\ f(y(n+1),...,y(N)|s(n)=1) \end{bmatrix} \\ = & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times F \times L(n+1) \times ... \times F \times L(N) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned} \tag{2.3}$$

And of course

$$\begin{aligned} & \begin{bmatrix} f(y(1),...,y(n-1)|s(n)=0) \\ f(y(1),...,y(n-1)|s(n)=1) \end{bmatrix} \\ = & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times B \times L(n-1) \times ... \times B \times L(1) \times \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} \end{aligned} \tag{2.4}$$

Note that we can add an initial distribution Π if we wish. We obtain the inference on $S(n)$ by

$$\begin{aligned} f(\mathbf{y}_N|s(n) &= i) = f(y(1),...,y(n-1) \mid s(n)=i) \\ \times f(y(n)|s(n) &= i) \times f(y(n+1),...,y(N) \mid s(n)=i) \end{aligned}$$

The inferred distribution is given by

$$\Pr[S(n)=i|\mathbf{y}_N] = \frac{f(\mathbf{y}_N|s(n)=i)}{f(\mathbf{y}_N|s(n)=0)+f(\mathbf{y}_N|s(n)=1)}$$

At first sight this may not appear to achieve anything new. Indeed all we have done is perform the same calculation using matrices. The time saving can be found when we evaluate and then record two series of matrices, one moving forward through the series and

the other moving backwards. This is done step by step by defining a Forward and Backward matrix for each observation. For the n^{th} observation, $y(n)$ we would have matrices $F(n)$ and $B(n)$ respectively. We start with stationary distributions Π_F and Π_B that we can either choose ourselves or obtain using the Filter.

$$F(N) = diag(\Pi_F) \; and \; B(1) = diag(\Pi_B)$$

Where I is the Identity Matrix.

We can then we find the other matrices using

$$\begin{aligned} F(n-1) &= F \times L(n) \times F(n) \\ B(n+1) &= B \times L(n) \times B(n) \end{aligned}$$

This will give us a set of N Forward Matrices and N Backward Matrices. We then store these and use them to determine the smoothed probabilities. We will require one more sweep through the data set to do this. Evaluating

$$f(\mathbf{y}_N|s(n)=0) = \begin{bmatrix} 1 & 0 \end{bmatrix} \times B(n) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times f(y(n)|s(n)=0) \times \begin{bmatrix} 1 & 0 \end{bmatrix} \times F(n) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and so on to find the probability densities for each regime. From these we can evaluate the distribution between states for this observation by rescaling so

$$\Pr[s(n)=0|\mathbf{y}_N] = \frac{f(\mathbf{y}_N|s(n)=0)}{f(\mathbf{y}_N|s(n)=0) + f(\mathbf{y}_N|s(n)=1)}$$

We obtained this set of smoothed probabilities using

1 pass of N steps to obtain the Forward Matrices

1 pass of N steps to obtain the Backward Matrices

1 pass of N steps to derive the probabilities

Requiring a total of $3N$ steps, which is of order $o(N)$. This is a significant improvement over the original application of the smoother.

2.2.4 Future Observations

Of course this matrix approach is not limited to using the Full-sample. We can apply the same method when adding a fixed number of future observations to the inference. If we retain the notation of the last section and define

$$F_{ij} = \Pr[S(n+1) = j | S(n) = i]$$

and

$$\begin{aligned} L_{ik}(n) &= \Pr[Y(n) = y(n)|S(n) = k] \; if \; i = k \\ and \; L_{ik}(n) &= 0 \; otherwise \end{aligned}$$

And we are at time t, and have our inferred distribution (based on all data up to and including time t) for the current time point

$$\begin{bmatrix} \Pr[S(n) = 0|\mathbf{y}_n] \\ \Pr[S(n) = 1|\mathbf{y}_n] \end{bmatrix}$$

In order to incorporate r future value in the inference we work forward from t

$$\begin{bmatrix} f(y(n+1),...y(n+r)|S(n) = 0, \mathbf{y}_n \\ f(y(n+1),...y(n+r)|S(n) = 1, \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \Pr[S(n) = 0|\mathbf{y}_n] & \Pr[S(n) = 1|\mathbf{y}_n] \end{bmatrix}$$
$$\times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times F \times L(n+1) \times ... \times F \times L(n+r) \times \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

thus including the future values in the inference.

This is almost the same method as for the Full Sample Smoother. Where the programming language chosen allows matrix operations this approach greatly simplifies the process (and performance) for more complex cases.

2.2.5 Performance

The improvement in performance can be easily demonstrated by measuring the time taken to derive the inferred distributions for the same data using the two different methods. We compare the evaluation of the likelihoods using both looping and matrices when working with a Markov process with $N_{(R)}$ regimes.

First we examined the increasing run times for both methods as the number of levels increases (see Table 2.1). We can see a huge difference between the times taken to obtain the inference. In both cases, though, the expected time seems to be increasing in a linear fashion.

Next we took the same 2-regime Markov Switching model and obtained full sample inference with sample of different lengths to see how this affected the performance. The results are displayed in Table 2.2. This time we find the expected time increases linearly for the Matrix method but exponentially for the Looping method.

The difference in performance is so great that it reduces the expected time for full sample inference down to roughly three times that required for running on the Basic Filter. When working with the Basic Filter it is often desirable to reduce the number of parameters requiring estimation by running the Filter in reverse to obtain the initial distribution. This can be done as part of the process of obtaining Full Sample Inference, allowing both methods

Number of Regimes (R)	Time in Seconds Matrices	Loop
1	0.2700	34.6500
2	0.2700	72.1730
3	0.2910	109.0070
4	0.3000	146.1100
5	0.3110	184.1050
6	0.3100	228.6590

Table 2.1: A comparison of the performance of the Matrix Method and the original Looping algorithm for performing Full Sample Inference for a Markov Switching Model with N(R) regimes.

Length of Series(N)	Time in Seconds Matrix	Loop
10	0.0100	0.1210
40	0.0500	0.9510
70	0.0800	2.8240
100	0.1100	5.7080
130	0.1500	9.5640
160	0.1700	14.4610

Table 2.2: A comparison of the performance of the Matrix Method and the original Looping algorithm for performing Full Sample Inference for a 2 regime Markov Switching Model with time series of length N.

to be applied simultaneously, with considerable time savings.

2.3 The Pointer Filter

In Hamilton's model we have two regimes, upper and lower, each consisting of one level. This allows the hidden Markov chain $S(n)$ two possible regimes to occupy for each point in the time series. But in order to correctly model the autoregression at the current time point we also require knowledge of its recent history. It is much easier to work with a Markov chain so we construct one chain from the current state and recent history. For instance, if we have a r^{th} order autoregression then we define a chain $V(n) = [S(n), S(n-1), ..., S(n-r+1)]$

If we continue to work with a 2-regime chain $S(n)$ then we can define the Markov chain, $V(n)$, as a 1-dimensional vector.

$$\begin{array}{rll} V(n) = 0 & if & s(n) = 0, ..., s(n-r+2) = 0, s(n-r+1) = 0 \\ V(n) = 1 & if & s(n) = 0, ..., s(n-r+2) = 0, s(n-r+1) = 1 \\ V(n) = 2 & if & s(n) = 0, ..., s(n-r+2) = 1, s(n-r+1) = 0 \\ \vdots & & \vdots \\ V(n) = 2^{r+1} - 1 & if & s(n) = 1, ..., s(n-r+2) = 1, s(n-r+1) = 1 \end{array}$$

It is quite straightforward to construct a transition matrix for the process as each level of V can lead only to two others. Then we can obtain an initial distribution, as Hamilton did, by starting the process in the stationary state, between $V(-r+1) = 0$ and $V(-r+1) = 1$ say, and applying the transition matrix r times.

The only problem with this approach arises as we increase the complexity of the model. The dimensions of the distributions of V, and its transition matrix, increase with the power of the order of the autoregression.

$$\text{The number of levels of } V = 2^{r+1}$$

As r rises the number of levels (and hence calculations required) becomes very large very quickly. This explosion of levels can occur in another way, if we try to introduce more levels into the driving Markov signal for instance. If we were to introduce more levels to allow us to model a gradual transition in the growth rate $g(n)$, we find the levels rising much faster. For instance if we increase r by 2 we quadruple the number of levels, but if we double r we multiply the number of levels by 2^r.

The kind of econometric behaviour we are interested in modelling is the existence of periodic changes of behaviour. In general we are not so concerned with cases where these regime changes happen too frequently as this would give rise to mixtures of distributions less suitable for the application of our Bayesian Filters. If we consider only the cases where the switches are few and far between, there is an approximation that may help us. Rather than recording the precise state history in the distribution of the Markov process we retain

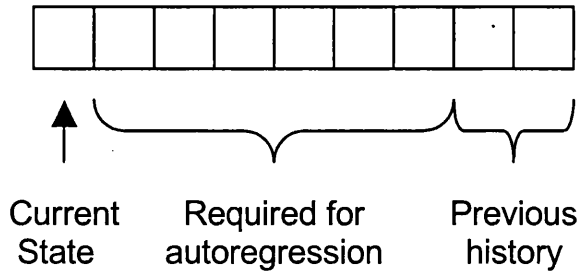

Figure 2.1: A visual representation of recent history of the Markov Chain $S(n)$ that is recorded using a Pointer chain $W(n)$.

information about only the most recent switchpoints. In this case we construct a different Markov chain, $W(n)$ that will record the history of $S(n)$ We shall construct a chain that records only the two most recent switchpoints.

We define $W(n) = [u(n), d(n)]$ where $u(n)$ is the measurement at time n of the time that has elapsed since the most recent upward switch (from $S(k-1) = 0$ to $S(k) = 1$) and $d(n)$ as the time since the most recent switch down (from $S(k-1) = 1$ to $S(k) = 0$).

We are still concerned only with an autoregression of lag r and a driving signal $S(n)$ of 2 levels. The most recent switch that could have taken place is at $k = n$ and if the last switch occured before $n - r + 1$ it will not directly affect the autoregression. If the last two switches occured before this point then we shall record the most recent as having occured at time $n - r$ and the other at time $n - r - 1$. In order to use a Bayesian Filter on this we require a number of possible values of $u(n)$ and $d(n)$ of $N_{(W)} = r + 3$. A visual representation for this history is given in Figure 2.1.

The Markov chain we define to contain this information is $W(n)$ and we would define the levels in the following way:

$$
\begin{array}{rlll}
W(n) = 0 & if & u(n) = n - N_{(W)} + 1 \quad and & d(n) = n - N_{(W)} + 1 \\
W(n) = 1 & if & u(n) = n - N_{(W)} + 1 \quad and & d(n) = n - N_{(W)} + 2 \\
W(n) = 2 & if & u(n) = n - N_{(W)} + 1 \quad and & d(n) = n - N_{(W)} + 3 \\
\vdots & \vdots & & \\
W(n) = N_{(W)} & if & u(n) = n - N_{(W)} + 2 \quad and & d(n) = n - N_{(W)} + 1 \\
W(n) = N_{(W)} + 1 & if & u(n) = n - N_{(W)} + 2 \quad and & d(n) = n - N_{(W)} + 2 \\
\vdots & \vdots & & \\
W(n) = N_{(W)}^2 - 1 & if & u(n) = n \quad and \quad d(n) = n &
\end{array}
$$

Of course certain of these levels will remain at probability 0 as two switches will not be allowed to occur at the same point. Examples of such levels are $W(n) = 0$, $W(n) = N_{(W)}+1$

etc. The transition matrix is easy to construct since

$$u(n+1) = \min(N_{(W)}, u(n)+1) \; and \; d(n+1) = \min(N_{(W)}, d(n)+1)$$

As long as

$$d(n+1) \neq u(n+1)$$

Unless a new switch occurs. We also know the regime at each point as if

$$u(n) < d(n) \; then \; s(n) = 1$$

And vice-versa for $u(n) > d(n)$.

Overall this is hardly a simple alternative and care must be taken when transforming from regime representation $\{s(n), ..., s(n-r+1)\}$ to Pointer representation $W(n)$, but it has the advantage of being less affected by the demands of incorporating the recent switching history into the Filter. The model was termed the Pointer model and an example of its use can be found on page 177. The results were encouraging in that it returned very similar estimates to the original smoother. Further tests to use the method to introduce gradual switching were less successful when the higher order autoregressions were added as the likelihood space appeared very complex and recognising convergence proved difficult. Despite this the results were positive enough to suggest that for very complex cases this may prove a useful way to make the burden of calculation manageable.

Chapter 3

Maximising Likelihood

Maximum Likelihood is a relatively straightforward and flexible method for fitting non-linear time series models. A model can be quickly applied to find the Likelihood of a particular set of parameter values using numerical methods. The more problematic part is exploring the parameter space to maximise this function. In this chapter we examine some of the options available for identifying the Maximum Likelihood Estimate.

3.1 Nature of the Modelling

The type of problem we are facing is that of parameter estimation with incomplete data. The nature of the missing data is a hidden 'driving' process influencing the observed series. We shall be assuming this hidden process to be Markov chain with constant transition probabilities. Of course because this process is not visible the regimes occupied by the process at different times are undefined variables. We are faced with a choice of whether or not to treat the history of regimes as parameters to be estimated. If so, then for a series of length N we would add N extra parameters. When working with Hamilton's filter we avoid doing so but we are still faced with a parameter space of high dimension. To make matters worse there may be a number of dimensions which are strongly dependent on each other. In this kind of situation direct methods of obtaining parameter estimates are rarely available. We have to resort to the likelihood function to measure the suitability of a set of parameters and then use numerical optimisation methods to find the most suitable set. The most suitable set, as judged by this method, are known as the Maximum Likelihood Estimate (MLE). The process of optimising any function brings difficulties of its own. We examine several numerical methods for finding the MLE and discuss their suitability.

3.2 Hill Climbing

We need to find the 'best' estimates for the parameters we are studying. We do this by finding which set of parameter values gives the greatest likelihood value for the given

sample. For many dimensioned parameter spaces the simple approach of exploring every point is impossible when considering continuous or unbounded variables as parameters. We need, then, to find another way of finding the MLE. Many methods are possible, with hill climbing being the simplest. The idea behind the hill climbing algorithm is that if we move closer to the MLE our likelihood value will rise, due to the continuous nature of the likelihood function. In fact if the distribution of the likelihood function is unimodal then this is a reliable and effective method. In order to maximise the MLE we simply look at gradient of the MLE and look at which 'direction' in parameter space will give us the greatest gain.

3.2.1 The Hill Climbing Algorithm

The concept behind this method is simple but there are many ways of applying it. In simple terms the likelihood is calculated at a point in parameter space. If we obtain the likelihood in a selection of local points then we can estimate the 'gradient' of the likelihood function relative to changes in each of the parameters. From this an optimum vector can be found that will give the direction of movement in parameter space that will result in the greatest increase in likelihood. By repeatedly moving and then recalculating the optimum vector a journey is taken through parameter space. In a simple distribution with only one local maximum this sequence of steps should always lead us towards the MLE. As the gradient at the current point refers to an arbitrarily small step, if our step does not take us to a higher level of likelihood then our step was too long. By only stepping to a point with a higher likelihood value we move 'uphill' (eventually) taking us close to the MLE. This approach sounds straightforward to apply and indeed can be. The real complexity arises from the many techniques that are available for optimising the movement of the process, by controlling the length of the steps taken. The real weakness of this method is exposed when there are multiple local maxima in the likelihood function. In this case the journey, blindly climbing uphill, can become 'trapped' in one of the local maxima and never find its way to the global maximum, the MLE. Whether or not this happens depends solely on the choice of starting point, and so a wide range of possible starting points would have to be used to ensure all maxima were encountered. Even if we are careful, using several alternative starting points, if we are dealing with a sufficiently complex likelihood function then this method could not be relied upon to locate the MLE.

3.2.2 Hill Climbing in practice

The algorithm is relatively simple to understand and implement. The only complexity is deciding on the length of the steps. If step lengths are chosen too short the journey to the MLE is very slow but if the step lengths are too long then the process can become trapped by the topography of the likelihood function. For this reason the algorithm will require the capacity to adjust the step length when a proposed movement is rejected. As there

is no random element in the algorithm the movement is largely deterministic (subject to numerical precision) with the result being determined by the choice of starting point. If you do not require any information other than the MLE then this method can be very effective.

When first presenting the Markov switching model Hamilton used a hill climbing algorithm to maximise the likelihood of a parameter set. In particular he applied a Davidson-Fletcher-Powell routine. He found the models were "...relatively robust with respect to a broad range of start-up values". As we were less concerned by the efficiency of the model fitting we applied a somewhat cruder routine which, nevertheless, used the hill climbing principle. We found the same robustness, not just with the US GNP data but with data sets generated for the purpose of testing convergence.

3.3 Markov Chain Monte Carlo

We are sometimes required to make inference about parameters for a given set of data where the model has a complex, high dimensional probability distribution. We do not always find these distributions easy to work with, or even defined in some cases. We may be faced, for instance, with hierarchical models, mixture models or an incomplete set of data. In these situations we need a simple and flexible method which, nevertheless has the power to provide answers to a wide range of different problems. The rapid advance in processor speeds and storage capacity in the last few decades has made it practical to use iterative methods evaluated through simple computer programs. One way we can achieve this is by utilising the Monte Carlo principle. According to this we can learn anything we want about a variable by sampling its probability density function many times. This approach was first anticipated in the first half of the 20^{th} century by Metropolis & Ulam (1949).

3.3.1 The MCMC Algorithm

In Monte Carlo integration we draw samples from the required distribution and form sample averages to approximate expectations. The principle of the Markov Chain Monte Carlo (MCMC) is to obtain these samples using a specially constructed Markov chain. There are many special cases of MCMC with distinct ways of constructing these chains, each having its own strengths and weaknesses. In the course of our work we used one particular version of MCMC. This was the Metropolis algorithm called random walk Metropolis. Using this method we initiate a random walk around the model's parameter space in such a way as to concentrate on the high probability areas. From the journey of the process we can recover the posterior distribution without any recourse to asymptotically valid approximations about the shape of the likelihood. At some points we utilised the Single Component Metropolis-Hastings that updated only one parameter at a time while at other points we allowed the process to change several parameters simultaneously.

The process starts by choosing the initial point. From here new parameters are chosen at random by drawing a value from the proposal density. The proposal density of a point

is some distribution (its actual form is not too important though under the Metropolis algorithm it must be symmetric) that may depend on the previous value of the parameter. The method we chose was to use independent Normal distributions to update each parameter seperately. As a result the acceptance probability was then calculated by taking the ratio of the likelihood values of the model at the current and the proposed point. This will be set to 1 if the move increases likelihood. The new point is then accepted with this probability and rejected otherwise. It is easy to see that this process is a Markov Chain because the choice of proposed point depends only upon the current location. By this series of proposed and accepted points the process will 'walk' around the parameter space concentrating on those areas with high likelihood. In fact once the process has settled down to the areas of high likelihood the process acquires the useful property of having as its stationary distribution the multivariate density we are wishing to study. Once in this state we can draw samples from this Markov Chain and use them to estimate any function of the distribution. We can, for instance, easily estimate the marginal distribution of the individual parameters to find their expected values and confidence intervals.

3.3.2 MCMC in Practice

There are potential problems when using MCMC methods. These are similar in nature to those faced when using hill climbing methods. If the proposal density is too narrow the Markov chain will not explore the distribution very quickly, leading to problem of slow 'mixing'. On the other hand if the proposal density is too wide then many of the proposed points will be simply rejected giving us fewer points. The points the process visits make up the sample that we use to represent the distribution and the more complex the shape of the distribution the more we require. Another problem shared with hill climbing methods concerns the unknown nature of the density function we are exploring. As we do not know its actual shape we cannot be sure that the process fully explores all key areas of the distribution. If it were to ignore certain alternative modes of a multimodal density we would not necessarily be aware of it as the process may converge to a different portion of the distribution. We can run multiple chains from different starting points but this will not always solve the problem as they may also converge incorrectly. As we increase the number of chains we are using we are slowing the process down and collecting progressively smaller samples to make our inference from. So although multiple chains can be a useful tool it is difficult to know how many would be enough. We can try to broaden the support of the proposal function to allow the process to 'jump' between alternative modes but this will increase the number of rejected moves which may reduce the rate of mixing. In more complex cases it may be necessary first to explore the rough characteristics of the multivariate distribution to identify potential problems before we start inference in earnest. There are also many 'fixes' that we can apply, but none of these will fully overcome the weaknesses of the system when dealing with potentially multi-modal densities.

Despite these concerns the method is incredibly simple to apply although to ensure

convergence and obtain good estimates of confidence intervals, for instance, can be quite time consuming. The algorithm also has the advantage that it can be easily adapted to work as a random walking hill climber by only accepting points that increase likelihood making a seperate algorithm for hill climbing unecessary. We used this random walking version for finding MLEs and also in its true form for obtaining marginal distributions of the parameter estimates (and hence confidence intervals) when fitting models.

Hansen (1992) applied Monte Carlo simulation methods to Hamilton's model and discussed the suitability of those methods to this kind of problem. Although we are concerned with MCMC methods some of his criticisms are of interest to us. He points out that the likelihood function of Hamilton's model is extremely ill behaved with numerous local optima and the question of which of these are found is heavily dependent on the starting point of the chain. He advises cautious interpretation of the results.

3.4 The EM algorithm

Dempster & Rubin (1977) describe their EM algorithm as "a broadly applicable algorithm for computing maximum likelihood estimates from incomplete data". The principle behind the algorithm was not entirely new, Hartley (1958) had applied similar methods, while some of the theory had been presented by Orchard & Woodbury (1972) and Sundberg (1974). When working with Hidden Markov Models it is sometimes known as the Baum-Welch algorithm, in reference to the ideas proposed in Baum *et al.* (1970). It is on this, specialised, version we will be concentrating. The purpose of the algorithm is to ensure better performance of Maximum Likelihood Esimation when working with complex likelihood functions that cannot be easily maximised numerically.

3.4.1 The EM Algorithm

Application of the algorithm involves two steps, the 'E' (or Expectation step) and the 'M' (or Maximisation) step. The term 'algorithm' may be misleading since no single procedure of steps can be specified and that "detailed applications vary widely in complexity and feasability". The method, as it concerns us, works in the following way. We have a set of observed data $\{y(n)\}$ and propose a set of starting values for the parameters Θ_1 describing the hidden Markov process. The filter proposed by Hamilton is used to collect the complete set of smoothed probabilities using our current parameter estimates Θ_1. These are then used to reweight the observed data. With this weighted data we perform a series of simple sample statistics of OLS regressions. These will give us new estimates of the parameters controlling the movement of the process, Θ_2. These new estimates are, in turn, used to produce more smoothed probabilities and so on. Each such calculation of probabilities and reweighting of the data can be shown to increase the value of the likelihood function. A fuller explanation of the process for this case, and others, can be found in Hamilton (1990).

3.4.2 The EM Algorithm in Practice

The EM algorithm has been successfully applied to a wide range of different problems. The area of hidden Markov models is just one of these. The algorithm is generally well regarded and this is in no small part due to the two monotonicity properties. The convergence to the MLE is monotone and the value of the likelihood function increases with each iteration. Hamilton concedes that he found the algorithm robust with respect to starting values and that it offered a "vast improvement in efficiency" over earlier methods.

There are two potential problems with the EM algorithm generally. The first is that each application requires a different specification and so small changes to a model will result in changes to the required form of the algorithm. The second criticism of the algorithm focuses on the rate of convergence. While the initial steps of the algorithm are well chosen the rate of convergence tends to slow as it approaches the MLE. These factors make the EM algorithm a powerful tool, but where the specification of the models changes frequently or the likely location of the MLE is already known simpler methods may be more desirable.

3.5 The MAP algorithm

Kim (1993) suggested an alternative two-stage estimation method. He bases this method on the Maximum A Posteriori decision rule.

The sequence of inferred states of the hidden Markov chain $\{s(n)\}$ can be obtained through the use of the MAP decision rule. Kim shows that it is a solution path of Bellman's Dynamic Programming algorithm. Monte Carlo experiments peformed by Kim demonstrate that this method will outperform Hamilton's method when measured against the misclassification of states. He also found that the dating of the business cycles of US-GNP generated by the MAP approach were closer to those published by the NBER that when using Hamilton's method.

3.6 Summary and Conclusions

We have looked at the main alternatives for optimising the likelihood of a model. At different times we will require different things from this fitting process but one thing remains true. As a whole in our research we have been less interested in the efficiency of the fitting methods and more interested in their robustness. We have also frequently had a rough idea of the location of the MLE and the time to explore the likelihood space.

The hill climbers offer the simplest implementation but are relatively poor when it comes to efficiency, being very prone to missing the global maxima.

The EM algorithm is extremely effective at finding the higher likelihood region of the parameter space but quite slow to converge.

MCMC methods can be extremely useful for obtaining information such as the shape of the likelihood space close to the MLE. This makes them invaluable for finding the standard

error or confidence intervals of estimates.

We have not implemented the MAP method in order to assess its suitability and only include mention of it here for sake of completeness.

Given that our requirement has been to test the difference between similar models rather than to analyse new time series we tend to favour the simpler methods. It is necessary for us to constantly adapt and develop our routines making any method with a comparatively high implementation time (such as the EM algorithm) undesirable. The approximate location of the solutions we seek are usually already known to us, through the results of the neighbouring models and we will usually start the optimisation algorithms from very close to their final solutions. As a result we will tend to work with very simple hill climbing algorithms, but perform repeated optimisations from different starting positions.

Chapter 4

The Filtered Markov Process

In this chapter we introduce a continuous time Markov process incorporating gradual switching dynamics, and discuss its interesting points. We also present theoretical expressions for the transition probabilities, distribution functions and moments.

4.1 Gradual Switching History

Hamilton's model represented a significant new chapter in the development of time series models, resulting from the challenge to address the existence of certain types of non-linearity. It is an example of one of the many steps that allow the subject area to progress. Each progression represents the availability of more sophisticated or flexible models. There is no end to this process, with the weaknesses of these new models becoming the research areas of the future. One such subject area is that of Markov Switching Regression models, like Hamilton's model. Within this subject area, one such area for future research is the assumption of instantaneous transition from regime to regime. Whether or not this assumption is realistic will depend upon what behaviour the model is being used to describe. In Hamilton's case the model is applied to the growth rate of US postwar GDP. In this kind of macroeconomic modelling is is unlikely that any regime transition will be immediate. The growth rate of an economy depends on the behaviour and output of countless businesses and industries. Whether or not we should consider this 'inertia' depends upon the relationship between the length of this transition period and the interval between observations. Despite this kind of structural model being fifteen years old surprisingly little work has been done on gradual switching models.

There are several decisions facing anyone wanting to model a gradually switching process:

- Is the form of the transition symmetric?
- How long should a transition take?
- What form should a transition take?

The answers to each of these questions will then have significant implications for the model and what tools are open to us. For instance if the transition interval does not have clear start or end points this raises real problems for the sequential Bayesian method of fitting used by Hamilton. It also raises the possibility of the end of one transition overlapping with the beginning of the next. How to handle this will be an important question that will need a clear answer in order for us to construct our process. The duration of a switch is intimately connected with the form the switch takes. Several, quite different, choices have been made by different authors to find a function that can provide a smooth transition from one state to another.

4.1.1 Existing Research

In Bacon & Watts (1971) we are presented with what appears to be a clear case of a straight line abruptly changing gradient mid-series. They felt that the change might not be so sudden as it appeared and so introduced a transition function, dependent upon the distance from the switch point. They were fairly unconcerned by the precise choice of function, from a list of suitable ones, arguing that this was not crucial given the similarities of shape. They fitted this model to data collected on the behaviour of stagnant surface layer height in a controlled flow of water and conclusively rejected the assumption of instantaneous transition. The advantages of the kind of data they were working with was good prior knowledge about the behaviour in the two regimes.

In Ohtani *et al.* (1990) a transition function is also used to manage the switch from one model to another but the author opted for a more complex arrangement. An assumption was made that the switch was symmetric. Instead of defining the transition around a clear switch point they defined a transition interval, measured by its start and end points. This had the advantage of restricting the effects of a switch to within certain bounds. This more general approach also allowed a wide range of different transition patterns with differing transition rates and symmetry. The exact order of the process is determined by the AIC criteria. In applying the model to an import demand function for Japan before and after the second oil crisis (data from 1975 to 1986) they claim only that the gradual switching model is 'plausible'.

Prompted by an inability to fit models to data for labour productivity growth Varoufakis & Sapsford (1991) tried developing their regression model. They had originally tried one strucutural break, then two, but were still unable to get a really convincing model. Accepting that the data displayed behaviour their models were unable to capture they introduced a gradual switch between the different regression models. The transition function they used was based on the Normal CDF. This development of their model resulted in a huge jump (upwards) for the likelihood function. They concluded that the switch was not immediate, suggesting that the series took eight years to fully settle down to the new regime after a switch.

In another example of a gradual switching model Konno & Fukushige (2002) focused on

bilateral import functions, and the effects on them of the Canada-US Free Trade Agreement (CUSTA). The particular set of data is concerned with the introduction of the CUSTA and the way in which the import functions adjusted during its implementation. An adaptation of Ohtani's model was used but the transition function was changed to one of arbitrary order. They conclude that the transition on the US side is almost immediate whereas on the Canadian side the adjustment is slow, taking almost the full 10 years the actual structural transition took.

It is worth noting, before we move on, that this is not the only way to model a gradual switch. Also available are models which incorporate parameter variation. These kinds of models, however, should probably not be considered as a simple extension of the Markov Switching Model.

4.1.2 Shortcomings of the Research

The study of gradual switching models is still in its infancy. Despite the first of these papers being 30 years old we have still not progressed much beyond a mixture of regression models. The range of papers available on the subject appears impressive but masks a major restriction. In almost all of the cases the data that are being analysed are known (or at least strongly believed) to contain one switch, and one switch only. As such the problem faced by the authors is to somehow reconcile the difference between the pre-switch and post-switch data. There do not appear to be any serious attempts to establish a model that incorporates gradual switching as an inherent property of its structure. It is likely that this is due to the effectiveness of the existing models and the limited advantages on offer for the majority of cases. Not least there is the problem of collecting sufficient evidence for inference. Even if the transition affects many observations after the switch, if the switching is infrequent there will be little information providing inference on the switching period. This could easily be swamped with only a moderate amount of noise. It is inevitable, however, that there would be challenges to overcome as in any open area of research. These models are worth pursuing for their own sake regardless of their possibly limited appeal.

4.2 Introducing the model

The path we have chosen in this research is to explore this idea of a gradual regime switching Markov model. To do this we must question an implicit assumption, namely the instantaneous nature of the change from one regime to another. If we assume that this change is not immediate we require a model that will allow the series to adjust, slowly moving from the level associated with the previous regime to that of the current regime. In particular it would be interesting for us to consider the possibility of a gradual switching model driven by a continuous time Markov process. We shall call this hidden driving process the signal (or driving signal). For the moment we shall consider the driving process to have two levels,

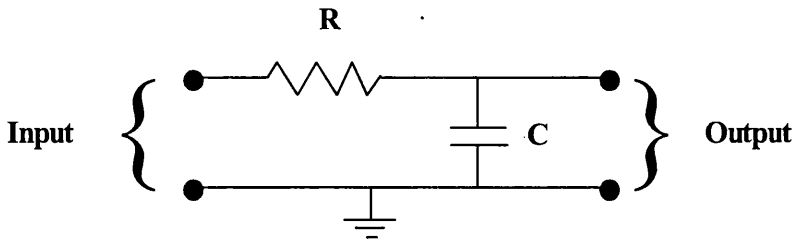

Figure 4.1: A representation of a simple RC network (containing a Resistor and Capacitor) with Input and Output terminals marked.

although we can generalise to include more if necessary. It is very important at this stage to make our notation clear when referring to the signal.

Definition 1 *When introducing Hamilton's model we used $S(n)$ to refer to the index of the current level of the signal where $S(n)$ could take the values $\{0, 1\}$. We then used l_0 and l_1 to refer to the level of the growth rate if $S(n) = 0$ and $S(n) = 1$ respectively. At times we will have to consider a more general case with $N_{(R)}$ regimes. In order to keep the notation simple, for $S(n) = i$, when we refer to the* **regime** *of the driving signal we shall always mean i, and when we refer to the* **level** *of the driving signal we always mean l_i. In some circumstances we may refer to the* **state** *of the signal. This will be taken to mean the combined state comprising the regime and level of the signal, $(S(n), l_{S(n)})$.*

To recognise the fact we are now considering a continuous time process, rather than a discrete one as in Hamilton, we shall use t to denote the time rather than n. The observable process is given by $X(t)$. We define the process by setting the rate of change of the observed signal $X(t)$ to be inversely proportional to the distance of $X(t)$ from the level of the driving signal $l_{S(t)}$ such that this separation will reduce with time. This will give us a behaviour pattern governed by the negative exponential distribution.

There already exists an analogous situation in physics where an electrical circuit consisting of a resistor and a capacitor (an RC network) adapts to a change in voltage. In Figure 4.1 we show an example of an RC network. The input voltage is applied across the 'Input' Terminals. The Output Signal is measured across the 'Output' Terminals. The principle behind the circuit is quite simple. When voltage is first applied the capacitor (C) is uncharged and does not impede the current flow towards the earth. As the capacitor charges it resists the flow of current leading to a greater voltage across the Output Terminals. Eventually, if the input voltage remains constant, the Output signal will seem to reach an equilibrium as the Input voltage becomes balanced by the Voltage across the capacitor. If the Input voltage is removed the capacitor now discharges itself slowly tending towards a state of complete discharge. The rates of charging and discharging are identical

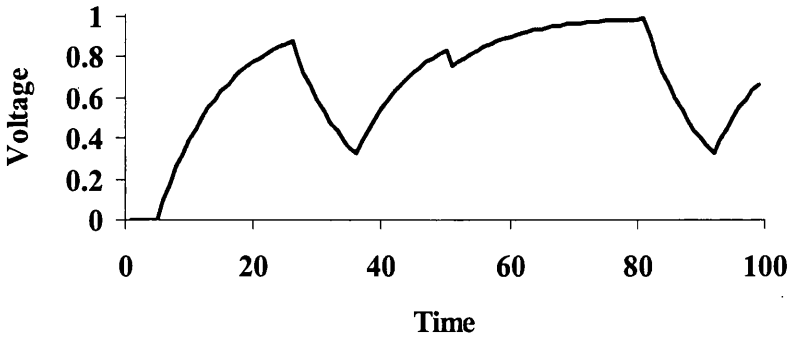

Figure 4.2: An example of the observed output voltage for the RC circuit if τ is large. The input voltage is a two-level Markov process.

(in magnitude), both being proportional to the difference between the Input Voltage and the Voltage across the capacitor. As a result we get a certain symmetry of style in the up and down movements of the Output Signal. This rate of dis/charging is affected by the properties of the Resistor and Capacity with the scaling constant τ given by.

$$\tau = RC$$

We shall often refer to this as the Time Scaling Factor. The greater the product of the capacitance and resistance the slower the process moves.

In Figure 4.2, we have given an example of a typical process. The Input signal can only takes voltages $l_0 = 0$ and $l_1 = 1$, and periodically changes from one level to another. When the voltage is 0 the process tends towards 0 ($X(t) \to 0$) and when the voltage is 1, we see $X(t) \to 1$. In the second example shown in Figure 4.3, a lower value of τ leads to a faster convergence between the Input and Output voltages.

We are not intending to work with electrical circuits but this method of visualising the process is an invaluable tool to understanding its dynamics. This idea is not a new one to the science of signal processing. Wonham (1959), Fitzhugh (1983) and Pawula (1986) have already worked with this type of process deriving transition probabilities and conditional distributions. We shall reproduce some of their work here, adjusting their notation to match ours. They focus on the two-regime case and mostly on a symmetric driving signal. This type of process also falls within the broad class of piecewise-determinstic Markov processes introduced and explored in Davis (1984) although we shall not be using any of his results here. The multi-state Input signal has been discussed in Jalali (2003c). Some of the results he obtained are discussed in the next Section.

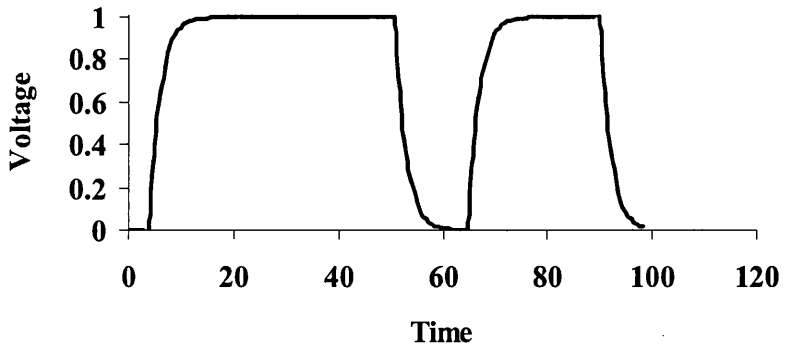

Figure 4.3: An example of the observed output voltage for the RC circuit if τ is small. The input voltage is a two-level Markov process.

4.3 Distributions of the Process

Now that we have proposed a stochastic process we should explore it theoretically. The analagy of the RC network can be used to obtain the fundamental rules of movement. From these we can find expression for the distributions and moments. The following results were proposed in Jalali (2003a).

4.3.1 Movement of the Process

Let us take a circuit, like that in Figure 4.1, with a resistor (of resistance R) and a capacitor (of Capacitance C). The voltage across the Input terminals is a constant V and that across the output terminals is X and a current i is flowing. The charge on the capacitor C is given by Q. Electrical charge accumulates at the capacitor at a rate of $i.dt$ and produces a potential across the capacitor (and output terminals) of Q/C. The current i can be written as

$$i = \frac{dQ}{dt} = C\frac{dX}{dt}$$

But also

$$i = \frac{V - X}{R}$$

Equating these two expressions we get

$$\frac{dX}{dt} = \frac{V - X}{RC}$$

Solving this we find an expression for the output voltage at time t, denoted by $X(t)$

$$X(t) = V - c\exp\left(-\frac{t}{RC}\right) \quad \text{where } c \text{ is a constant of integration}$$

Incorporating initial conditions ($X(0) = x(0)$), we can rewrite this

$$X(t) = V - (V - x(0))\exp\left(-\frac{t}{RC}\right)$$

Without loss of generality we can take a unit time to be RC seconds, in which case

$$X(t) = V - (V - x(0))e^{-t}$$

We can see clearly from this that if the Input voltage V is kept constant the Output voltage X_t will tend towards V regardless of which is greater. If we relax the condition that V be constant and instead allow it to alternate periodically between values 0 and 1, then once $0 < X(t_1) < 1$ then $0 < X(t_2) < 1$ for $t_2 > t_1$. As a result the stationary distribution will only take values for X in the interval $0 < x < 1$.

We now go further and replace V with our continuous time Markov process $S(t)$, with regimes i that correspond to levels of voltage l_i. The transition intensities of $S(t)$ while in regime 0 and regime 1 are given by α and β respectively. We have introduced a time-scaling factor RC and we shall represent this by the constant τ. In general it is often easier for us to rescale our measurement of time to ensure that $\tau = 1$. In this standardised case we shall replace our transition intensities with their rescaled counterparts a and b, where

$$\begin{aligned} a &= \alpha\tau \\ b &= \beta\tau \end{aligned}$$

The level of the voltage at time t will be given by $X(t)$, which will be controlled by the Markov process $S(t)$ through $l_{S(t)}$ (which is defined by $l_{S(t)} = l_i$ when $S(t) = i$) and the following movement rule

$$X(t+dt) = l_{S(t)} - (l_{S(t)} - x(t))\exp\left(-\frac{dt}{\tau}\right) \tag{4.1}$$

Where the regime of $S(t)$ (and hence the level of $l_{S(t)}$) is constant during the interval $[t, t+dt)$. We can now drop the analagy of the RC circuit as we have Markov process, such as we would find if we viewed the driving signal $S(t)$ through a low-pass filter. It is from this the name Filtered Markov process derives.

4.3.2 The Stationary Distribution

We shall use the notation defined above and add several new terms. We shall take $x(0)$ to be equal to the value of the process $X(t)$ at time $t = 0$. Given the movement rule (4.1) we know that eventually, with probability 1, we must have $0 < X(t) < 1$. However we shall

insist on the initial restriction $0 < X(0) < 1$. To simplify notation in this section we shall take $l_0 = 0$, $l_1 = 1$.

We shall denote by $F(x)$ the stationary cumulative distribution function (CDF). The CDFs conditional on $S(t) = i$ are denoted by $F_i(x)$. Since $S(t)$ is a two-regime Markov process we can express the transition probabilities in the form of the infinitessimal transition matrix. As we will be working with the standardised process this will be

$$\begin{pmatrix} -a & a \\ b & -b \end{pmatrix}$$

Note that this will have equilibrium probabilities

$$\Pr[S(t) = 0] = \frac{b}{a+b} \tag{4.2}$$

$$\Pr[S(t) = 1] = \frac{a}{a+b} \tag{4.3}$$

Using these equilibrium probabilities (4.2) and (4.3) we can say that

$$\begin{aligned} \Pr[S(t) = 0 \ \& \ X(t) \leqslant x] &= \frac{b}{a+b} F_0(x) \\ \Pr[S(t) = 1 \ \& \ X(t) \leqslant x] &= \frac{a}{a+b} F_1(x) \end{aligned}$$

Given the fact that this solution is stationary we can also say that

$$\begin{aligned} \Pr[S(t-dt) = 0 \ \& \ X(t-dt) \leqslant x^*] &= \frac{b}{a+b} F_0(x^*) \\ \Pr[S(t-dt) = 1 \ \& \ X(t-dt) \leqslant x^*] &= \frac{a}{a+b} F_1(x^*) \end{aligned}$$

Now if no transitions have occured during the interval $[t-dt, t)$, then

$$\{S(t) = 0 \ \& \ X(t) \leqslant x\} \text{ is equivalent to } \{S(t-dt) = 0 \ \& \ X(t-dt) \leqslant xe^{-(-dt)}\}$$

Whereas if one transition did take place during the interval $[t-dt, t)$ then

$$\{S(t) = 0 \ \& \ X(t) \leqslant x\} \text{ is equivalent to } \{S(t-dt) = 1 \ \& \ X(t-dt) \leqslant x + O(dt)\}$$

The probability of more than one switch inside $[t-dt, t)$ is $o(dt)$ and we can safely ignore it.

Hence we can write

$$\begin{aligned} \frac{b}{a+b} F_0(x) &= (1 - adt) \frac{b}{a+b} F_0(x + xdt) + \frac{a}{a+b} F_1(x + O(dt)) bdt + o(dt) \\ \frac{a}{a+b} F_1(x) &= (1 - bdt) \frac{a}{a+b} F_1(x - (1-x)dt) + \frac{b}{a+b} F_0(x + O(dt)) adt + o(dt) \end{aligned}$$

or

$$F_0(x) - F_0(x + xdt) = -aF_0(x + xdt)dt + aF_1(x + O(dt))dt + o(dt) \quad (4.4)$$

$$F_1(x) - F_1(x - (1-x)dt) = -bdtF_1(x - (1-x)dt) + bF_0(x + O(dt)) + o(dt) \quad (4.5)$$

We can obtain the derivative of F_0 by dividing (4.4) by xdt and letting $xdt \to \infty$. This is given in (4.6). Similarly we can obtain (4.7) from (4.5).

$$-xf_0(x) = a(F_1(x) - F_0(x)) \quad (4.6)$$

$$(1-x)f_1(x) = -b(F_1(x) - F_0(x)) \quad (4.7)$$

From these we also obtain

$$f_1(x) = \frac{bx}{a(1-x)} f_0(x) \quad (4.8)$$

To solve these equations we find the first derivative of both sides of (4.6) and (4.7) with respect to x and rearrange terms

$$-x\frac{d}{dx}f_0(x) = -(a-1)f_0(x) + af_1(x)$$

$$(1-x)\frac{d}{dx}f_1(x) = bf_0(x) - (b-1)f_1(x)$$

Then we substitute (4.8) to obtain

$$\frac{d}{dx}f_0(x) = \frac{(a-1)}{x} f_0(x) - \frac{b}{1-x} f_0(x) \quad (4.9)$$

Or

$$\left(\frac{1}{f_0(x)}\right)\frac{d}{dx}f_0(x) = \frac{(a-1)}{x} - \frac{b}{1-x} \quad (4.10)$$

Equation (4.9) can be easily integrated to give us

$$f_0(x) = acx^{a-1}(1-x)^b$$

from which it follows (due to (4.8)) that

$$f_1(x) = bcx^a(1-x)^{b-1}$$

The value of c can be obtained from the properties of a PDF since

$$\int_0^1 f_0(x)dv = 1$$

giving us

$$c = \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b+1)}$$

So the conditional PDFs of the process can be written

$$f_0(x) = \frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)} x^{a-1}(1-x)^b \tag{4.11}$$

$$f_1(x) = \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^{a}(1-x)^{b-1} \tag{4.12}$$

The unconditional PDF of the output is then

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \tag{4.13}$$

A quick application of Bayes Theorem to (4.11) and (4.13) gives us the probability of being in either regime based only on an observed value of the process. This probability is given by

$$\Pr[S(t)=0 \mid X(t)=x] = \lim_{dx\to 0} \frac{\Pr[S(t)=0 \;\&\; \{x \leqslant X(t) < x+dx\}]}{\Pr[x \leqslant X(t) < x+dx]} \tag{4.14}$$

$$\begin{aligned} &= \frac{\left(\frac{b}{a+b}\right) f_0(x)}{f(x)} \\ &= \frac{\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) x^{a-1}(1-x)^b}{\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) x^{a-1}(1-x)^{b-1}} \\ &= 1-x \end{aligned}$$

Similarly we find

$$\Pr[S(t)=1 \mid X(t)=x] = x \tag{4.15}$$

It is interesting to note that the conditional and unconditional distributions are all Beta distributions. For a standardised process ($\tau = 1$) we have

$$\begin{aligned} f_0(x) &\sim Beta(a, b+1) \\ f_1(x) &\sim Beta(a+1, b) \\ f(x) &\sim Beta(a, b) \end{aligned}$$

For the case where $\tau \neq 1$ we simply transform $a = \alpha\tau$ and $b = \beta\tau$.

$$\begin{aligned} f_0(x) &\sim Beta(\alpha\tau, \beta\tau+1) \\ f_1(x) &\sim Beta(\alpha\tau+1, \beta\tau) \\ f(x) &\sim Beta(\alpha\tau, \beta\tau) \end{aligned}$$

These solutions presume that we are dealing with levels $l_0 = 0$ and $l_1 = 1$. It is straight-

forward to adapt these distributions to a more general case with levels l_i by scaling and shifting.

4.3.3 The Transition Probabilities

This type of process has been studied before, mostly in the field of signal processing in engineering. In earlier works Wonham (1959) and Pawula (1970) studied the transition probabilities of the Symmetric Filtered Random Telegraph signal. Where we refer to a symmetric signal we mean one in which the rate of switching is independent of regime. In a symmetric process the stationary probability of occupying each regime is identical. The symmetric Filtered Random Telegraph signal is the particular case of this Filtered Markov (FM) process where $\alpha = \beta$. More recently Fitzhugh (1983) produced a comprehensive study of Asymmetric Filtered random Telegraph signals. He gave expressions for the various density functions, the Fokker-Planck-Kolmogorov equations and their transient solution along with an appraisal of its suitability for application to Channel Signal Analysis. Fitzhugh also extended the idea to the case of two arbitrary levels (rather than the 0 and 1 we had been working with previously). We now give the expression he obtained for the transition function of the process.

The notation we shall use is as follows.

$\{y(t)\}$ represent the observations in our series.

$\{s(t)\}$ represent the regimes of the driving signal.

l_0, l_1 are the levels of the driving signal (assumed $l_0 < l_1$).

As the transition-probability function is time invariant we will consider only the transition between times $t = 0$ and $t = t$.

The variables x and y refers to the values of the process at time $t = 0$ and $t = t$ respectively.

The time scaling factor we ignored (without loss of generality) in the last section is given the constant τ. .

The transition-probability function between times $t = 0$ and $t = t$ is defined as

$$r_{ij}(x, y, t)dy = \Pr[s(t) = j \ \& \ y(t) \in (y, y + dy) \mid s(0) = i \ \& \ y(0) = x]$$

To keep the expressions manageable we shall introduce some specific notation to this

section.

$$\begin{aligned}
\eta_{0i} &= \frac{|x-l_i|}{(l_1-l_0)} \\
\eta_{1i} &= \frac{|y-l_i|}{(l_1-l_0)} \\
\lambda &= e^{-\frac{t}{\tau}} \\
m_0(x,t) &= l_0+\lambda(x-l_0) \\
m_1(x,t) &= l_1-\lambda(l_1-x) \\
z &= \frac{(\lambda(l_1-x)-(l_1-y))(\lambda(x-l_0)-(y-l_0))}{(\lambda(x-l_0)+(l_1-y))(\lambda(l_1-x)+(y-l_0))}
\end{aligned}$$

And as we have already used F to represent the CDF of $X(t)$ then we shall take H to represent the Hypergeometric Function. Two terms depend on this

$$\begin{aligned}
H_1(z) &= H(1-\beta\tau,-\alpha\tau;1;z) \\
H_2(z) &= H(2-\beta\tau,1-\alpha\tau;2;z)
\end{aligned}$$

Fitzhugh derived the forward Kolmogorov equations for a process discrete in level and continuous is time and the forward Kolmogorov (Fokker-Planck) equations for a process continuous in level and time. By combining these he obtained the Fokker-Planck-Kolmogorov (FPK) equations for r_{0j}.

$$\begin{aligned}
\frac{\partial}{\partial t}r_{00}(x,y,t) &= \frac{1}{\tau}\frac{\partial}{\partial y}\left[(y-l_0)r_{00}(x,y,t)\right]-\alpha r_{00}(x,y,t)+\beta r_{01}(x,y,t) \\
\frac{\partial}{\partial t}r_{01}(x,y,t) &= \frac{1}{\tau}\frac{\partial}{\partial y}\left[(y-l_1)r_{01}(x,y,t)\right]-\beta r_{01}(x,y,t)+\alpha r_{00}(x,y,t)
\end{aligned}$$

By Riemann's method these can be solved exactly. The derivation is rather lengthy. Given the initial conditions

$$r_{ij}(x,y,0)=\delta_{ij}\delta(y-x)$$

He obtains the solutions

$$\begin{aligned}
r_{00}(x,y,t) &= \lambda^{\alpha\tau}\delta(y-m_0(x,t))+\frac{\alpha\tau}{(l_1-l_0)}(\lambda\eta_{00}+\eta_{11})^{\beta\tau-1}(\lambda\eta_{01}+\eta_{10})^{\alpha\tau-1}(\lambda\eta_{01}-\eta_{11}) \\
&\quad\times\left\{\frac{\lambda(1-\beta\tau)}{(\lambda\eta_{00}+\eta_{11})(\lambda\eta_{01}+\eta_{10})}H_2(z)-H_1(z)\right\} \qquad (4.16) \\
r_{01}(x,y,t) &= \frac{\alpha\tau}{(l_1-l_0)}(\lambda\eta_{00}+\eta_{11})^{\beta\tau-1}(\lambda\eta_{01}+\eta_{10})^{\alpha\tau}H_1(z) \qquad (4.17)
\end{aligned}$$

The two other solutions $r_{10}(x,y,t)$ and $r_{11}(x,y,t)$ can be obtained by symmetry.

$$\begin{aligned} r_{11}(x,y,t) &= \lambda^{\beta\tau}\delta(y-m_1(x,t)) + \frac{\beta\tau}{(l_1-l_0)}(\lambda\eta_{01}+\eta_{10})^{\alpha\tau-1}(\lambda\eta_{00}+\eta_{11})^{\beta\tau-1}(\lambda\eta_{00}-\eta_{10}) \\ &\quad \times\left\{\frac{\lambda(1-\alpha\tau)}{(\lambda\eta_{01}+\eta_{10})(\lambda\eta_{00}+\eta_{11})}H_2(z) - H_1(z)\right\} \\ r_{10}(x,y,t) &= \frac{\beta\tau}{(l_1-l_0)}(\lambda\eta_{01}+\eta_{10})^{\alpha\tau-1}(\lambda\eta_{00}+\eta_{11})^{\beta\tau}H_1(z) \end{aligned}$$

When working with the standardised case, where $l_0=0$, $l_1=1$ and $\tau=1$ the equations 4.16 and 4.17 simplify to

$$\begin{aligned} r_{00}(x,y,t) &= e^{-at}\delta(y-xe^{-t}) + a(xe^{-t}+1-y)^{b-1}(e^{-t}(1-x)+y)^{a-1}(e^{-t}(1-x)+(1-y)) \\ &\quad \times\left\{\frac{e^{-t}(1-b)}{(xe^{-t}+1-y)(e^{-t}(1-x)+y)}H_2(z) - H_1(z)\right\} \\ r_{01}(x,y,t) &= a((xe^{-t}+1-y)^{b-1}(e^{-t}(1-x)+y)^{a}H_1(z) \end{aligned}$$

where

$$z = \frac{(e^{-t}(1-x)-(1-y))(xe^{-t}-y)}{(xe^{-t}+1-y)(e^{-t}(1-x)+y)}$$

He also provides the stationary conditional distributions as

$$\begin{aligned} f_0(y) &= \frac{\Gamma(\alpha\tau+\beta\tau+1)}{\Gamma(\alpha\tau)\Gamma(\beta\tau+1)}\cdot\frac{1}{(l_1-l_0)}\cdot\left(\frac{|y-l_0|}{l_1-l_0}\right)^{\alpha\tau-1}\left(\frac{|y-l_1|}{l_1-l_0}\right)^{\beta\tau} \\ f_1(y) &= \frac{\Gamma(\alpha\tau+\beta\tau+1)}{\Gamma(\alpha\tau+1)\Gamma(\beta\tau)}\cdot\frac{1}{(l_1-l_0)}\cdot\left(\frac{|y-l_0|}{l_1-l_0}\right)^{\alpha\tau}\left(\frac{|y-l_1|}{l_1-l_0}\right)^{\beta\tau-1} \end{aligned}$$

The unconditional output is then easily found

$$f(y) = \frac{\Gamma(\alpha\tau+\beta\tau)}{\Gamma(\alpha\tau)\Gamma(\beta\tau)}\cdot\frac{1}{(l_1-l_0)}\cdot\left(\frac{|y-l_0|}{l_1-l_0}\right)^{\alpha\tau-1}\left(\frac{|y-l_1|}{l_1-l_0}\right)^{\beta\tau-1}$$

Which reduce to the distributions (4.11), (4.12) and (4.13) obtained in the previous section when $l_0=0$, $l_1=1$ and $\tau=1$.

4.3.4 Moments of the Process

A powerful method for determining the moments of the, more general, $N_{(R)}$-regime process can be found in Jalali (2003c). We shall outline the method and show how he uses it to obtain conditional moments, transient moments and the Covariance of the process. The author considers as an input signal a Markov process with $N_{(R)}$ regimes.

The Transient Moments

The notation we shall use here will require the definition of new terms.

$\{S(t)\}$ is the regime of the driving signal, a $N_{(R)}$-regime continuous time Markov process.
Q is the matrix of transition intensities for $S(t)$ with entries q_{ij}.
P is the regime transition matrix, given by $P(t) = e^{\mathbf{Q}t}$. The entries are $p_{ij}(t)$.
$\{X(t)\}$ is the level of the process.
$\{l_i : i = 0, 1, 2, ..., N_{(R)} - 1\}$ are the levels of driving signal.

We also need to define three key conditional terms.

$$\begin{aligned} p_{ij}(t) &= \Pr[S(t) = j \mid S(0) = i] \\ X_i(t) &= X(t) \mid S(0) = i \\ S_i(t) &= S(t) \mid S(0) = i \\ X_{ij}(t) &= X(t) \mid S(0) = i, S(t) = j \end{aligned}$$

We are then able to define a new concept, the contribution to the expectation for a specified start and end regime (m_{ij}). For the first moment we have

$$m_{ij}(x,t) = p_{ij}(t)E[X_i(t) \mid X(0) = x, S(t) = j, S(0) = i]$$

But more generally we shall work with the r^{th} moment. So $m_{ij}(x,t;r)$ is defined by

$$m_{ij}(x,t;r) = p_{ij}(t)E[X_i^r(t) \mid X(0) = x, S(t) = j, S(0) = i]$$

In order to derive an expression for $m_{ij}(x,t;r)$ we need to introduce another concept. This is the value of the process conditional on its starting point, as well as starting regime. For the value of a process conditional on the initial value we use

$$X_i(x,t) = X(t) \mid X(0) = x \ \& \ S(0) = i$$

We find the following relationship between the $X(x,t)$ and $X(0,t)$

$$X(0,t).X(x,t) = X(0,t) + x\exp\left(-\frac{t}{\tau}\right) \tag{4.18}$$

Equation (4.18) is easily proved using induction.
If we have an interval $[0,t)$ containing no switches (changes of regime) then we know

that

$$\begin{aligned} X_i(x,t) &= l_i - (l_i - x)e^{-\frac{t}{\tau}} \\ &= l_i(1 - e^{-\frac{t}{\tau}}) + xe^{-\frac{t}{\tau}} \\ &= X_i(0,t) + xe^{-\frac{t}{\tau}} \end{aligned}$$

We then assume that (4.18) holds for an interval containing k switches and add another (final) switch at time $t = s$. Obviously then

$$X_i(x,s) = X_i(0,s) + xe^{-\frac{s}{\tau}}$$

If the process is in regime j during the final interval $[s,t)$ then

$$\begin{aligned} X_i(x,t) &= l_j - (l_j - X_i(x,s))\, e^{-\frac{t-s}{\tau}} \\ &= l_j - \left(l_j - X_i(0,s) - xe^{-\frac{s}{\tau}}\right) e^{-\frac{t-s}{\tau}} \\ &= l_j - (l_j - X_i(0,s))e^{-\frac{t-s}{\tau}} + xe^{-\frac{t}{\tau}} \\ &= X_i(0,t) + xe^{-\frac{t}{\tau}} \end{aligned}$$

Obviously then

$$E[X(x,t)] = E[X(0,t)] + x\exp\left(-\frac{t}{\tau}\right) \tag{4.19}$$

Now we can begin to construct an expression for the transient moments of the process. As we have the relationship 4.18 we can work with $X(0,t)$ rather than $X(x,t)$, and $m_{ij}(0,t;r)$ instead of $m_{ij}(x,t;r)$.

In order to proceed we need to find an expression for $m_{ij}(0,t)$.

We must consider the value of $m_{ij}(0,t)$ after a time interval $[0,t]$. We will condition on the value shortly after the start (at time dt). At this point we will find ourselves in one of two positions. Either the process will have switched (once) or not at all. If $X_{ij}(0,t)$ is the process $X(t)$ which started from $X(0) = x$ in regime $S(0) = i$ and ended in regime $S(t) = j$ then we wish to study $E[X_{ij}(0,t)]$

If no switching has occured by time dt then

$$X_i(0,dt) = X_{ii}(0,dt)$$

The probability of this is $(1+q_{ii}dt)$, where q_{ij} is defined earlier as an element of the transition rate matrix of $\{S(n)\}$, namely Q. The expectation will then be

$$E[X_{ij}(0,t)|S(dt) = i] = E[X_{ij}(X_{ii}(0,dt), t - dt)]$$

We also have the matching probability (with the $o(dt)$ term representing the possibility of multiple switches)

$$p_{ij}(t)|\{S(dt) = i\} = (1 + q_{ii}dt)p_{ij}(t - dt) + o(dt)$$

If however switching has occured by time dt then

$$X_i(0, dt) = X_{ik}(0, dt) < l_i(1 - e^{-dt})$$

This has probability $q_{ik}dt$ and the expectation will be

$$E[X_{ij}(0, t)|S(dt) = k] = E[X_{kj}(X_{ik}(0, dt), t - dt)]$$

With

$$p_{ij}(t)|\{S(dt) = k\} = q_{ik}dt.p_{kj}(t - dt) + o(dt)$$

Combining all these terms we obtain an expression for the probability moment $m_{ij}(0, t)$

$$\begin{aligned} p_{ij}(t)E[X_{ij}(0, t)] &= (1 + q_{ii}dt)p_{ij}(t - dt)E[X_{ij}(X_{ii}(0, dt), t - dt)] \\ &\quad + \sum_{k \neq i} q_{ik}dt.p_{kj}(t - dt)E[X_{kj}(X_{ik}(0, dt), t - dt)] + o(dt) \end{aligned} \tag{4.20}$$

If dt is small then we can safely make the following changes

$$\begin{aligned} m_{ij}(0, t) &= (1 + q_{ii}dt)m_{ij}(X_{ii}(0, dt), t - dt) + \sum_{k \neq i} q_{ik}dt.m_{kj}(0, t) + o(dt) \\ &= (1 + q_{ii}dt)m_{ij}(0, t - dt) + (1 + q_{ii}dt)l_i(1 - e^{-dt})e^{-(t-dt)}p_{ij}(t) \\ &\quad + \sum_{k \neq i} q_{ik}dt.m_{kj}(0, t) + o(dt) \end{aligned}$$

If we take $dt \to 0$ we can drop any terms of the form $o(dt)$

$$m_{ij}(0, t) - m_{ij}(0, t - dt) = l_i e^{-t}dt.p_{ij}(t) + \sum_{k=0}^{N_{(R)}-1} q_{ik}dt.m_{kj}(0, t)$$

Clearly $m_{ij}(0, t)$ is differentiable close to t so if we represent the derivative with respect to time by $\mathring{m}_{ij}(0, t)$ then we have

$$\mathring{m}_{ij}(0, t) = l_i e^{-t}p_{ij}(t) + \sum_{k=0}^{N_{(R)}-1} q_{ik}m_{kj}(0, t)$$

In matrix form this can be written

$$\mathring{\mathbf{M}}(0,t) = \mathbf{L}e^{-t}\mathbf{P}(t) + \mathbf{Q}\mathbf{M}(0,t)$$

To solve this we take Laplace transforms of both sides and take $\mathbf{M}(0,0) = 0$

$$\begin{aligned} s\hat{\mathbf{M}}(0,s) &= \mathbf{L}[(\mathbf{s}+\mathbf{1})\mathbf{I} - \mathbf{Q}]^{-1} + \mathbf{Q}\hat{\mathbf{M}}(0,s) \\ \hat{\mathbf{M}}(0,s) &= [(s\mathbf{I} - \mathbf{Q})]^{-1}\mathbf{L}[((s+1)\mathbf{I} - \mathbf{Q})]^{-1} \end{aligned}$$

Where $\mathbf{L}$ contains the levels of the process, $\mathbf{M}$ the contributions to the expectation and $\mathbf{Q}$ the infinitesimal transition probabilities.

$$\mathbf{L} = \begin{bmatrix} l_0 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & l_{N_{(R)}-1} \end{bmatrix}, \mathbf{M}(0,t) = \begin{bmatrix} m_{0,0}(0,t) & \cdots & m_{0,N_{(R)}-1}(0,t) \\ \vdots & \ddots & \vdots \\ m_{N_{(R)}-1,0}(0,t) & \cdots & m_{N_{(R)}-1,N_{(R)}-1}(0,t) \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} q_{0,0} & \cdots & q_{0,N_{(R)}-1} \\ \vdots & \ddots & \vdots \\ q_{N_{(R)}-1,0} & \cdots & q_{N_{(R)},N_{(R)}-1} \end{bmatrix}$$

We can now repeat this process starting by finding an expression such as Eqn. 4.20 but for the r^{th} moment. It is

$$\begin{aligned} p_{ij}(t)E[(X_{ij}(0,t))^r] &= (1+q_{ii}dt)p_{ij}(t-dt)E[(X_{ij}(X_{ii}(0,dt),t-dt))^r] \\ &\quad + \sum_{k\neq i} q_{ik}dt.p_{kj}(t-dt)E[(X_{kj}(X_{ik}(0,dt),t-dt))^r] \end{aligned} \quad (4.21)$$

We need a little creative algebra, expanding

$$\begin{aligned} E[(X_{ij}(X_{ii}(0,dt),t-dt))^r] &= E[(X_{ij}(0,t-dt) + X_{ii}(0,dt)e^{-(t-dt)})^r] \\ &= E[X_{ij}^r(0,t-dt)] \\ &\quad + X_{ii}(0,dt)e^{-(t-dt)}rE[X_{ij}^{r-1}(0,t-dt)] + o(dt) \end{aligned} \quad (4.22)$$

and

$$dt.E[(X_{kj}(X_{ik}(0,dt),t-dt))^r] = dt.E[X_{kj}^r(0,t-dt)] + o(dt) \quad (4.23)$$

Substituting Eqn. 4.22 and 4.23 in Eqn. 4.21.

$$\begin{aligned} m_{ij}(0,t;r) &= (1+q_{ii}dt)m_{ij}(0,t-dt;r) + r.dt.l_i(1-e^{-dt})e^{-(t-dt)}m_{ij}(0,t-dt;r-1) \\ &\quad + \sum_{k\neq i} q_{ik}dt.m_{ij}(0,t-dt;r) + o(dt) \end{aligned}$$

Ignoring smaller terms and letting $dt \to 0$, as before, we obtain

$$\mathring{m}_{ij}(0,t;r) = rl_i e^{-t} m_{ij}(0,t;r-1) + \sum_{k=0}^{N_{(R)}-1} q_{ik} m_{kj}(0,t;r) \tag{4.24}$$

Which can be expressed in matrix form

$$\mathring{\mathbf{M}}(0,t;r) = r\mathbf{L}e^{-t}\mathbf{M}(0,t;r-1) + \mathbf{Q}\mathbf{M}(0,t;r)$$

To solve this equation, we take the Laplace transform of both sides of (4.24), noting that $m_{ij}(0,0;r) = 0$.

$$s\hat{m}_{ij}(0,s;r) = rl_i\hat{m}_{ij}(0,s+1;r-1) + \sum_{k=0}^{N_{(R)}-1} q_{ik}\hat{m}_{kj}(0,s;r)$$

Writing this in matrix form we obtain

$$\begin{aligned} s\hat{\mathbf{M}}(0,s;r) &= r\mathbf{L}\hat{\mathbf{M}}(0,s+1;r-1) + \mathbf{Q}\hat{\mathbf{M}}(0,s;r) \\ (s\mathbf{I}-\mathbf{Q})\hat{\mathbf{M}}(0,s;r) &= r\mathbf{L}\hat{\mathbf{M}}(0,s+1;r-1) \\ \hat{\mathbf{M}}(0,s;r) &= r(s\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}\hat{\mathbf{M}}(0,s+1;r-1) \end{aligned}$$

When r is an integer we have, by recursion

$$\begin{aligned} \hat{\mathbf{M}}(0,s;r) &= r!(s\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}((s+1)\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}...((s+r-2)\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}\hat{\mathbf{M}}(s+r-1;1) \\ &= r!(s\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}((s+1)\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}... \qquad (4.25) \\ &\quad ...((s+r-2)\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}((s+r-1)\mathbf{I}-\mathbf{Q})^{-1}\mathbf{L}((s+r)\mathbf{I}-\mathbf{Q})^{-1} \qquad (4.26) \end{aligned}$$

By substituting the appropriate matrices into (4.25) we can obtain $\hat{\mathbf{M}}(0,s;r)$ which contains the higher moments of the transient distributions in the $N_{(R)}$-regime case. Once an expression for $\hat{\mathbf{M}}(0,s;r)$ has been found it will be necessary to perform an inverse Laplace Transform. The conclusion of this process will be an expression for the transitory moments of the process. Most of the terms should drop out on subsitution of $t = 0$ giving us the stationary conditional moments. Problems can occur in one of two areas, inverting each of the $((s+r)\mathbf{I}-\mathbf{Q})$ matrices and applying the inverse Laplace transform. While the inversion of the matrices is usually simple enough for 2×2 matrices, it may prove much more challenging when the process has a higher number of levels. The ease of application of the inverse Laplace transform depends heavily on the form of the inverted matrices. In the example below they take on a form that makes a partial fraction expansion easy and the expanded expression very amenable to the inverse transform. This is unlikely to be the case generally and other, more sophisticated, methods may need to be used.

Example:

To keep the calculations simple and to obtain a result that we can use later, we shall restrict ourselves to the two-regime case with levels $l_0 = 0$ and $l_1 = 1$. In which case if we are interested in the first moments of the transient distributions we can define the matrices

$$\mathbf{L} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{M}(0,t;r) = \begin{bmatrix} m_{00}(0,t;r) & m_{01}(0,t;r) \\ m_{10}(0,t;r) & m_{11}(0,t;r) \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} -a & a \\ b & -b \end{bmatrix}$$

This gives us

$$\begin{aligned}
(s\mathbf{I} - \mathbf{Q}) &= \begin{bmatrix} s+a & -a \\ -b & s+b \end{bmatrix} \\
((s+1)\mathbf{I} - \mathbf{Q}) &= \begin{bmatrix} s+a+1 & -a \\ -b & s+b+1 \end{bmatrix} \\
(s\mathbf{I} - \mathbf{Q})^{-1} &= \frac{1}{s(s+a+b)} \begin{bmatrix} s+b & a \\ b & s+a \end{bmatrix} \\
((s+1)\mathbf{I} - \mathbf{Q})^{-1} &= \frac{1}{(s+1)(s+a+b+1)} \begin{bmatrix} s+b+1 & a \\ b & s+a+1 \end{bmatrix}
\end{aligned}$$

Substitute these into 4.25, and replace $a+b$ with θ

$$\begin{aligned}
\hat{\mathbf{M}}(0,s) &= \hat{\mathbf{M}}(0,s;1) = \frac{1}{s(s+1)(s+\theta)(s+\theta+1)} \begin{bmatrix} 0 & a \\ 0 & s+a \end{bmatrix} \begin{bmatrix} s+b+1 & a \\ b & s+a+1 \end{bmatrix} \\
&= \frac{1}{s(s+1)(s+\theta)(s+\theta+1)} \begin{bmatrix} ab & a(s+a+1) \\ (s+a)b & (s+a)(s+a+1) \end{bmatrix}
\end{aligned}$$

If we assume that $a+b \neq 1$, then by partial fraction expansion and inversion we obtain a lengthy, but very useful, expression

$$\begin{aligned}
\mathbf{M}(0,t) &= \frac{1}{\theta(\theta+1)} \begin{bmatrix} ab & a(a+1) \\ ab & a(a+1) \end{bmatrix} - \frac{1}{\theta(\theta-1)} \begin{bmatrix} ab & a^2 \\ (a-1)b & a(a-1) \end{bmatrix} e^{-t} \\
&+ \frac{1}{\theta(\theta-1)} \begin{bmatrix} ab & -a(b-1) \\ -b^2 & b(b-1) \end{bmatrix} e^{-\theta t} - \frac{1}{\theta(\theta+1)} \begin{bmatrix} ab & -ab \\ -b(b+1) & b(b+1) \end{bmatrix} e^{-(\theta+1)t}
\end{aligned} \tag{4.27}$$

The cells of (above) gives us the transitory moments of a process at time t, given that $X(0) = 0$. The four cells of the matrix represent the four possible combinations of initial and final regime.

It is not much more difficult to work with higher moments in this case. When we pair the terms in (4.25) they become much more manageable. We can see this in Equation

(4.28)

$$\begin{aligned}((s+n)\mathbf{I}-\mathbf{Q})^{-1} &= \frac{1}{(s+n)(s+a+b+n)}\begin{bmatrix} s+b+n & a \\ b & s+a+n \end{bmatrix} \\ ((s+n)\mathbf{I}-\mathbf{Q})^{-1}.\mathbf{L} &= \frac{1}{(s+n)(s+a+b+n)}\begin{bmatrix} 0 & a \\ 0 & s+a+n \end{bmatrix} \end{aligned} \tag{4.28}$$

It follows from substituting (4.28) in (4.25) that

$$\begin{aligned}\hat{\mathbf{M}}(0,\mathbf{s};\mathbf{r}) &= r!\left(\prod_{n=0}^{r}\frac{1}{(s+n)(s+\theta+n)}\right) \\ &\quad\times\begin{bmatrix} 0 & a(s+a+1)...(s+a+r-1) \\ 0 & (s+a)...(s+a+r-1) \end{bmatrix}\begin{bmatrix} s+b+r & a \\ b & s+a+r \end{bmatrix} \\ &= r!\left(\prod_{n=0}^{r}\frac{1}{(s+n)(s+\theta+n)}\right)\left(\prod_{n=1}^{r-1}(s+a+n)\right) \\ &\quad\times\begin{bmatrix} ab & a(s+a+r) \\ (s+a)b & (s+a)(s+a+r) \end{bmatrix} \quad for\ r>1 \end{aligned} \tag{4.29}$$

From Equation (4.29) we can once again expand by partial fractions and invert. Simple substitutions will give us a solution of the form seen in Equation (4.27). For a stationary solution we are interested only in the first term of such an expansion.

Stationary Moments

We already know the stationary moments of the process, as they are those of a Beta distribution. We can easily check the validity of the above method by obtaining them from (4.27). We continue to work with the two-regime case.

We know that the cells of $\mathbf{M}(0,t)$ represent weighted conditional moments

$$\mathbf{M}(0,t) = \begin{bmatrix} p_{00}(t).E[X_0(0,t)|S(t)=0] & p_{01}(t).E[X_0(0,t)|S(t)=1] \\ p_{10}(t).E[X_1(0,t)|S(t)=0] & p_{11}(t).E[X_1(0,t)|S(t)=1] \end{bmatrix}$$

It is therefore easy to obtain the moments conditional only on the initial state

$$\begin{bmatrix} E[X_0(0,t)] \\ E[X_1(0,t)] \end{bmatrix} = \mathbf{M}(0,t).\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

But three of the four terms in (4.27) are dependent upon t. As $t \to \infty$, $e^{-t} \to 0$ leaving us with only one term to contribute to the stationary distribution.

$$\lim_{t\to\infty}\mathbf{M}(0,t) = \frac{1}{(a+b)(a+b+1)}\begin{bmatrix} ab & a(a+1) \\ ab & a(a+1) \end{bmatrix}$$

From this we can obtain two things.

First from either of the two rows we have the unconditional moment

$$E[X(t)] = \frac{a}{a+b}$$

Secondly, we obtain the moments conditional on the final state

$$\begin{aligned} E[X(t) \mid S(t) &= 0] = \frac{a}{a+b+1} \\ E[X(t) \mid S(t) &= 1] = \frac{a+1}{a+1+b} \end{aligned}$$

Of course these (and those for higher moments) are precisely what we would expect,. But this confirms the method, which could then be used to obtain results for the moments of processes with greater numbers of levels.

Covariance of the (2 regime) Process

Another result we can obtain using (4.27) has been taken from Jalali (2003d). After deriving an expression for the product of two successive observations from the same process (a time k apart) it is possible to obtain the following expression for the covariance of $X(t)$ and $X(t+k)$.

$$\begin{aligned} E[X(t)X(t+k)] &= \frac{a^2}{\theta^2} + \frac{1}{\theta-1}\left(\frac{ab}{\theta^2(\theta+1)}\right)\left[\theta\rho^k - \rho^{\theta k}\right] \\ where\ \rho &= e^{\frac{1}{\tau}} \end{aligned}$$

Since we know that

$$\begin{aligned} E[X(t+k)] &= E[X(t)] \\ Var(X(t)) &= \frac{ab}{\theta^2(\theta+1)} \end{aligned}$$

The correlation coefficient can also be found

$$Corr(X(t), X(t+k)) = \frac{\theta\rho^k - \rho^{\theta k}}{\theta - 1}$$

Although this is not defined for $\theta = 1$, by an application of L'Hôpital's rule it reduces to

$$(1 - k\ln\rho)\rho^k$$

4.4 The Discrete Approximation

Until this point we have been working with a continuous process. The observations are not assumed to occur at fixed intervals. The length of time the process resides in each state (the sojourns) have also been assumed to be continuous. The transition functions for this process are rather complex. Worse still, we find considerable problems with inference, something that we will deal with in the next chapter. Many of these problems may be exacerbated by the continuous nature of the model. We should consider how useful it would be to construct discrete models to approximate the behaviour of the Filtered Markov model. There are two elements of the process that are continuous in nature, the time scale and the level. We have the option of restricting either or both of these to a discrete set.

4.4.1 Discrete Time

Our first attempt to simplify the Filtered Markov process is to replace the continuous process with a discrete time process. In the continuous case we measure time using the index t, and our observations are taken at t_n where

$$\begin{aligned} t_n &\in \mathbb{R}^+ \\ n &\in \{1, 2, ..., N\} \\ and\ t_0 &< t_1 < t_2 < ... < t_n \end{aligned}$$

In order to keep our notation simple we have usually referred to time t rather than time t_n. In our discrete time approximation we shall have a fixed interval between our observations Even though our observations are still made at t_n we know that

$$\begin{aligned} t_n &= n.dt \qquad for\ n = \{1, 2, ..., N\} \\ where\ dt &= t_n - t_{n-1} \qquad for\ all\ n \end{aligned}$$

So when working with discrete time we shall simplify our notation by writing n when we mean t_n. We also retain a and b but they now refer to switching probabilities rather than swiching intensities.

This introduction of discrete time has a profound effect on the equations describing the movement of the process. In the continuous case we relied on the following rule of movement

$$X(t + dt) = S(t) - (S(t) - X(t)) \exp\left(-\frac{dt}{\tau}\right) \tag{4.30}$$

as long as $S(t)$ remained unchanged during $(t, t + dt]$. Now that $S(t)$ cannot change between observations due to the discrete nature of the driving signal (4.30) can be replaced

by (4.31)

$$\begin{aligned} X(n+1) &= S(n+1) - (S(n+1) - X(n)).\rho \\ Where\ \rho &= \exp\left(-\frac{1}{\tau}\right) \end{aligned} \tag{4.31}$$

Distribution of the Discrete Time Process

The combination of the regime and level $(S(n), X(n))$ is still Markov as we can obtain the future distributions from just these two pieces of information. However, while the support of the stationary distribution of the continuous time model is a dense set that of the discrete time model is of a Cantor type. For a finite number of steps the transitional distribution is a countable set, but the set becomes uncountable for the stationary distribution.

For instance

$$if\ X(0) = x(0)\ and\ S(0) = 0\ then\ X(1) = \begin{cases} x(0)\rho & with\ probability\ 1-a \\ 1-(1-x(0))\rho & with\ probability\ a \end{cases}$$

We can generate a representation of a process with levels 0 and 1, of length 50 with $a = b = 0.5$ and $x(0) = 0.5$. A PDF is estimated using 5000 realisations of this process and displayed in the histogram in Figure 4.4. In a visual representation such as this we can show only a limited number of bars. Even though much of the complexity of the distribution at this point is lost in this visual representation the nature of the distribution is apparent. Despite the quite different structure at lower levels the behaviour of the process results in similar cumulative distribution functions to the original model. This similarity is quite good for large values of τ.

We shall show two examples of comparisons made between the theoretical CDF of the continuous time case with an ECDF drawn from the equivalent discrete time case. These ECDFs were obtained using 1000 values of $X(50)$ from a continuous time process with fixed starting point $x(0) = 0.5$. In the first example (shown in Figure 4.5) we use $\alpha = \beta = 0.2$ and $\tau = 5$ ($\rho = 0.8187$).and find a slight deviation from the Uniform distribution.

In the second example (Figure 4.6) we use $\alpha = \beta = 0.4$ and $\tau = \frac{5}{4}$ ($\rho = 0.4493$). Due to the faster switching we find a much less smooth ECDF than in the first example and one that is characterised by an central interval of zero probability.

Fitting the Discrete Time Process

We have two main options when it comes to fitting the discrete time process. The first option is to work with the Markov process $(S(n), X(n))$. This allows very straightforward generation of a series but creates problems when trying to fit the model to an observed series. Whether we are attempting to fit the model as an approximation to a true Filtered Markov process or to an observed process contaminated by noise we will find less than perfect data. For a Bayesian Filter, such as that proposed by Hamilton, to work we need to be able to

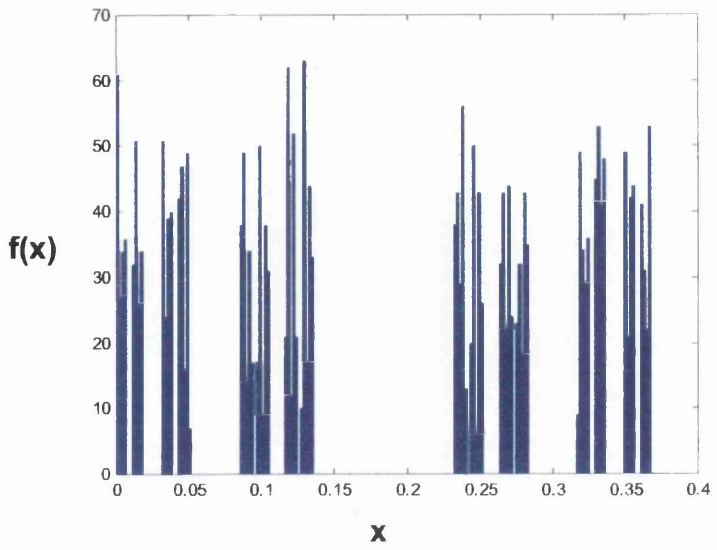

Figure 4.4: A representation of the Cantor type distribution of the PDF of the discrete time, continuous level model.

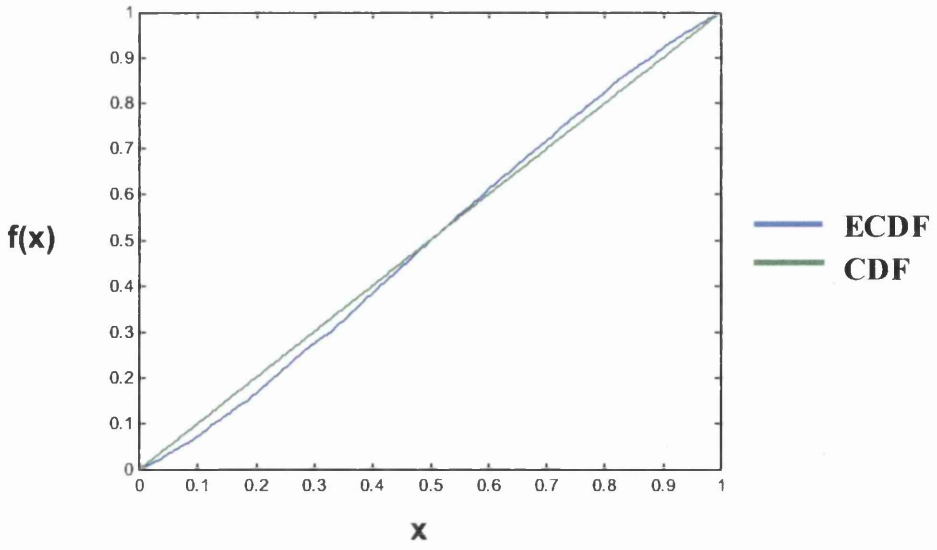

Figure 4.5: A comparison between the ECDF of a discrete time, continuous level model (with parameters $\alpha = \beta = 0.2, \tau = 5$) and the CDF of the $Beta[1, 1]$.

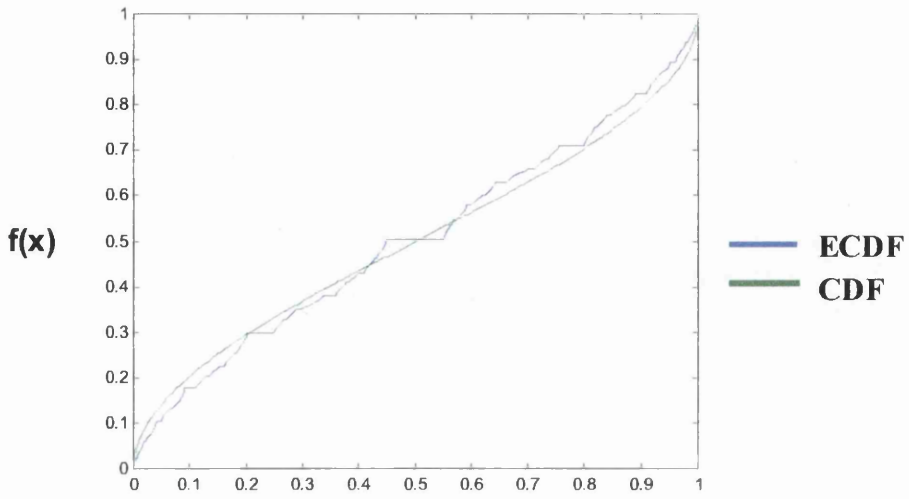

Figure 4.6: A comparison between the ECDF of the discrete time, continuous level model (with parameters $\alpha = \beta = 0.4$, $\tau = \frac{5}{4}$) and the CDF of $Beta\left[\frac{1}{2}, \frac{1}{2}\right]$.

attach probabilities to a finite number of states of the combined process $(S(n), X(n))$. As $X(n)$ is continuous there is no such structure of finite states and so we would be forced to impute a historical sequence of regimes $(S(n))$ that maximised the likelihood. This introduces an incredibly complex likelihood space with numerous local maxima, making obtaining the MLE considerably more difficult.

A second option to avoid these complications is to consider the Markov chain consisting of the history of the regime occupied during the r most recent intervals. We can call this Markov chain $H(n)$ where

$$H(n) = [S(n), S(n-1), ..., S(n-r+1)]$$

Our transition probabilities are easy to caculate given that they depend on the transition from $S(n)$ to $S(n+1)$. We can find $H(n+1)$ (using the Markov property) by

$$\begin{aligned}
\Pr[S(n+1) &= s(n+1), ..., S(n-r+1) = s(n-r+1)] \\
&= \sum_{s(n-r+1)=0}^{1} \Pr\left[S(n) = s(n), ..., S(n-r+1) = s(n-r+1)\right] \\
&\qquad \times \Pr\left[S(n+1) = s(n+1) | S(n) = s(n)\right]
\end{aligned}$$

The only problem we face is that the recent history of regimes (alone) does not precisely determine the level of the process. If we take $l_0 = 0$ and $l_1 = 1$ then this position is given by

$$x(n) = x(n-r).\rho^r + \sum_{k=0}^{r-1} s(n-k)\rho^k(1-\rho)$$

Which requires knowledge of $x(n-r)$, which is unobservable. Instead it gives us an interval I in which the level of the process must lie. This interval will be given by

$$I(n) = \left(\sum_{k=0}^{r-1} s(n-k)\rho^k(1-\rho) \ , \ \rho^r + \sum_{k=0}^{r-1} s(n-k)\rho^k(1-\rho) \right)$$

The interval I is of length ρ^r and if this is small enough we can simply take the midpoint as our estimate of the level of the process. How good an approximation this is depends upon the relationship between r and ρ. This process is easy and efficient to model although the matrices required for a fitting algorithm are of the order $o(2^r)$. If ρ is large then we require a high value of r in order to ensure the interval $I(n)$ is narrow. This has profound consequences for the computational demands of the method.

4.4.2 Discrete Range

As a result of our experimentation we concluded that discrete time continuous level models were generally problematic. It seems likely that a better approximation of the Filtered Markov model can be found working with a continuous time, discrete level process. We replace the fully continuous model with a continuous time Markov process $(S(t), L(t))$, where $S(t)$ is a two-regime continuous time Markov process in its own right. As in the FM process S_t represents the regime of the series (growth or decay) while $L(t)$ is a finite level process whose transition probabilities are controlled by $S(t)$. The intention is to use $L(t)$ as a discrete approximation of $X(t)$. The levels of $L(t)$ represent sub-intervals of the range of the process $X(t)$ between its upper and lower boundaries. It would be possible to evaluate the true transition probabilities between intervals by integrating across both. In general, however, as long as the number of levels is large enough, taking the probability density of a transition between the midpoints of the intervals will give us a reasonable estimate. Allowance must also be made for the impulse part of the transition function resulting from a switch free interval between observations. This can be done by adding the probability of this occuring to the appropriate cell of the transition matrix.

Distribution of the Discrete Range Model

The quality of the approximation of the distribution function to its theoretical counterpart will depend solely on the number of levels we choose to represent the continuous range of $X(t)$. If this is set to a number as low as 10 we will find quite grave shortcomings with our approximation. A more reasonable number would be 50. Even though the number of levels may seem large this does not pose a serious problem for the Bayesian Filter. When using a Filter we only need record the inferred distribution of the process at the previous observation, with the Matrix representation of the Markov process $(S(t), L(t))$ is the tensor product of the vectors of the possible values of $S(t)$ and $L(t)$.

Fitting the Discrete Range Model

We provide a full explanation of the algorithm for fitting this model to data on page 74 in the section titled 'Approximate Likelihood Method'. The process is relatively time efficient, given the complexity of the transition probabilities and we have explored its use for a range of different cases. By varying the noise levels, parameter values and even the length of the sample data set and number of levels we are able to provide guidelines for the practical limits on the use of this algorithm.

4.4.3 A Fully Discrete Model

One of our original intentions was to modify the kind of Markov Switching model proposed by Hamilton to allow the introduction of gradual transitions to regime changes. The Filtered Markov model was one such attempt to do this but represents a radical departure from the original specification of a Markov Switching model, in which both time and level were discrete in nature. In the previous sections we have been concerned with simplifications of the Filtered Markov model motivated by the simpler inference the discrete models offered. There is also room for development in the other direction, constructing new Markov Switching models to capture the gradual switching dynamics of the Filtered Markov process. On page 143 we introduce a class of models with these characteristics. These Ladder models will be define in Chapter 8 and will bridge the gap between the Filtered Markov model and Hamilton's Markov switching model.

Chapter 5

Fitting the Filtered Markov Model

We have introduced a continuous time, continuous level Markov process we call the Filtered Markov (FM) process (or model). In this chapter we shall be looking at the process of fitting this model to data using Maximum Likelihood. We shall also outline several new methods for parameter estimation for a sample from the Beta distribution where the observations are contaminated by noise.

5.1 Introduction

The intention is this chapter is to develop a method, or methods, for fitting the two-regime FM model. This full model consists of 8 parameters, 6 main parameters and 2 to determine the initial conditions. The six main parameters the model requires are the two switching intensities α and β, the two levels l_0 and l_1, the time scaling factor τ and σ the standard deviation of the added noise. The initial conditions consist of the position of the process, $x(0)$, and the regime of the driving signal, $s(0)$. We shall eventually propose methods for estimating all 8 parameters simultaneously, but shall start with simpler methods for fitting a reduced parameter model. We shall begin with the simplest case of a two-parameter estimation problem and build up to the full model.

5.2 Generation of data

The data is generated using the programming language MATLAB, as are many of the other routines we have used. Simulating data from a FM process is very straighforward and the algorithm used is presented here.

The first series we generate is a summary of the switchpoints, since the behaviour between these points is deterministic. We shall denote this series $Z(t_n)$. The initial conditions, $z(0)$ and $s(0)$, will be chosen at random (with appropriate stationary probabilities). The generation process involves producing a series of values drawn from an exponential distribution with alternating means (determined by the appropriate switch intensity) to represent

the intervals between switches. From these intervals we have the timings of switches. This series of switch points, and the parameter τ, will then allow us to move forward through the series determining the position of the process at any of the switch points. The final part of the process is converting from our recording of the series on a continuous scale via the position at switchpoints to recording the process using values over a predetermined fixed interval scale. The observations will occur at times $\{t_n\}$ where $n = \{1, 2, ..., N\}$

If we are choosing the intial conditions randomly then we choose the initial regime from the stationary distribution of the driving signal first

$$\begin{aligned} \Pr(S(0) &= 0) = \frac{\beta}{\alpha+\beta}\ ,\ \Pr(S(0)=1) = \frac{\alpha}{\alpha+\beta} \\ If\ S(0) &= 0 \quad then \quad Z(0) \sim Beta(a\tau, b\tau+1) \\ If\ S(0) &= 1 \quad then \quad Z(0) \sim Beta(a\tau+1, b\tau) \end{aligned}$$

First we must generate the sequence of intervals that will (together with the initial conditions) define the process. The initial regime $s(0)$ determines the regime occupied during each of the successive time intervals between switch points. The regime occupied by the driving signal during the k^{th} interval is given by $S(k)$. The interval lengths are denoted by $\{u(1), u(2), ...\}$

The number of switches required in a simulated sample, J, is determined by the final observation time N. We keep incrementing J until the following condition is met

$$\sum_{j=1}^{J-1} u(j) < N \leqslant \sum_{j=1}^{J} u(j)$$

Each of the intervals represented by $u(j)$ for $j = \{1, 2, ..., J\}$ are exponentially distributed

with alternating means. Due to the lack of memory property of the Exponential distribution the first of these intervals, $u(1)$, can be represented by the same distribution even though $t = 0$ may not be a switch point. The mean of this first Exponentially distributed interval time $u(1)$ is dependent upon the initial regime.

$$\begin{aligned} u(1), u(3), ... &\sim Exp(1/\alpha\tau) \quad and \quad u(2), u(4), ... \sim Exp(1/\beta\tau) \quad if \quad S(0)=0 \\ u(1), u(3), ... &\sim Exp(1/\beta\tau) \quad and \quad u(2), u(4), ... \sim Exp(1/\alpha\tau) \quad if \quad S(0)=1 \end{aligned}$$

Then the switch times can be easily found

$$\begin{aligned} V &= \{v(1), v(2), ..., v(J)\} \\ v(n) &= \sum_{k=1}^{n} u(k) \end{aligned}$$

So

$$v(n) = v(n-1) + u(n)$$

It is simple to determine the positions of the process at switch points using these time intervals, and $z(0) = x(0)$. We represent the position of the process at the k^{th} switch point (at time $v(k)$) by $z(k)$

$$\begin{aligned} z(k) &= l_0 - (l_0 - z(k-1)).\exp(-\tfrac{u(k)}{\tau}) \qquad if\ S(k) = 0 \\ z(k) &= l_1 - (l_1 - z(k-1)).\exp(-\tfrac{u(k)}{\tau}) \qquad if\ S(k) = 1 \end{aligned}$$

Finally we wish to work with a process that is observed at points separated by fixed intervals.

This process, $\{x(t_n)\}$, is found using

$$\begin{aligned} x(t_n) &= l_0 - (l_0 - z(k-1)).\exp\left(-\tfrac{(t_n - v(k-1))}{\tau}\right) \qquad if\ t_n \in (v(k-1), v(k)]\ and\ S(k) = 0 \\ x(t_n) &= l_1 - (l_1 - z(k-1)).\exp\left(-\tfrac{(t_n - v(k-1))}{\tau}\right) \qquad if\ t_n \in (v(k-1), v(k)]\ and\ S(k) = 1 \end{aligned}$$

The observed process $\{y(t_n)\}$ is the simulated process we will use for model testing. It is obtained by adding Gaussian noise (if required) to the series $\{x(t_n)\}$.

$$y(t_n) = x(t_n) + \epsilon(t_n) \qquad Where\ \epsilon(t_n) \sim N(0, \sigma^2)\ and\ \epsilon(t_1), \epsilon(t_2), ...\ are\ independent$$

5.3 Model Fitting using Maximum Likelihood

When it comes to obtaining parameter estimates from a set of sample data from the FM process, there are two main options available to us. These two methods rely on Maximum Likelihood (ML) and the Method of Moments. As with any stochastic process a large amount of data is contained in the order in which the observations are made. By choosing to use ML we ensure that we will be able to utilise at least some of this information. Our notation is as follows:

We shall assume without loss of generality that our observation times t are equally spaced at integer values of t. We have a set of data $\{x(t)\}$ generated by the filtered Markov process with no noise added. We use $\mathbf{x}_t$ to represent the history of observed values of the process at time t. So

$$\mathbf{x}_t = \{x(t), x(t-1), ..., x(1)\}$$

We wish to obtain a likelihood value for a given set of values of the parameters.

The process is defined using 5 parameters.

These are α and β, the switching intenstities up and down respectively,

The upper and lower levels of the process l_0 and l_1,

and τ, the time scaling factor.

The set of parameters $(\alpha, \beta, l_0, l_1, \sigma)$ as a set will be denoted by $\mathbf{\Theta}$. When combined with the initial distribution $\mathbf{\Pi}$ we shall use $\mathbf{\Sigma} = (\mathbf{\Theta}, \mathbf{\Pi})$.

We can construct a simple algorithm for determining the likelihood.

As input it will require

$$\Pr(S(t-1) = i | \mathbf{x}_{t-1})$$

And will produce as output

$$\Pr(S(t) = i | \mathbf{x}_t)$$

and

$$f(x(t) | \mathbf{x}_{t-1})$$

We will require the use of $r_{ij}(x, y, dt)$ defined in Section 4.3.3. We simplify this to $r_{ij}(x, y)$ as we shall always take $dt = 1$

$$r_{ij}(x, y)dy = \Pr[s(t+1) = j \ \& \ x(t+1) \in (y, y+dy) \mid s(t) = i \ \& \ x(t) = x]$$

which is dependent upon all of the parameters of the model.

Step 1

We can write our inferred distribution of the process at time $t-1$ as

$$\begin{bmatrix} \Pr(S(t-1) = 0 | \mathbf{x}_{t-1}) \\ \Pr(S(t-1) = 1 | \mathbf{x}_{t-1}) \end{bmatrix}$$

At time $t = 0$ we take instead

$$\begin{bmatrix} \Pr(S(0) = 0) \\ \Pr(S(0) = 1) \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix},$$

where π_i is the stationary probability of the process occupying the i^{th} state.

We can then obtain

$$\begin{bmatrix} f(x(t), S(t)=0|\mathbf{x}_{t-1}) \\ f(x(t), S(t)=1|\mathbf{x}_{t-1}) \end{bmatrix} = \begin{bmatrix} r_{00}(x(t-1);x(t)) & r_{01}(x(t-1),x(t)) \\ r_{10}(x(t-1),x(t)) & r_{11}(x(t-1),x(t)) \end{bmatrix}' \times \begin{bmatrix} \Pr(S(t-1)=0|\mathbf{x}_{t-1}) \\ \Pr(S(t-1)=1|\mathbf{x}_{t-1}) \end{bmatrix}$$

Step 2

From which we infer

$$f(x(t)|\mathbf{x}_{t-1}) = f(x(t), S(t)=0|\mathbf{x}_{t-1}) + f(x(t), S(t)=1|\mathbf{x}_{t-1})$$

and finally

$$\begin{bmatrix} \Pr(S(t)=0|\mathbf{x}_t) \\ \Pr(S(t)=1|\mathbf{x}_t) \end{bmatrix} = \frac{1}{f(x(t)|\mathbf{x}_{t-1})} \times \begin{bmatrix} f(x(t), S(n)=0|\mathbf{x}_{t-1}) \\ f(x(t), S(n)=1|\mathbf{x}_{t-1}) \end{bmatrix}$$

We then return to Step 1, using this as an input, and repeat until we reach $t = T$ when we proceed instead to Step 3

Step 3

The likelihood, and log-likelihood, are evaluated by

$$L(\mathbf{\Sigma}) = \prod_{t=0}^{T} f(x(t)|\mathbf{x}_{t-1})$$

$$l(\mathbf{\Sigma}) = \sum_{t=0}^{T} \ln(f(x(t)|\mathbf{x}_{t-1}))$$

Care must be taken during the process to ensure that numerical errors do not influence the result. The transitional distribution function contains a significant impulse element to model switch free observation intervals. Even a small rounding error in the data during a true switch free interval may result in this not matching precisely the position of the impulse in the distribution function. This can be remedied by broadening the impulse into a (very) short interval.

The likelihood method can now be applied to some simple estimation problems.

5.3.1 Case 1: 2 Parameters (α, β)

We can now generate data easily and we have a method for inferring the value of the parameter values used to generate the simulated data At first we shall examine a very

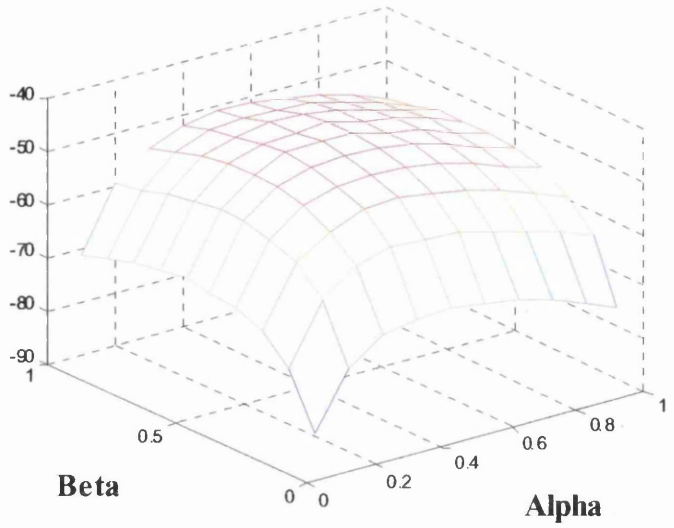

Figure 5.1: Log-Likelihood surface for a simulated symmetric process ($\alpha = \beta = \frac{1}{2}$) derived using the 2-parameter fitting method outlined in Case 1.

simple estimation problem before adding in extra parameters. The simplest case we shall concern ourselves with is where we have an asymmetric, standardised, noise-free Filtered Markov process That is, where:

$$\alpha \neq \beta \ , \ l_0 = 0 \ , \ l_1 = 1 \ , \ \tau = 1 \ , \ \sigma = 0$$

As such we have only two parameters to estimate, α and β. The initial conditions were randomly chosen and the remainder of the parameters were fixed. A short series was generated (100 values) over a range of possible values for α and β, although we have restricted ourselves mostly to values in the range [0,1]. For the upper end of this scale we will find the process switching so frequently that we will rarely find a switch free interval. This is not really in keeping with the original concept of this research of using the model to represent gradual transitions between infrequent changes of regime.

In general we had little problem obtaining good estimates for the parameter values. This was true even for the middle or upper end of the scale and even for short series such as this one. In Figure 5.1 we show the log-likelihood for one such example, for $\alpha = \beta = 0.5$. The clarity of the estimate is better demonstrated by looking instead at the likelihood function, shown here in Figure 5.2.

In general working with the Likelihood was a very effective way of solving for this type of problem.

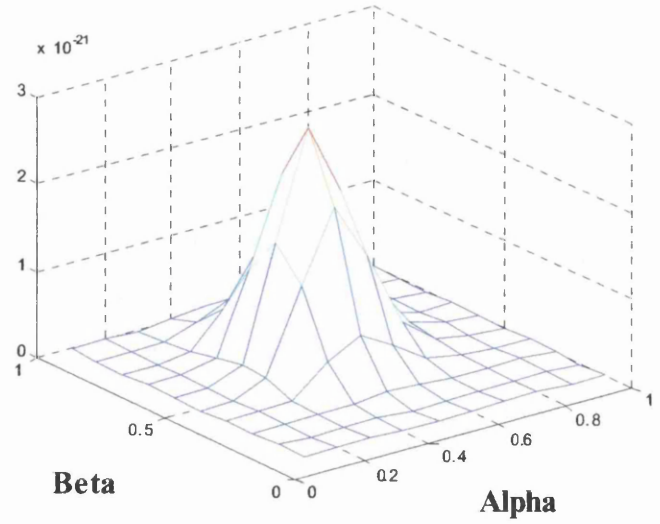

Figure 5.2: The Likelihood surface for a simulated symmetric process ($\alpha = \beta = \frac{1}{2}$) derived using the 2-parameter fitting method outlined in Case 1.

5.3.2 Case 2: 2 Parameters (α, τ)

Given that we can deal effectively with the the first estimation problem, we shall now move onto another case. A decision had to be made as to which parameter to add next. Experiments were conducted with several different three parameter models with mixed results. Of these by far the most intriguing concerned the addition of the time-scaling factor τ. When this was added to the original two-parameter problem great difficulties were found in obtaining an estimate. To examine this further we simplified this case to another two-parameter case by taking the symmetric case of $\alpha = \beta$. The early examples for lower values of of α and β did yield a maximum likelihood close to the values used to generate the process. Figure 5.3 shows the results of one such example, for $\alpha = \beta = 0.1$ and $\tau = 1$. The surface was generated over a range of values for α falling between 0 and 1, against values for τ between 0.9 and 1. When working with a noise free series such as this it is unlikely that we will have a problem with the overestimation of τ. This is due to the influence of any switch free intervals in the series, whose deterministic movement is controlled by the τ parameter. If the proposed value of τ is too large to allow for the transition then the new value will fall outside the transitional distribution (conditional on the previous value) and so will return a 0 probability.

We can observe that there is a second, local, maximum log-likelihood value in the region covered by the graph. This represents the solution that does not depend on the impulse part of the transitional distribution function, that models switch free intervals. As α increases in magnitude this maximum increases in likelihood eventually dominating the 'true' solution. A second example, shown in Figure 5.4, uses a data set generated using intensities only

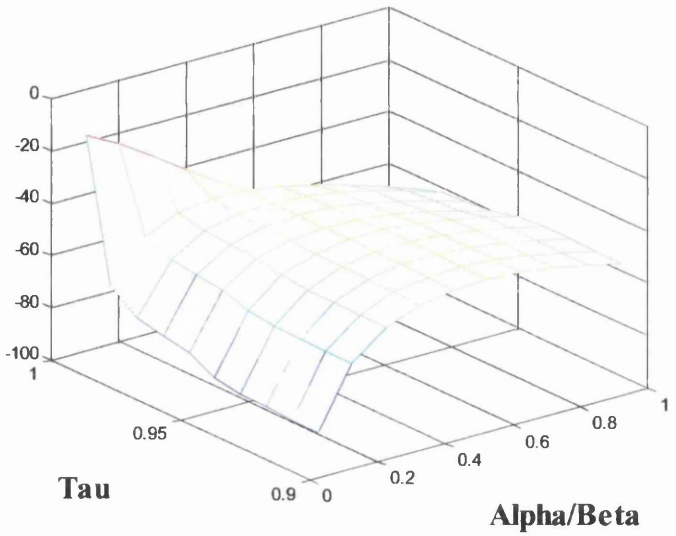

Figure 5.3: The Log-Likelihood surface for a simulated symmetric process ($\alpha = \beta = \frac{1}{10}$) derived using the 2-parameter fitting method outlined in Case 2.

slightly higher at $\alpha = 0.2$.

By taking two slices through the graph corresponding to $\tau = 0.95$ and $\tau = 1$ we can see that there is a second solution that now dominates the first (see Figure 5.5). We can also see that this does actually give an estimate but this is near $\alpha = 1.3$ and does not relate to the value of the parameters used to generate the process.

This problem is not simply one created by numerical errors or short data sets. Even if we specify greater precision in our evaluation of the distribution and use 1000 values instead of 100 we still find exactly the same result. We also find the same result for the asymmetric case, again dependent upon the magnitude of the parameters. Were this problem limited only to values of α close to 1 we could probably accept it, given that we are unlikely to model data exhibiting such high frequency switching. In practice finding data with switching intensities higher than 0.1 would hardly be seen as extreme. To put them in context, the solutions Hamilton obtained for the (discrete) two-regime Markov model of US GNP had switching probabilities of 0.2450 and 0.0951.

The problems we encounter when dealing with the time scaling factor τ are of a different nature to those produced by other parameters. We cannot simply increase the sample size and refine our estimates. It will be necessary to reconsider the methods we use for estimating τ.

5.3.3 Estimating the Time Scaling Factor

It seems that the use of likelihoods alone is unlikely to allow us to solve estimation problems containing the time scaling factor τ. Once we have an estimate for τ we can still use

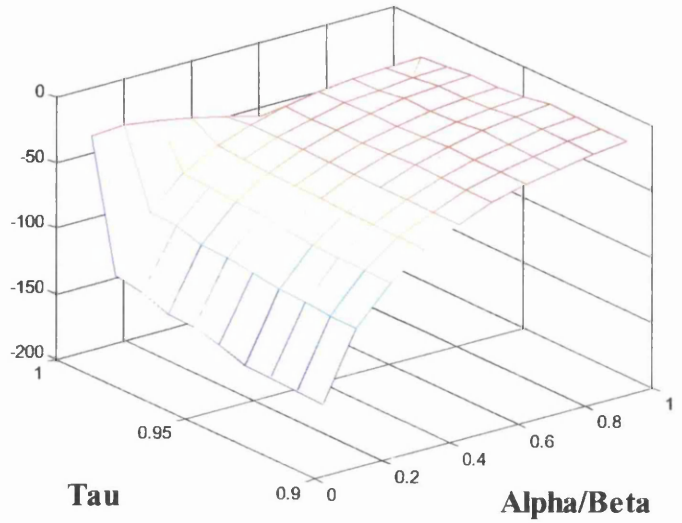

Figure 5.4: The Log-Likelihood surface for a simulated symmetric process ($\alpha = \beta = \frac{2}{10}$) derived using the 2-parameter fitting method outlined in Case 2.

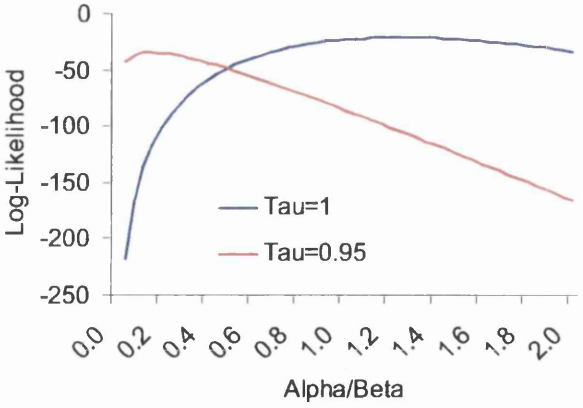

Figure 5.5: A Comparison of the Log-Likelihood values obtained for a simulated symmetric sample ($\alpha = \beta = 0.2$) along the planes $\tau = 0.95$ and $\tau = 1$.

likelihoods to obtain good estimates for the other parameters. Therefore we need to explore other methods for finding τ.

Where we have a pure estimation problem i.e. where our sample is uncontaminated by noise and we know l_0 and l_1, we use a method is proposed in Jalali (2003e) for estimating τ. The sample size is taken as T. For any two consecutive points $x(t_1)$ and $x(t_2)$ there are upper and lower bounds of the possible values that can be taken by $x(t_2)$ given $x(t_1)$.

$$\begin{aligned}\text{If } S(t) &= 0 \text{ for all } t_1 \leqslant t < t_2 \text{ then } x(t_2) = l_0 - (l_0 - x(t_1))\, exp\left(-\frac{t_2 - t_1}{\tau}\right)\\ \text{If } S(t) &= 1 \text{ for all } t_1 \leqslant t < t_2 \text{ then } x(t_2) = l_1 - (l_1 - x(t_1))\, exp\left(-\frac{t_2 - t_1}{\tau}\right)\end{aligned}$$

Therefore

$$l_0 - (l_0 - x(t_1))\, exp\left(-\frac{t_2 - t_1}{\tau}\right) \leqslant x(t_2) \leqslant l_1 - (l_1 - x(t_1))\, exp\left(-\frac{t_2 - t_1}{\tau}\right)$$

We can define our estimators

$$\begin{aligned}R(t_2) &= \left(\frac{x(t_2) - l_0}{x(t_1) - l_0}\right)^{\frac{1}{t_2 - t_1}} \qquad \text{if } x(t_2) < x(t_1)\\ R(t_2) &= \left(\frac{l_1 - x(t_2)}{l_1 - x(t_1)}\right)^{\frac{1}{t_2 - t_1}} \qquad \text{if } x(t_2) > x(t_1) \quad ,\end{aligned}$$

where l_0 and l_1 are the two levels of the process. Then we have an estimator for $\rho = \exp(-\frac{1}{\tau})$ by taking

$$\begin{aligned}\hat{\rho} &= \min\{R(t_1), R(t_2), ..., R(t_T)\}\\ \text{and then } \hat{\tau} &= -\frac{1}{\ln(\hat{\rho})}\end{aligned}$$

If the sample size is large enough and the intervals small enough relative to the frequency of the switching this should, with high probability, give a 100% accurate estimate. If not then at least we know that we have an overestimate, and hence an upper bound, for $\hat{\tau}$. In practice we find numerical errors result in a few rogue values falling below this theoretical estimate. We can see an example of this in Figure 5.6, which displays the CDF of the R values for a sample of 200 with $\alpha = \beta = 0.9, l_0 = 0, l_1 = 1$ and $\tau = 1$. We obtain $\hat{\rho} = 0.35$ where the 'true' value of ρ is 0.37.

Unfortunately this method is not particularly robust if the data is contaminated with any form of noise. In Figure 5.7 we see the CDFs of the test statistics R when the data is contaminated with Gaussian noise. We can see the complete absence of the step in the CDF once the standard deviation of the noise level reaches 0.05. In making these estimates we have full knowledge of the levels of the model and so it appears that this type of approach is not very robust. One possibility would be to introduce some form

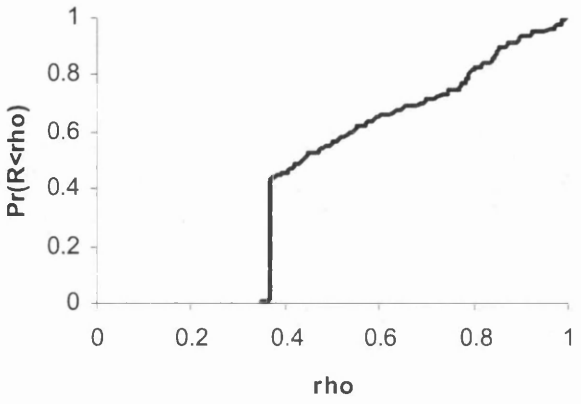

Figure 5.6: The ECDF of the test statistic R, used in the process of estimating Tau for simulated data (details given in text)

of parameter augmentation into the estimation procedure, but this could bring its own problems and so we first try an approximate method.

5.3.4 Approximate Likelihood

At least one of the properties of the transitional distributions is creating estimation problems. This is the inclusion of the impulse function to represent the deterministic movement of the process when no switch occurs during an interval. Although the addition of noise to a previously clean signal is in general a hindrance, it may alleviate the consequences by reducing this singular element. It will of course also deny us the precise knowledge of the position of the process that is available to us with a clean signal. Either we will require the transitional distribution function for a noisy process or we will be forced to seek a solution dependent upon numerical approximation. The complexity of the distribution functions makes the former seem rather imposing and we provide instead an algorithm for achieving the latter.

When obtaining the true value of the Likelihood with clean data we only made inference about the regime of the hidden Markov signal. Now we must also rely on inference about the current position of the clean (noise free) process $\{x(t)\}$, as we have only the observations of the observable process $\{y(t)\}$ where

$$\begin{aligned} y(t) &= x(t) + \epsilon(t) \\ &\quad \textit{and } \epsilon(t) \backsim N(0, \sigma^2) \textit{ and } \epsilon(s) \textit{ and } \epsilon(t) \textit{ are independent for any } s \neq t \end{aligned}$$

In order to do this we must decide on the number of levels we wish to use to approximate

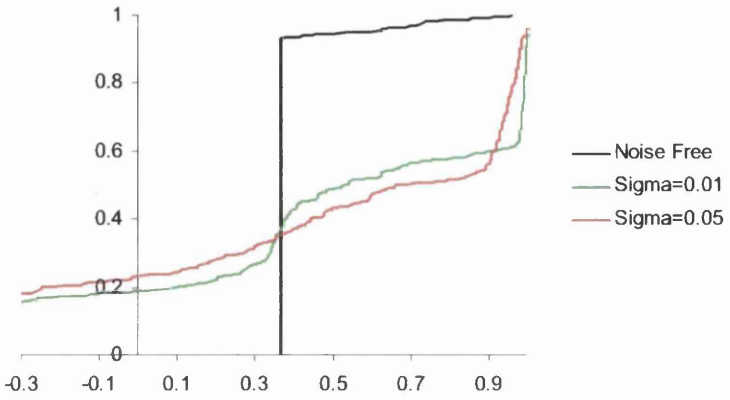

Figure 5.7: The ECDFs of R, the estimator for Rho in the presence of different levels of Gaussian noise.

to the continuous range of the clean process and we call this $N_{(L)}$. The greater the number of levels the greater the accuracy of the approximation but the slower the calculations. Experimentation suggests that 10 levels is too few for the approximation to work and 50 is probably a more reasonable lower bound.

The range of the clean process $\{x(t)\}$ is $[l_0, l_1]$. This interval is then split into $N_{(L)}$ sub-intervals I_i. for $i = \{1, 2, ..., N_{(L)}\}$. We shall take $l_0 = 0, l_1 = 1$ to simplify the notation, wlog. The intervals I_i are given by

$$I_i = \left[\frac{(i-1)}{N_{(L)}}, \frac{i}{N_{(L)}}\right)$$

and i is then used as the index of the level. From then on we treat the clean process $x(t)$ as if it were discrete rather than continuous.

To complicate matters further we need to model both the regime of the Markov driving signal $(S(t))$ and the level of the clean process $X(t)$.

We do this by defining a new process $L(t)$ (which has $2N_{(L)}$ possible levels) to replace $S(t)$ and $X(t)$. The first $N_{(L)}$ levels of $L(t)$ represent the i intervals that can be occupied by $\{x(t)\}$ while the driving signal has regime $S(t) = 0$. The the second set of $N_{(L)}$ levels represent those same intervals while $S(t) = 1$. In notation

$$\begin{aligned} &if\ i \in \{1, 2, ..., N_{(L)}\}\ then\ L(t) = i\ represents\ the\ case\ where\ S(t) = 0\ and\ x(t) \in I_i \\ &if\ i \in \{N_{(L)} + 1, ..., 2N_{(L)}\}\ then\ L(t) = i\ represents\ S(t) = 1\ and\ x(t) \in I_{i-N_{(L)}} \end{aligned}$$

We then apply a version of the Basic Filter (as defined on page 17) but make inference on the process $L(t)$ rather than $S(t)$.

We require as input

$$\Pr(L(t-1) = i|\mathbf{x}_{t-1})$$

And will produce as output

$$\Pr(L(t) = i|\mathbf{x}_t)$$

and

$$f(x(t)|\mathbf{x}_{t-1})$$

When it comes to determining the transitional probabilities we again turn to an approximation.

Instead of using the density $p_{ij}(x(t_1), x(t_2), t_2 - t_1)$ for the transitions we introduce a matrix T where:

$$T_{ij} = \Pr(L(t) = j|L(t-1) = i)$$

We define $p_{ij}^{(s)}(x(t_1), x(t_2), t_2 - t_1)$ to represent the transitional distribution that requires at least one switch to have occured during the interval (hence no impulse element). It is presumed that each interval I_i can be represented by its midpoint $\bar{I}_i$. These midpoints are given by

$$\begin{aligned} \bar{I}_k &= \tfrac{2k-1}{2N_{(L)}} && if\ k \in \{1, 2, ..., N_{(L)}\} \\ \bar{I}_k &= \tfrac{2(k-N_{(L)})-1}{2N_{(L)}} && if\ k \in \{N_{(L)}+1, N_{(L)}+2, ..., 2N_{(L)}\} \end{aligned}$$

We evaluate the transition probabilities between the levels of L

$$\Pr(L(t) = j|L(t-1) = i) = p_{ij}^{(s)}\left(\bar{I}_i, \bar{I}_j, 1\right) . \left(\frac{1}{N_{(L)}}\right)$$

To complete T we need to incorporate the behaviour when no switching occurs (the impulse element). For each level we increase the appropriate cells of T by the probability of no switch $P^{(ns)}$

And if $x_i^{(ns)}$ represents the position of the process after a switch free interval starting in interval i and we continue to represent every interval by its midpoint, then

$$\begin{aligned} x_i^{(ns)} &= \bar{I}_i \exp(-\tfrac{1}{\tau}) && for\ i \in \{1, ..., N_{(L)}\} \\ x_i^{(ns)} &= 1 - \left(1 - \bar{I}_i\right) \exp(-\tfrac{1}{\tau}) && for\ i \in \{N_{(L)}+1, ..., 2N_{(L)}\} \end{aligned}$$

And the probability of a switch free interval is

$$\begin{aligned} P^{(ns)} &= \exp(-\alpha) && if\ i \in \{1, ..., N_{(L)}\} \\ P^{(ns)} &= \exp(-\beta) && if\ i \in \{N+1, ..., 2N_{(L)}\} \end{aligned}$$

Then we increase the appropriate cell in T by this probability

$$T_{ij} = T_{ij} + P^{(ns)} \qquad if\ x_i^{(ns)} \in I_j$$
$$T_{ij} = T_{ij} \qquad if\ x_i^{(ns)} \notin I_j$$

This approximation is quite satisfactory for our purposes if $N_{(L)}$ is large enough, although to ensure that we have a proper transition matrix we will need to rescale the rows of T slightly..

We have constructed our process by considering a clean signal $\{x(t)\}$. The observations we will have to work with will probably be contaminated by Gaussian noise, of standard deviation σ. This observed series is denoted by $\{y(t)\}$.

Step 1

We have our inferred distribution of the process $L(t)$ as

$$\begin{bmatrix} \Pr[L(t-1)=1|\mathbf{y}_{t-1}] \\ \Pr[L(t-1)=2|\mathbf{y}_{t-1}] \\ \vdots \\ \Pr(L(t-1)=2N_{(L)}|\mathbf{y}_{t-1}) \end{bmatrix}$$

At time $t=0$ we take

$$\begin{bmatrix} \Pr(L(0)=1) \\ \Pr(L(0)=2) \\ \vdots \\ \Pr(L(0)=2N_{(L)}) \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{2N_{(L)}} \end{bmatrix}$$

Then we apply the transition matrix to obtain the expected distribution for time t.

$$\begin{bmatrix} \Pr[L(t)=1|\mathbf{y}_{t-1}] \\ \Pr[L(t)=2|\mathbf{y}_{t-1}] \\ \vdots \\ \Pr[L(t)=2N_{(L)}|\mathbf{y}_{t-1}] \end{bmatrix}' = \begin{bmatrix} \Pr[L(t-1)=1|\mathbf{y}_{t-1}] \\ \Pr[L(t-1)=2|\mathbf{y}_{t-1}] \\ \vdots \\ \Pr[L(t-1)=2N_{(L)}|\mathbf{y}_{t-1}] \end{bmatrix}' \times T$$

Step 2

The next step is similar to the Basic Filter but does not have to incorporate AR noise. We evaluate

$$f(y(t)|L(t)=i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y(t) - \frac{2i-1}{2N_{(L)}}\right)^2\right]$$

This is then included in the process by mulitplying the cells of the output of Step1 by the appropriate probability.

$$\begin{bmatrix} f(y(t), L(t) = 1|\mathbf{y}_{t-1}) \\ f(y(t), L(t) = 2|\mathbf{y}_{t-1}) \\ \vdots \\ f(y(t), L(t) = 2N_{(L)}|\mathbf{y}_{t-1}) \end{bmatrix} = \begin{bmatrix} \Pr[L(t) = 1|\mathbf{y}_{t-1}).f(y(t)|L(t) = 1] \\ \Pr[L(t) = 2|\mathbf{y}_{t-1}).f(y(t)|L(t) = 2] \\ \vdots \\ \Pr[L(t) = 2N_{(L)}|\mathbf{y}_{t-1}).f(y(t)|L(t) = 2N_{(L)}] \end{bmatrix}$$

Step3

From here we can infer that

$$f(y(t)|\mathbf{y}_{t-1}) = \sum_{i=1}^{2N_{(L)}} f(y(t), L(t) = i|\mathbf{y}_{t-1})$$

and

$$\begin{bmatrix} \Pr[L(t) = 1|\mathbf{y}_t) \\ \Pr[L(t) = 2|\mathbf{y}_t) \\ \vdots \\ \Pr[L(t) = 2N_{(L)}|\mathbf{y}_t) \end{bmatrix} = \frac{1}{f(y(t)|\mathbf{y}_{t-1})} \times \begin{bmatrix} f(y(t), L(t) = 1|\mathbf{y}_{t-1}) \\ f(y(t), L(t) = 2|\mathbf{y}_{t-1}) \\ \vdots \\ f(y(t), L(t) = 2N_{(L)}|\mathbf{y}_{t-1}) \end{bmatrix}$$

Step 4

Finally we can evaluate our likelihood in the same way as before, using

$$\begin{aligned} L(\Theta) &= \prod_{t=1}^{T} f(y(t)|\mathbf{y}_{t-1}) \\ l(\Theta) &= \sum_{t=1}^{T} \ln(f(y(t)|\mathbf{y}_{t-1})) \end{aligned}$$

It is also possible to easily convert back to produce inference on the distribution between regimes of the $S(t)$ process, since:

$$\begin{bmatrix} \Pr[S(t) = 0|\mathbf{y}_t] \\ \Pr[S(t) = 1|\mathbf{y}_t] \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N_{(L)}} \Pr[L(t) = i|\mathbf{y}_t) \\ \sum_{i=N_{(L)}+1}^{2N_{(L)}} \Pr[L(t) = i|\mathbf{y}_t) \end{bmatrix}$$

This method does allow us to fit the full model, including Gaussian noise. For reasonably large $N_{(L)}$ the approximation is good enough to obtain results. The method seems more limited by the practicalities of fitting the model than by the approximation required. In the next chapter a study is made of the effectiveness of this method and we provide practical guidelines to its limitations.

5.4 Model Fitting using Moments

Although ML is probably the more suitable of the two options when fitting time series models, the estimation problem is of more general interest. As the stationary distribution of the process is *Beta* we could consider the observations from a stationary series as approximating a random sample from the Beta distribution. So we are also considering the closely related problem of parameter estimation for the Beta distribution. Furthermore we are concerned with the estimation problem when the original sample is contaminated with noise. Although the use of the method of moments for the Beta distribution can be traced back to 1895 there is not much to be found on this noisy sample problem in the literature and is therefore worthy of consideration. The estimation problem is trivial if the levels l_0 and l_1 are known so we shall start with the 4 parameter case.

Note that the stationary distribution of the Filtered Markov model with parameters α, β and τ has stationary distribution $Beta(\alpha\tau, \beta\tau)$, and so if these methods were applied to that model then the parameters of interest would be $a = \alpha\tau$ and $b = \beta\tau$ rather than α and β. We shall continue to use the parameter θ to simplify notation, where

$$\theta = a + b$$

It should also be made clear that if we were to apply these methods to a time series we would require that the observations were obtained from a stationary FM process (to ensure convergence to the Beta distribution) and depend on the series being a good approximation to a sample of independent $Beta(a, b)$ variates.

5.4.1 Case 1: Four Parameters(a, b, l_0, l_1)

The solution to this four parameter problem can be found in Elderton & Johnson (1969) and requires the estimation of the first four moments, which may have straightforward although lengthy expressions. Although the two levels of the process l_0 and l_1 are natural parameters of the stochastic model it is often simpler to consider scale (c) and shift (d). If we choose to reparameterise the distribution we find the moments take a much simpler form, and so we define

$$d = l_0 \tag{5.1}$$
$$c = l_1 - l_0 \tag{5.2}$$

This allows us to represent the mean (μ) and variance (v) more easily

$$\mu = c\frac{a}{\theta} + d \tag{5.3}$$
$$v = c^2\frac{ab}{\theta^2(\theta+1)} \tag{5.4}$$

We can also write the 3^{rd} and 4^{th} central moments in similar form, greatly simplifying the process of finding the higher central moments.

$$E[(X-\mu)^3] = \frac{2ab(b-a)}{\theta^3(\theta+1)(\theta+2)}$$
$$E[(X-\mu)^4] = \frac{3ab[ab\theta + 2(\theta^2 - 3ab)]}{\theta^4(\theta+1)(\theta+2)(\theta+3)}$$

In particular we will also require estimates of the coefficients of Skewness and Kurtosis, γ and κ (the estimates will be denoted by $\hat{\gamma}$ and $\hat{\kappa}$). We know that these two depend only on a and b since

$$\frac{E[(X-\mu)^3]^2}{v^3} = \gamma^2 = \frac{4(\theta+1)(b-a)^2}{ab(\theta+2)^2} \tag{5.5}$$
$$\frac{E[(X-\mu)^4]}{v^2} = \kappa = \frac{3(ab\theta + 2(\theta^2-3ab))(\theta+1)}{ab(\theta+2)(\theta+3)} \tag{5.6}$$

Solving (5.5) and (5.6) gives us an estimator for θ

$$\hat{\theta} = \frac{6(\hat{\kappa} - 1 - \hat{\gamma}^2)}{3\hat{\gamma}^2 + 6 - 2\hat{\kappa}}$$

And then substitution of $\hat{\theta}$ into (5.6) leads us to

$$\hat{ab} = \frac{6\hat{\theta}^2(\hat{\theta}+1)}{\hat{\theta}^2(\hat{\kappa}-3) + (\hat{\kappa}-3)(5\hat{\theta}+6)}$$

Given estimates of both $a.b$ and $a+b$ (namely $\hat{\theta}$) we can obtain two possible solutions for a. From the sign of the Skewness we can tell whether $a > b$ (or vice versa), allowing us to determine the values of a and b individually. Substitution of the estimates of a, b and the variance (v) into (5.4) gives us an estimate for c. This can then be used with (5.3) to solve for d, giving us estimates of all the parameters.

5.4.2 Case 2: Four Parameters($a = b, l_0, l_1, \sigma$)

We already have clear expressions for the moments of the uncontaminated Beta distribution in 5.3 to 5.6. In this case we are now faced with a new problem, that of incorporating Gaussian noise (of standard deviation σ) into our moments. The resultant expressions make a simple solution of the form we saw in Case 1 unlikely and force us to look for alternative approaches. One such solution, proposed in Jalali (2003b), utilises cumulants to circumvent this problem. The advantage they possess is that the higher order (3^{rd} order or higher) cumulants are unaffected by Gaussian noise. In this case we are concerned with the symmetric case, for which all odd moments of higher order are equal to zero. With four parameters to estimate we will need four non-zero cumulants and will therefore use the 1^{st}, 2^{nd}, 4^{th} and 6^{th}. We shall work with the scale and shift parameters c and d, as defined in (5.1) and (5.2). We shall denote the i^{th} cumulant by K_i, which have the following expressions:

$$K_1 = \frac{c}{2} + d \tag{5.7}$$

$$K_2 = \frac{c^2}{4(\theta+1)} + \sigma^2 \tag{5.8}$$

$$K_4 = -\frac{3c^4}{8(\theta+1)^2(\theta+3)} \tag{5.9}$$

$$K_6 = \frac{15c^6}{4(\theta+1)^3(\theta+3)(\theta+5)} \tag{5.10}$$

We still have four unknowns and four cumulants. We can remove the scale coefficient c by taking ratios of cumulants. For instance using (5.9) and (5.10) we can obtain an expression in θ only.

$$\frac{K_6^2}{K_4^3} = -\frac{800(\theta+3)}{3(\theta+5)^2} \tag{5.11}$$

Now

$$\frac{d}{d\theta}\left[\frac{\theta+3}{(\theta+5)^2}\right] \leqslant 0 \qquad for\ all\ \theta \geqslant 0 \tag{5.12}$$

So the right hand side of (5.11) is monotonically increasing with θ. Also

$$-32 \leqslant \frac{K_6^2}{K_4^3} < 0 \tag{5.13}$$

This gives us a method for finding a unique estimate for θ. After finding $\hat{\theta}$ through solving (5.11) we can obtain the estimate of the scaling factor c from (5.9) thus

$$\hat{c} = \left[\frac{K_4 8(\hat{\theta}+1)^2(\hat{\theta}+3)}{3}\right]^{\frac{1}{4}}$$

And finally we find the variance of the noise σ^2 by substitution into (5.8)

$$\hat{\sigma}^2 = K_2 - \frac{\hat{c}^2}{4(\hat{\theta}+1)}$$

The levels are easily found from here.

The major problem facing any method using cumulants is that of obtaining good estimates from the data. An examination of the practical limits for the use of this method have been made in the next chapter.

5.4.3 Case 3: 5 Parameters (a, b, l_0, l_1, σ)

Finally we find ourselves facing the full set of parameters to estimate. In working with moments this is the largest number of parameters we shall have to worry about. This process settles down very quickly to the stationary distribution, becoming independent of the initial state and so, for all but the shortest samples, we can safely ignore this. We are also not concerned with the time-scaling factor τ directly here as our parameters of interest are $a = \alpha\tau$ and $b = \beta\tau$.

The method that has been developed for the asymmetric case of this problem does not utilise cumulants as was done for the symmetric case. This was motivated in part by the increased complexity of the expressions involved but also due to the difficulty in obtaining good estimates for the higher, odd cumulants. These generally are of very small magnitude and therefore the consequences of any inaccuracy can be quite significant. Instead a method, proposed in Jalali (2003d), makes use of the expressions for the moments, adjusted to incorporate the effects of the observation error (Gaussian noise).

The key innovation is to consider the (standardised central) sample moments of the contaminated observations $\{y(t)\}$ and obtain estimates of those of the uncontaminated $\{x(t)\}$. Once we have done this the problem reduces to that in Case 1. We shall differentiate between the two sets of moments by denoting them as $\mu', v', \gamma', \kappa', \eta'$ (contaminated) and $\mu, v, \gamma, \kappa, \eta$ (uncontaminated) respectively. We begin with the central moments of the Beta distribution. They are given below, up to the 5^{th} order and are written in terms of the standardised central moments of an uncontaminated variable.

$$\begin{aligned} E[X] &= \mu \\ E[(X-\mu)^2] &= v \\ E[(X-\mu)^3] &= \gamma v^{\frac{3}{2}} \\ E[(X-\mu)^4] &= kv^2 \\ E[(X-\mu)^5] &= \phi v^{\frac{5}{2}} \end{aligned}$$

We can then introduce the observation error $\epsilon \tilde{} N(0, \sigma^2)$ by substituting $Y = X + \epsilon$ into $E[(Y-\mu)^n]$. This allow us to express the moments of the contaminated sample ($E[(Y-\mu')^n]$)

in terms of the uncontaminated standardised central moments. These are given below

$$E[Y] = \mu' = \mu \tag{5.14}$$
$$E[(Y-\mu)^2] = v' = v + \sigma^2 \tag{5.15}$$
$$E[(Y-\mu)^3] = \gamma v'^{\frac{3}{2}} = \gamma v^{\frac{3}{2}} \tag{5.16}$$
$$E[(Y-\mu)^4] = \kappa v^2 = \kappa v^2 + 6v\sigma^2 + 3\sigma^4 \tag{5.17}$$
$$E[(Y-\mu)^5] = \phi v'^{\frac{5}{2}} = \phi v^{\frac{5}{2}} + 10\sigma^2\gamma v^{\frac{3}{2}} \tag{5.18}$$

From these we can obtain the standardised central moments of orders 3 to 5

$$\gamma' = \gamma\left(\frac{v}{v'}\right)^{\frac{3}{2}} \tag{5.19}$$
$$\kappa' = \kappa\left(\frac{v}{v'}\right)^2 + 6\sigma^2\left(\frac{v}{v'^2}\right) + \frac{3\sigma^4}{v'^2} \tag{5.20}$$
$$\phi' = \phi\left(\frac{v}{v'}\right)^{\frac{5}{2}} + 10\gamma\left(\frac{v}{v'}\right)^{\frac{3}{2}}\frac{\sigma^2}{v'} \tag{5.21}$$

We can then solve for γ, κ, η in terms of γ', κ', η'.

$$\gamma = \left(\frac{v'}{v}\right)^{\frac{3}{2}}\gamma' \tag{5.22}$$
$$\kappa = \left(\frac{v'}{v}\right)^2\kappa' - 6\left(\frac{v'}{v}-1\right) - 3\left(\frac{v'}{v}-1\right)^2 \tag{5.23}$$
$$\phi = \left(\frac{v'}{v}\right)^{\frac{5}{2}}\phi' - 10\left(\frac{v'}{v}\right)^{\frac{3}{2}}\left(\frac{v'}{v}-1\right)\gamma' \tag{5.24}$$

To simplify the notation, and reduce this equation to one parameter, we can now introduce

$$x^2 = \frac{v'}{v}$$

Which simplifies (5.22)-(5.24) to

$$\gamma = x^3\gamma' \tag{5.25}$$
$$\kappa = x^4\kappa' - 6(x^2-1) - 3(x^2-1)^2 \tag{5.26}$$
$$\phi = x5\phi' - 10x^3(x^2-1)\gamma' \tag{5.27}$$

Each of these three estimators on the left hand side of (5.25)-(5.27) is expressible in terms of a and b with x the only variable on the right hand side.

It is possible, at some length, to reduce this to a cubic equation in x^2 with coefficients provided by functions of γ', κ', ϕ'.

$$x^6\phi' - 15(x^2-1)x^4\kappa' + 45(x^2-1)^2 + 30(x^2-1)^3 \tag{5.28}$$

We can use the original sample to provide us with estimates of γ', κ', ϕ', and this equation can then be solved for $\hat{x}^2$. Once we have solved (5.28) and have a measure of the noise level we can use (5.22)-(5.24) to obtain estimates of the noise free parameters $\hat{\mu}, \hat{v}, \hat{\gamma}, \hat{\kappa}, \hat{\phi}$. Once we have these the problem reverts to that we dealt with in Case 1 of four parameters without noise.

It is also possible to obtain a result for non-Gaussian noise as long as the distribution of the noise is symmetric. All we require in addition is the Kurtosis of the distribution of the noise, which is represented by λ. In this case we obtain a more general version of (5.28) which is given below in (5.29)

$$\begin{aligned} 0 &= 2(x^4\kappa' - 6(x^2-1) - \lambda(x^2-1)^2)^2\gamma' \\ &\quad +(x^4\kappa' - 6(x^2-1) - \lambda(x^2-1)^2)(x^2\phi' - 10(x^2-1)\gamma' + 14\gamma') \\ &\quad -6x^6\gamma'^3 - 3x^6\gamma'^2(y\phi' - 10(x^2-1)\gamma') - 9(x^2\phi' - 10(x^2-1)\gamma') \end{aligned} \tag{5.29}$$

Which, with the inclusion of a suitable value for λ can be solved as before giving us our uncontaminated estimates, after which the distribution of the noise becomes irrelevant.

Case 5: The Symmetric Beta$(a = b, l_0, l_1, \sigma)$

The obvious weakness of this method stems from the fact that it relies on information derived from the third and 5th moments. In the case of a symmetric Beta distribution these will be zero and any deviation from zero is due to estimation error. So for this case we need an additional even moment, the next one being the 6^{th} central moment. This normalised central moment ψ is defined as

$$\begin{aligned} \psi &= \frac{E[(Y-\mu)^6]}{E[(Y-\mu)^2]^3} \\ or\ \psi v^3 &= E[(Y-\mu)^6] \end{aligned} \tag{5.30}$$

If we use the substitution $a = b = \frac{\theta}{2}$ the right hand side of (5.30) can be expressed entirely in terms of θ. Some algebraic manipulation eventually gives us

$$\psi v^3 = \frac{15}{64(\theta+1)(\theta+3)(\theta+5)} \quad ,$$

and using the fact that $v = \frac{1}{4(\theta+1)}$ we obtain the following expression for ψ

$$\psi = \frac{15(\theta+1)^2}{(\theta+3)(\theta+5)} \tag{5.31}$$

Another result we will require is that in the symmetric case (5.6) reduces to (5.32)

$$\kappa = \frac{3(\theta+1)}{(\theta+3)} \tag{5.32}$$

Equation (5.32), together with (5.31), gives a relationship between κ and ψ that holds for symmetric Beta distributions. That expression is given below in (5.33)

$$\psi(6-\kappa) = 5\kappa^2 \tag{5.33}$$

When noise ($\epsilon \in N(0,\sigma^2)$) is added to the 6^{th} moment we obtain a new expression.

$$E[((Y+\epsilon)-\mu)^6] = E[(Y-\mu)^6] + 15E[(Y-\mu)^4]\sigma^2 + 45E[(Y-\mu)^2]\sigma^4 + 15\sigma^6$$

Substituting in for the expectations we find

$$\begin{aligned} \psi' v'^3 &= \psi v^3 + 15\kappa v^2\sigma^2 + 45v\sigma^4 + 15\sigma^6 \\ \psi &= \left(\frac{v'}{v}\right)\psi' - 15\kappa\left(\frac{v'-v}{v}\right) - 45\left(\frac{v'-v}{v}\right)^2 - 15\left(\frac{v'-v}{v}\right)^3 \end{aligned}$$

But we also know from (5.23) that

$$\kappa = \left(\frac{v'}{v}\right)^2\kappa' - 6\left(\frac{v'}{v}-1\right) - 3\left(\frac{v'}{v}-1\right)^2$$

If we define

$$\zeta = \left(\frac{v'}{v}-1\right) \tag{5.34}$$

Then we substitute ζ into (5.31) and (5.23) to obtain the following

$$\psi = (1+\zeta)^3\psi' - 15(1+\zeta)^2\kappa' + 45\zeta + 30\zeta^3 \tag{5.35}$$

$$\kappa = (1+\zeta)^2\kappa' - 6\zeta - 3\zeta^2 \tag{5.36}$$

If we then substitute (5.35) and (5.36) in the identity (5.33) we obtain the desired equation for ζ

$$\begin{aligned} &\left[(1+\zeta)^3\psi' - 15(1+\zeta)^2\kappa' + 450\zeta + 30\zeta^3\right]\left(6-(1+\zeta)^2\kappa' + 6\zeta + 3\zeta^2\right) \quad (5.37) \\ =\ &5\left[(1+\zeta)^2\kappa' - 6\zeta - 3\zeta^2\right]^2 \end{aligned}$$

All we need to do is obtain the noisy sample moments $\hat{\psi}, \hat{\kappa}$ and find the positive real roots of (5.37). This gives us our estimates of the level of noise $\hat{\zeta}$ since

$$\zeta = \frac{\sigma^2}{v}$$

Then we can easily obtain $\hat{\kappa}$ from (5.36) and then $\hat{\theta}$ from (5.32). Our estimate of $\hat{\theta}$ will give us estimates of a and b (since $a = b = \frac{\theta}{2}$) and hence all the noise-free moments and from here it is simple to solve for the parameters of the model.

5.5 Summary and Conclusions

We started working with ML with the simplest model we could, which was an ideal, noise-free process. We immediately found problems with this simplest of methods that forewarned us of the problems ahead. Quite soon it became clear that, although there was no problem with identifiability, this kind of approach was unlikely to be of much use when more parameters were added (in particular τ). By treating the continuous process as a discrete one we were able to apply other Likelihood methods, using a version of Hamilton's filter. Using this method we were able to fit the model quite effectively when noise levels were low to moderate although some parameters proved easier to estimate than others. A thorough examination of the quality of the estimates obtained for different levels of noise is given in the next Chapter.

The Method of Moments was always unlikely to provide the most powerful method for working with time series models. This is due to the large amount of very useful information that is discarded in treating the process as a sample from a Beta distribution with the observations obscured by the presence of Gaussian noise. Due to the lack of any available methods for parameter estimation for a noisy Beta sample it was worth pursuing despite its limited suitability for working with time series modelling. What we obtained was a method that could be applied to time series but could only provide answers when noise levels are very low.

The Cumulants approach faced a similarly uphill struggle. The loss of useful information from the series makes this approach of theoretical interest only. What we found was that progress could be made but the difficulty of obtaining good estimates of the cumulants obstructed this method. Very large sample sizes or very little noise would be needed for this approach to work.

In all our approaches to parameter estimation involving the Beta distribution we have found similar problems. The presence of even a small amount of uncertainty in the observations complicates the estimation process considerably. The methods collected here have been tested to determine how robust they are and identify the limits of their practical use. The results of these experiments are summarised in the next chapter.

Chapter 6

Applying the Methods to Simulated Data

In the previous chapter we proposed a method for parameter estimation when working with the FM model and several others for a more general problem of parameter estimation for a contaminated Beta sample. In this chapter we apply the methods to simulated data to study their performance.

6.1 Model Testing

We have settled upon a likely method for fitting the FM model and proposed several methods for obtaining inference about a sample from a Beta distribution contaminated by Gaussian noise. It now remains to observe these methods and their performance on simulated data. This will enable us to draw up guidelines for their use. For each method we wish to measure reliability (how often we can rely on it to produce a meaningful answer) and effectiveness (how good are the estimates it produces when the method does not fail). We shall retain the order in which we introduced the methods, and due to the number of tables we shall provide a page index.

6.2 Approximate Likelihood

This method, proposed in Section 5.3.4, utilises a version of the Filter algorithm introduced in Hamilton (1989) to obtain a likelihood value for a data set and a given set of parameters. We use a series of observations from a simulated FM process to represent our data. Our fitting is done by an MCMC algorithm known as a random walk Metropolis (see Section 3.3) which enables us to measure the shape of the likelihood space around the MLE. There are many ways we could have presented the measurements we collected in this and the next section. We give the estimates as percentiles of the sample estimates rather than a point estimate and measure of deviation. This was motivated by the fact that for only some of the parameters could the marginal distribution be said to be normally distributed.

In order to present the information collected in an efficient and readable manner the tables are constructed in the following way. There are two sets of tables, the first set

present the results when dealing with a sample of 100 and the second set a sample of 1000. In all cases we use levels $l_0 = 0$, $l_1 = 1$ and $\tau = 1$. Each set consists of 7 tables for different noise levels $\sigma = \{0.1, 0.2, ..., 0.7\}$. For each line, indexed by (α, β, σ), of the table upper (95%) and lower (5%) percentiles are given for the first of the two intensities (α) and the first of the two levels (l_0). The confidence intervals for the second parameter and second level can be found in the line indexed by the model (β, α, σ). When we use this approach to study the upper level parameter we must remember that the confidence interval will be reflected and centred around 1 rather than 0. Limits for τ and σ can be found in both lines of the table. Each set of values (α, β, σ) was tested only once, due to the large time requirements, with the results being split between these two cases. The obvious exception to this is the symmetric case which is presented in one line of the table only.

6.2.1 MCMC Methodology

The walk was initiated at the 'true' parameter values used to generated the series and used a Normally distributed proposal distribution. A walk of 10,000 steps was used with a fixed burn-in period of 1,500 steps. This standard burn-in period for all runs was chosen by using the required burn-in period for the single run with slowest convergence. The lengths of the walk and burn-in period were necessary to ensure both convergence of the sampler to the posterior distribution and sufficient measurements to obtain a good estimate of the confidence interval for each parameter. The specification of the proposal distribution and the burn-in period were chosen after careful inspection of the output of the algorithm. The routine was written in C++ to maximise performance but the time taken to measure a single set of parameters was still measured in hours. In the example below we show the convergence plots and marginal posterior distributions for a sample case.

Example 11

In this example we take one of the cases shown in the first table, that of $\alpha = 0.1, \beta = 0.3, \sigma = 0.1$. We also take $l_0 = 0, l_1 = 1, \tau = 1$ and generate a series of length 100 When constructing the tables below we initialised the MCMC algorithm at the 'true' parameter values. In this case we do not do this, so as to demonstrate the convergence. The algorithm is instead initialised at $\alpha = 0.5, \beta = 0.5, \sigma = 0.5, l_0 = 0, l_1 = 1, \tau = 1$. We run the algorithm until we have accepted 10,000 steps and the burn-in period is chosen by eye. In this case we chose a burn-in period of around $1,500$ steps. The convergence plots are given in Figures 6.1 to 6.7 while the plots of the estimated marginal posterior distributions are shown in Figures 6.8 to 6.13. Marked on the convergence plots are the end of the burn-in period and the 'true' value used to generate the series.

It is noticeable that the quality of the inference is highly variable from parameter to parameter. In some cases the estimate is both accurate and robust. This is the case in estimating the noise level, σ. In other cases, such as the time scaling parameter τ, the

confidence interval is very broad. These estimation problems seem to be due to the quite subtle influence of the parameter tau. Even with full knowledge of the levels of the process (l_i) the switching behaviour is easily drowned out by the white noise. Where there is any doubt as to the precise location of one of the levels, such as in a strongly asymmetric case, little can be learned about tau.

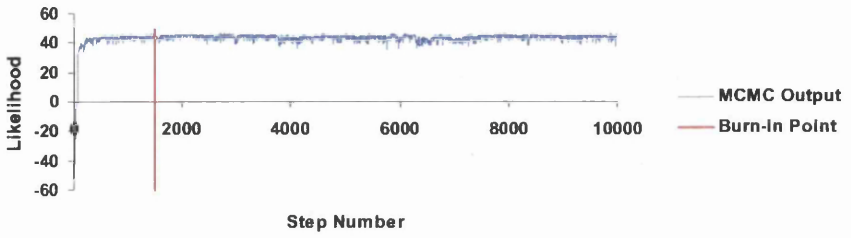

Figure 6.1: MCMC convergence plot for the log-likelihood of the series.

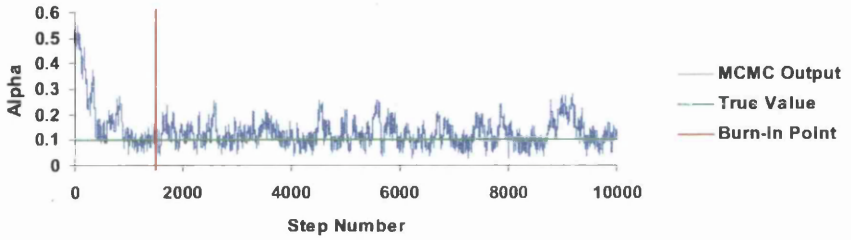

Figure 6.2: MCMC convergence plot for alpha.

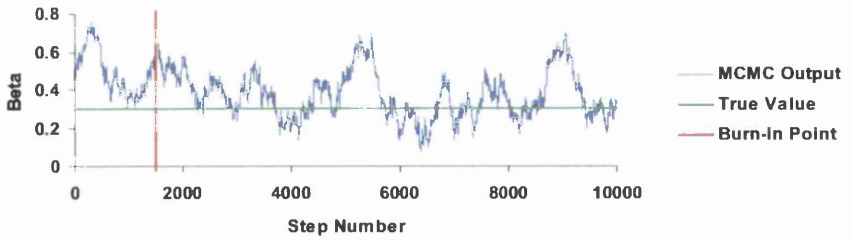

Figure 6.3: MCMC convergence plot for beta.

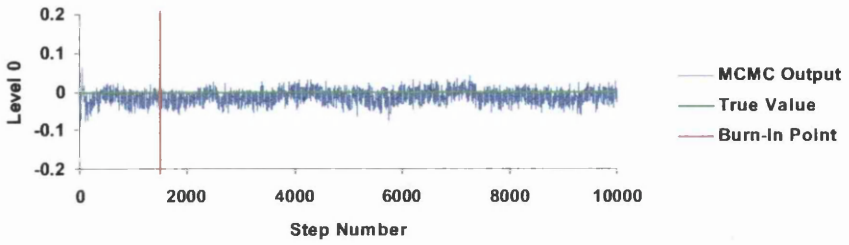

Figure 6.4: MCMC convergence plot for the level of the lower regime, l_0.

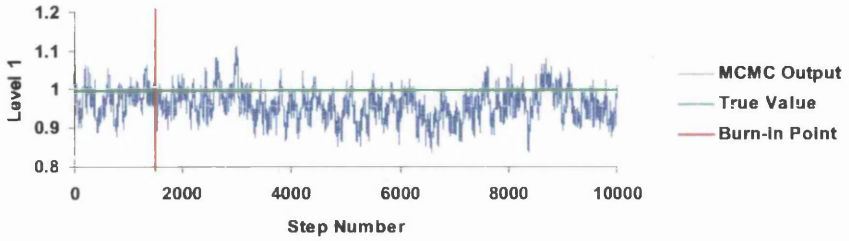

Figure 6.5: MCMC convergence plot for the level of the upper regime, l_1

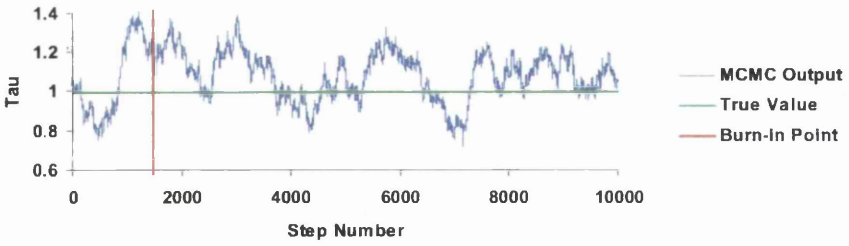

Figure 6.6: MCMC convergence plot for Tau.

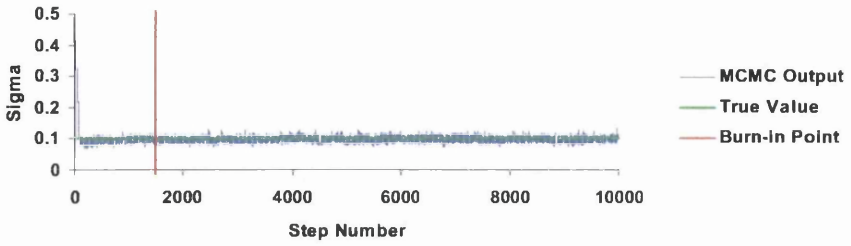

Figure 6.7: MCMC convergence plot for Sigma.

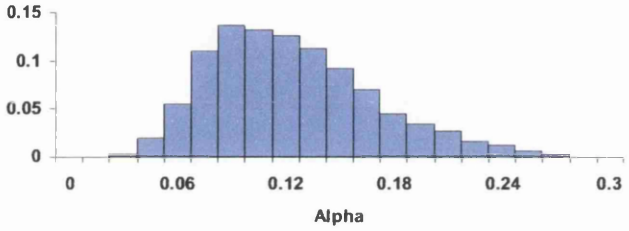

Figure 6.8: A plot of the estimate of the marginal posterior distribution for alpha.

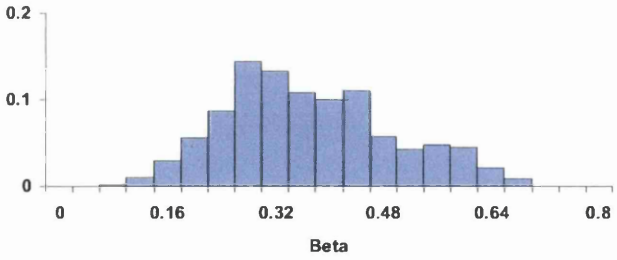

Figure 6.9: A plot of the estimate of the marginal posterior distribution for beta.

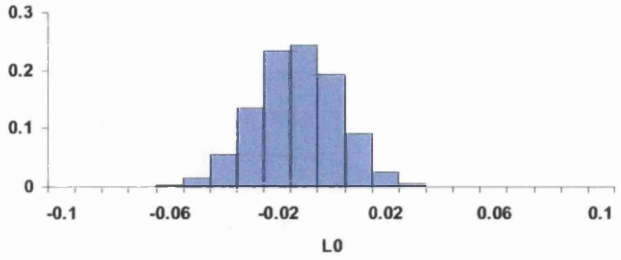

Figure 6.10: A plot of the estimate of the marginal posterior distribution for the level of the lower regime, l_0.

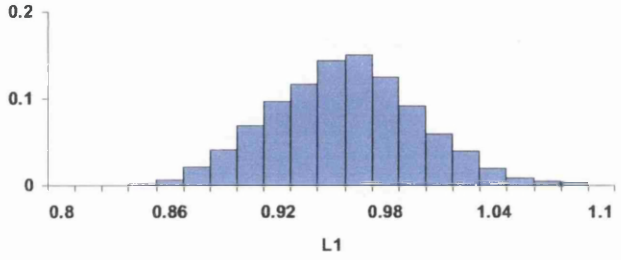

Figure 6.11: A plot of the estimate of the marginal posterior distribution for level of the upper regime, l_1.

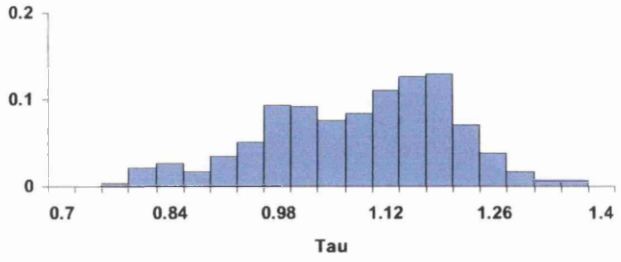

Figure 6.12: A plot of the estimate of the marginal posterior distribution for tau.

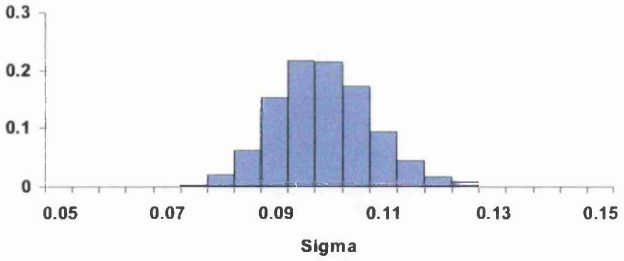

Figure 6.13: A plot of the estimate of the marginal posterior distribution for sigma.

Estimator Percentile (Sample Size 100)

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.1	0.0121	0.1006	-0.0072	0.0504	0.5021	1.0364	0.0950	0.1232
0.1	0.3	0.1	0.0487	0.1791	-0.0163	0.0295	0.8329	1.2261	0.0867	0.1171
0.1	0.5	0.1	0.0370	0.1675	-0.0295	0.0178	0.6582	1.0151	0.0794	0.1083
0.1	0.7	0.1	0.0517	0.1720	-0.0327	0.0063	0.7537	1.0403	0.0826	0.1083
0.3	0.1	0.1	0.2663	0.7450	0.9308	1.0824	0.8329	1.2261	0.0867	0.1171
0.5	0.1	0.1	0.2022	0.7001	0.8468	1.0124	0.6582	1.0151	0.0794	0.1083
0.7	0.1	0.1	0.2308	0.6824	0.8400	1.0160	0.7537	1.0403	0.0826	0.1083

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.2	0.0273	0.1799	-0.0563	0.0554	0.2816	0.6310	0.1699	0.2179
0.1	0.3	0.2	0.0501	0.2217	-0.0376	0.0796	0.7789	1.9807	0.1864	0.2430
0.1	0.5	0.2	0.0434	0.2241	-0.0554	0.0379	0.0534	0.8892	0.1616	0.2161
0.1	0.7	0.2	0.0011	0.0570	-0.0158	0.0591	0.0240	1.2682	0.1845	0.2391
0.3	0.1	0.2	0.1730	0.8462	0.9171	1.2398	0.7789	1.9807	0.1864	0.2430
0.5	0.1	0.2	0.1443	0.5029	0.7281	0.9252	0.0534	0.8892	0.1616	0.2161
0.7	0.1	0.2	0.0210	0.3378	0.2789	0.9268	0.0240	1.2682	0.1845	0.2391

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.3	0.0379	0.1870	-0.0657	0.0968	0.1498	0.7081	0.2413	0.3053
0.1	0.3	0.3	0.0236	0.1609	-0.0905	0.0252	0.0483	1.2406	0.2342	0.3114
0.1	0.5	0.3	0.0149	0.1542	-0.0781	0.0656	0.0236	0.5139	0.2834	0.3760
0.1	0.7	0.3	0.0418	0.1814	-0.0918	0.0166	0.0342	0.5991	0.2242	0.2940
0.3	0.1	0.3	0.0937	0.4404	0.8682	1.2556	0.0483	1.2406	0.2342	0.3114
0.5	0.1	0.3	0.0389	0.4359	0.4512	0.8061	0.0236	0.5139	0.2834	0.3760
0.7	0.1	0.3	0.1392	0.5428	0.6989	1.1075	0.0342	0.5991	0.2242	0.2940

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.4	0.0594	0.7767	-0.7420	0.1252	0.0288	0.4442	0.3765	0.4970
0.1	0.3	0.4	0.0805	0.7312	-0.2375	0.0714	0.0109	0.4174	0.3492	0.4727
0.1	0.5	0.4	0.0054	0.0945	-0.0919	0.0582	0.0677	1.1545	0.3628	0.4528
0.1	0.7	0.4	0.0117	0.4597	-0.0658	0.1834	0.0094	0.7292	0.3444	0.4576
0.3	0.1	0.4	0.1284	0.5186	0.5467	0.8594	0.0109	0.4174	0.3492	0.4727
0.5	0.1	0.4	0.0419	0.3579	0.4366	0.8064	0.0677	1.1545	0.3628	0.4528
0.7	0.1	0.4	0.0415	0.7599	0.2027	1.1513	0.0094	0.7292	0.3444	0.4576

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.5	0.0435	0.2965	-0.1662	0.1269	0.1043	0.7837	0.4009	0.5284
0.1	0.3	0.5	0.0345	0.4568	-0.4216	0.0401	0.0690	1.3012	0.4113	0.5749
0.1	0.5	0.5	0.0149	0.3311	0.0591	0.3070	0.0172	0.4921	0.3653	0.4908
0.1	0.7	0.5	0.0601	0.4458	-0.4029	-0.0115	0.0170	0.8988	0.4236	0.5681
0.3	0.1	0.5	0.0060	0.7066	0.1294	0.9350	0.0690	1.3012	0.4113	0.5749
0.5	0.1	0.5	0.0367	0.9687	0.2466	1.2203	0.0172	0.4921	0.3653	0.4908
0.7	0.1	0.5	0.0184	0.4001	0.2469	0.6933	0.0170	0.8988	0.4236	0.5681

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.6	0.0818	0.6227	-0.7518	-0.1848	0.0112	0.7723	0.5328	0.7222
0.1	0.3	0.6	0.1166	0.7114	-0.3652	0.0207	0.0240	0.6283	0.5085	0.7294
0.1	0.5	0.6	0.0201	0.3970	-0.6095	0.2116	0.0143	0.5615	0.5424	0.7135
0.1	0.7	0.6	0.0271	0.8756	-0.4428	0.0991	0.0232	0.8622	0.5288	0.6794
0.3	0.1	0.6	0.1698	0.7748	0.4784	1.0387	0.0240	0.6283	0.5085	0.7294
0.5	0.1	0.6	0.0044	0.2629	0.2217	0.6944	0.0143	0.5615	0.5424	0.7135
0.7	0.1	0.6	0.0158	0.9200	0.0276	0.5684	0.0232	0.8622	0.5288	0.6794

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.7	0.0844	0.5577	-0.3972	-0.0226	0.0313	0.6071	0.5320	0.6973
0.1	0.3	0.7	0.0226	0.3155	-1.7580	-0.2159	0.0121	0.6495	0.6298	0.7939
0.1	0.5	0.7	0.0240	0.9455	-2.6763	0.0014	0.0257	0.9637	0.6156	0.7719
0.1	0.7	0.7	0.0220	0.4541	-1.0303	0.0277	0.0357	0.8415	0.5737	0.7357
0.3	0.1	0.7	0.0008	0.0537	0.1195	0.3619	0.0121	0.6495	0.6298	0.7939
0.5	0.1	0.7	0.0021	0.3960	-0.0610	0.3598	0.0257	0.9637	0.6156	0.7719
0.7	0.1	0.7	0.0020	0.6299	-0.0067	0.8729	0.0357	0.8415	0.5737	0.7357

Estimator Percentile (Sample Size 1000)

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.1	0.0663	0.1221	-0.0229	-0.0005	0.8784	1.0023	0.0913	0.1020
0.1	0.3	0.1	0.0890	0.1334	-0.0160	0.0002	0.9270	1.0302	0.0904	0.1001
0.1	0.5	0.1	0.0752	0.1258	-0.0170	0.0028	0.8974	1.0692	0.0950	0.1075
0.1	0.7	0.1	0.0921	0.1439	-0.0076	0.0098	0.8608	1.0101	0.0882	0.0994
0.3	0.1	0.1	0.2812	0.4233	0.9802	1.0138	0.9270	1.0302	0.0904	0.1001
0.5	0.1	0.1	0.3902	0.5697	0.9566	1.0162	0.8974	1.0692	0.0950	0.1075
0.7	0.1	0.1	0.6084	0.8487	0.9356	1.0168	0.8608	1.0101	0.0882	0.0994

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.2	0.0708	0.1228	-0.0167	0.0163	0.8982	1.1512	0.1830	0.2013
0.1	0.3	0.2	0.0967	0.1705	-0.0241	0.0175	0.7413	1.0093	0.1930	0.2168
0.1	0.5	0.2	0.0763	0.1600	-0.0236	0.0164	0.6753	1.0563	0.1843	0.2094
0.1	0.7	0.2	0.0724	0.1331	-0.0248	0.0106	0.7809	1.3426	0.1833	0.2035
0.3	0.1	0.2	0.2533	0.4034	0.9300	1.0128	0.7413	1.0093	0.1930	0.2168
0.5	0.1	0.2	0.3441	0.6045	0.8956	1.0022	0.6753	1.0563	0.1843	0.2094
0.7	0.1	0.2	0.4266	0.6795	0.8326	0.9927	0.7809	1.3426	0.1833	0.2035

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.3	0.0496	0.0993	-0.0361	0.0248	0.8108	1.3296	0.2822	0.3111
0.1	0.3	0.3	0.0875	0.1887	-0.0455	0.0244	0.7630	1.4168	0.2649	0.2995
0.1	0.5	0.3	0.0747	0.2120	-0.0198	0.0402	0.4021	0.8992	0.2620	0.2976
0.1	0.7	0.3	0.0428	0.1186	-0.0319	0.0293	0.2422	0.8775	0.2920	0.3251
0.3	0.1	0.3	0.3014	0.5819	0.9839	1.1282	0.7630	1.4168	0.2649	0.2995
0.5	0.1	0.3	0.3651	0.8984	0.8115	1.0021	0.4021	0.8992	0.2620	0.2976
0.7	0.1	0.3	0.2941	0.6273	0.7223	0.9802	0.2422	0.8775	0.2920	0.3251

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.4	0.0785	0.1646	-0.0573	0.0367	0.8807	1.2265	0.3704	0.4106
0.1	0.3	0.4	0.0750	0.1711	-0.0628	0.0070	0.9448	0.6958	0.3536	0.3937
0.1	0.5	0.4	0.0752	0.2537	-0.0733	0.0354	0.7717	0.6918	0.3630	0.4116
0.1	0.7	0.4	0.0740	0.2676	-0.0993	0.0076	0.5764	1.1758	0.3659	0.4088
0.3	0.1	0.4	0.2312	0.4571	0.9448	1.1039	0.9448	0.6958	0.3536	0.3937
0.5	0.1	0.4	0.3236	0.8591	0.7717	1.0484	0.7717	0.6918	0.3630	0.4116
0.7	0.1	0.4	0.5353	0.9046	0.5764	0.9896	0.5764	1.1758	0.3659	0.4088

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.5	0.0605	0.1511	-0.0353	0.0894	0.2639	0.8950	0.4629	0.5137
0.1	0.3	0.5	0.0789	0.2181	-0.0904	0.0369	0.0589	0.7187	0.4619	0.5158
0.1	0.5	0.5	0.0342	0.2749	-0.0027	0.1030	0.0041	1.0111	0.4850	0.5383
0.1	0.7	0.5	0.0503	0.3109	-0.0785	0.0705	0.0120	0.8403	0.4691	0.5266
0.3	0.1	0.5	0.2294	0.5229	0.7398	1.0004	0.0589	0.7187	0.4619	0.5158
0.5	0.1	0.5	0.2608	0.8592	0.7181	1.1509	0.0041	1.0111	0.4850	0.5383
0.7	0.1	0.5	0.3249	0.9082	0.6278	1.0041	0.0120	0.8403	0.4691	0.5266

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.6	0.0550	0.1520	-0.0746	0.0887	0.0561	0.9395	0.5528	0.6119
0.1	0.3	0.6	0.0280	0.1165	-0.0649	0.0615	0.0101	0.9456	0.5651	0.6276
0.1	0.5	0.6	0.0272	0.2274	-0.1190	0.0231	0.0048	1.0507	0.5506	0.6193
0.1	0.7	0.6	0.0940	0.4577	-0.1504	0.0052	0.0115	0.8447	0.5488	0.6087
0.3	0.1	0.6	0.0857	0.2936	0.7002	0.9641	0.0101	0.9456	0.5651	0.6276
0.5	0.1	0.6	0.0606	0.5248	0.3610	0.8867	0.0048	1.0507	0.5506	0.6193
0.7	0.1	0.6	0.1597	0.8674	0.4231	0.8040	0.0115	0.8447	0.5488	0.6087

α	β	σ	$\hat{\alpha}_{0.05}$	$\hat{\alpha}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\tau}_{0.05}$	$\hat{\tau}_{0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.1	0.1	0.7	0.0448	0.1166	-0.1086	1.0664	0.0312	1.2101	0.6355	0.7013
0.1	0.3	0.7	0.0590	0.3917	-0.2087	0.8871	0.0092	1.2351	0.6621	0.7434
0.1	0.5	0.7	0.0354	0.2131	-0.0659	1.0912	0.0107	0.9345	0.6333	0.7001
0.1	0.7	0.7	0.0876	0.8916	-1.1267	0.7844	0.0066	1.0552	0.6385	0.7242
0.3	0.1	0.7	0.1293	0.4441	0.5060	0.9263	0.0092	1.2351	0.6621	0.7434
0.5	0.1	0.7	0.2371	0.7249	0.6420	1.0672	0.0107	0.9345	0.6333	0.7001
0.7	0.1	0.7	0.0007	0.6694	0.0331	1.4837	0.0066	1.0552	0.6385	0.7242

Summary

One notable difference between this method and those involving moments is the absence of any figure to represent the failure rate. Even when the series under examination proves problematic for the fitting method some kind of estimate is always obtained.

On examining these tables we notice that not all the parameters are estimated with equal precision. For instance the standard deviation of the noise σ tends to be estimated with a high degree of precision, even for high noise levels. This is true even when the precision of our estimates of other parameters is quite poor. This cannot be said for our estimates of the time scaling factor τ, for which our estimates tend to be very poor for all but the lowest noise levels. For the other parameters our estimate tend to be quite good until the noise levels reach $\sigma = 0.3$ or higher. In particular the estimates of the noise level are generally far better than the other estimates, even for a highly asymmetric or noisy model. The estimates are generally reasonably symmetric except for the most noisy and most asymmetric cases.

6.3 Methods of Moments

For these methods to work with a time series we would expect to require quite low noise levels. In practice we found that they were of little use when dealing with time series data except in the most ideal circumstances and we would not recommend them for this purpose. As a consequence of this we include an examination of these methods for their general interest in terms of obtaining inference about contaminated Beta samples rather than for their original purpose. For this reason we do not use a series from the FM model, which may display some correlation between consecutive values, and instead use a sample drawn directly from the appropriate Beta distribution.

6.3.1 Case 1: Four Parameter Asymmetric (a, b, l_0, l_1)

This algorithm is intended for use with a noise-free sample from the Beta distribution. We include here an appraisal of its performance when the sample is contaminated for the purpose of measuring the worth of the Five Parameter model (which attempts to incorporate noise into the inference). The confidence intervals are taken as the 5th and 95th percentiles of parameter estimates obtained from successful inference. Unsuccessful inference includes any case where there are non-real solutions (or negative values for certain parameters). The failure rate of the method for a set of parameters gives the proportion of the samples that resulted in a failed inference (on a scale of 0 to 1). Each set of parameters was tested 1000 times to produce these estimates, with the length of the samples generated being either 1000 or 10,000 as specified in the table.

Estimator Percentiles (Sample Size 1000) for $\sigma = 0$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.01	0.01	0.00	0.00	0.0049	0.0229	-0.0005	0.0004
0.01	0.10	0.00	0.00	0.0058	0.0173	-0.0003	0.0003
0.01	0.50	0.00	0.00	0.0054	0.0175	-0.0004	0.0004
0.01	1.00	0.00	0.00	0.0045	0.0177	-0.0003	0.0004
0.01	2.00	0.00	0.00	0.0039	0.0179	-0.0002	0.0004

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.10	0.01	0.00	0.00	0.0613	0.1875	-0.0041	0.0066
0.10	0.10	0.00	0.00	0.0778	0.1289	-0.0018	0.0021
0.10	0.50	0.00	0.00	0.0809	0.1241	-0.0018	0.0016
0.10	1.00	0.00	0.00	0.0803	0.1232	-0.0016	0.0017
0.10	2.00	0.00	0.00	0.0749	0.1261	-0.0014	0.0017

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.50	0.01	0.00	0.00	0.2970	0.8941	-0.0358	0.1238
0.50	0.10	0.00	0.00	0.4074	0.6121	-0.0149	0.0185
0.50	0.50	0.00	0.00	0.4356	0.5787	-0.0080	0.0084
0.50	1.00	0.00	0.00	0.4380	0.5707	-0.0066	0.0072
0.50	2.00	0.00	0.00	0.4298	0.5873	-0.0061	0.0058

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
1.00	0.01	0.00	0.00	0.5393	1.6390	-0.0753	0.3171
1.00	0.10	0.00	0.00	0.7999	1.2332	-0.0426	0.0638
1.00	0.50	0.00	0.00	0.8717	1.1398	-0.0227	0.0253
1.00	1.00	0.00	0.00	0.8750	1.1385	-0.0169	0.0173
1.00	2.00	0.00	0.00	0.8637	1.1532	-0.0133	0.0128

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
2.00	0.01	0.00	0.00	0.7524	2.6524	-0.1241	0.5989
2.00	0.10	0.00	0.00	1.4194	2.6398	-0.1770	0.1843
2.00	0.50	0.00	0.00	1.6754	2.4134	-0.0893	0.0702
2.00	1.00	0.00	0.00	1.7134	2.3543	-0.0545	0.0488
2.00	2.00	0.00	0.00	1.6883	2.3717	-0.0411	0.0327

Estimator Percentiles (Sample Size 1000) for $\sigma = 0.01$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.01	0.01	0.01	0.00	0.0060	0.0241	-0.0019	0.0002
0.01	0.10	0.01	0.00	0.0064	0.0180	-0.0012	0.0001
0.01	0.50	0.01	0.00	0.0072	0.0189	-0.0014	0.0000
0.01	1.00	0.01	0.00	0.0076	0.0228	-0.0016	-0.0001
0.01	2.00	0.01	0.04	0.0088	0.0341	-0.0022	-0.0003

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.10	0.01	0.01	0.00	0.0661	0.1930	-0.0091	0.0032
0.10	0.10	0.01	0.00	0.0795	0.1299	-0.0028	0.0013
0.10	0.50	0.01	0.00	0.0818	0.1246	-0.0025	0.0011
0.10	1.00	0.01	0.00	0.0824	0.1265	-0.0025	0.0010
0.10	2.00	0.01	0.00	0.0817	0.1332	-0.0026	0.0008

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.50	0.01	0.01	0.00	0.3329	1.0674	-0.0624	0.0683
0.50	0.10	0.01	0.00	0.4222	0.6291	-0.0179	0.0158
0.50	0.50	0.01	0.00	0.4462	0.5725	-0.0092	0.0084
0.50	1.00	0.01	0.00	0.4354	0.5769	-0.0078	0.0060
0.50	2.00	0.01	0.00	0.4353	0.5895	-0.0068	0.0051

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
1.00	0.01	0.01	0.00	0.6529	2.5509	-0.1866	0.2096
1.00	0.10	0.01	0.00	0.8183	1.2739	-0.0577	0.0505
1.00	0.50	0.01	0.00	0.8777	1.1565	-0.0241	0.0221
1.00	1.00	0.01	0.00	0.8824	1.1460	-0.0185	0.0163
1.00	2.00	0.01	0.00	0.8651	1.1647	-0.0144	0.0124

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
2.00	0.01	0.01	0.04	1.3020	10.7617	-0.8841	0.3252
2.00	0.10	0.01	0.00	1.4874	2.9255	-0.2167	0.1585
2.00	0.50	0.01	0.00	1.6731	2.4410	-0.0903	0.0702
2.00	1.00	0.01	0.00	1.6966	2.4035	-0.0609	0.0479
2.00	2.00	0.01	0.00	1.7082	2.4168	-0.0438	0.0334

Estimator Percentiles (Sample Size 1000) for $\sigma = 0.05$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.01	0.01	0.05	0.00	0.0271	0.0845	-0.0284	-0.0127
0.01	0.10	0.05	0.00	0.0285	0.0429	-0.0175	-0.0112
0.01	0.50	0.05	0.13	0.0496	0.1318	-0.0243	-0.0145
0.01	1.00	0.05	0.71	0.0776	0.1786	-0.0281	-0.0174
0.01	2.00	0.05	0.99	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.10	0.01	0.05	0.00	0.1725	0.7616	-0.1858	-0.0380
0.10	0.10	0.05	0.00	0.1129	0.1709	-0.0246	-0.0142
0.10	0.50	0.05	0.00	0.1213	0.1724	-0.0224	-0.0133
0.10	1.00	0.05	0.00	0.1487	0.2184	-0.0267	-0.0170
0.10	2.00	0.05	0.01	0.2405	0.4438	-0.0401	-0.0246

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
0.50	0.01	0.05	0.11	1.1770	18.0417	-4.4097	-0.2497
0.50	0.10	0.05	0.00	0.5613	0.8974	-0.0892	-0.0351
0.50	0.50	0.05	0.00	0.5064	0.6664	-0.0365	-0.0127
0.50	1.00	0.05	0.00	0.5220	0.6859	-0.0352	-0.0145
0.50	2.00	0.05	0.00	0.5810	0.8124	-0.0422	-0.0214

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
1.00	0.01	0.05	0.70	3.7364	182.3911	-37.0644	-0.8524
1.00	0.10	0.05	0.00	1.2648	2.4756	-0.2932	-0.0824
1.00	0.50	0.05	0.00	1.0244	1.3652	-0.0693	-0.0082
1.00	1.00	0.05	0.00	0.9926	1.3323	-0.0568	-0.0088
1.00	2.00	0.05	0.00	1.0260	1.4383	-0.0559	-0.0154

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$
2.00	0.01	0.05	1.00	NA	NA	NA	NA
2.00	0.10	0.05	0.01	4.2018	28.8325	-4.4826	-0.3923
2.00	0.50	0.05	0.00	2.1557	3.3808	-0.2243	-0.0139
2.00	1.00	0.05	0.00	1.9616	2.9233	-0.1343	0.0022
2.00	2.00	0.05	0.00	1.9187	2.9271	-0.0999	-0.0019

Estimator Percentiles (Sample Size 1000) for $\sigma = 0.1$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$
0.01	0.01	0.10	0.00	0.0983	0.2920	-0.1196	-0.0538
0.01	0.10	0.10	0.01	0.1085	0.2144	-0.0695	-0.0517
0.01	0.50	0.10	0.82	0.2458	0.4735	-0.0925	-0.0675
0.01	1.00	0.10	1.00	NA	NA	NA	NA
0.01	2.00	0.10	0.98	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$
0.10	0.01	0.10	0.01	0.4789	3.9393	-1.0941	-0.1574
0.10	0.10	0.10	0.00	0.2156	0.3201	-0.0916	-0.0618
0.10	0.50	0.10	0.00	0.2591	0.3630	-0.0839	-0.0626
0.10	1.00	0.10	0.01	0.4172	0.7954	-0.1181	-0.0806
0.10	2.00	0.10	0.98	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$
0.50	0.01	0.10	0.81	4.8084	238.5905	-63.6395	-1.2658
0.50	0.10	0.10	0.00	1.0537	1.9885	-0.3747	-0.1781
0.50	0.50	0.10	0.00	0.7187	0.9647	-0.1222	-0.0776
0.50	1.00	0.10	0.00	0.7844	1.0693	-0.1233	-0.0796
0.50	2.00	0.10	0.00	1.1189	1.7908	-0.1686	-0.1081

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$
1.00	0.01	0.10	1.00	NA	NA	NA	NA
1.00	0.10	0.10	0.00	3.3000	17.0639	-3.2426	-0.5594
1.00	0.50	0.10	0.00	1.4732	2.1618	-0.2286	-0.1126
1.00	1.00	0.10	0.00	1.3664	1.9395	-0.1758	-0.0878
1.00	2.00	0.10	0.00	1.5603	2.4930	-0.1961	-0.1079

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$
2.00	0.01	0.10	0.98	NA	NA	NA	NA
2.00	0.10	0.10	0.99	NA	NA	NA	NA
2.00	0.50	0.10	0.00	3.7522	8.3712	-0.8965	-0.2597
2.00	1.00	0.10	0.00	2.7928	5.0991	-0.4121	-0.1287
2.00	2.00	0.10	0.00	2.6540	4.9028	-0.3061	-0.1113

Summary

As we would expect this method gives good results for the noise free case. The estimates appear generally unbiased and are pretty good, even for the a and b. As we introduce

Gaussian noise the method becomes more unstable, especially in the strongly asymmetric case. The failure rates in these extremes rise steadily until we find almost guaranteed failure under these circumstances. Failure of the method is defined as the case where the solutions fall outside the acceptable range, for instance when we obtain a negative estimate for a.

The other consideration is where the model ceases to produce useful estimates of the parameters. In the asymmetric case this probably occurs somewhere before reaching $\sigma \approx 0.05$ while the symmetric case will go on producing reasonable estimates slightly longer.

Of course we are able to make no estimate of the level of noise contaminating the sample.

6.3.2 Case 2: Four Parameters($a = b, l_0, l_1, \sigma$)

This is the first of the new methods. It is intended for use with Symmetric Beta distributions and uses estimates of Cumulants in order to make inference about the parameters.

Estimator Percentiles (Sample Size 1,000)

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.00	0.11	0.0058	0.0141	-0.0255	0.0289	0.0035	0.0312
0.10	0.10	0.00	0.24	0.0835	0.1130	-0.0240	0.0282	0.0046	0.0305
0.50	0.50	0.00	0.39	0.4221	0.5408	-0.0180	0.0286	0.0055	0.0326
1.00	1.00	0.00	0.45	0.8079	1.0602	-0.0118	0.0397	0.0087	0.0408
2.00	2.00	0.00	0.50	1.3361	2.1178	-0.0057	0.0799	0.0116	0.0544

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.01	0.00	0.0061	0.0149	-0.0250	0.0258	0.0082	0.0308
0.10	0.10	0.01	0.07	0.0833	0.1143	-0.0228	0.0258	0.0065	0.0304
0.50	0.50	0.01	0.32	0.4250	0.5404	-0.0192	0.0277	0.0065	0.0337
1.00	1.00	0.01	0.42	0.8037	1.0617	-0.0096	0.0377	0.0077	0.0413
2.00	2.00	0.01	0.46	1.3909	2.1502	-0.0095	0.0733	0.0115	0.0542

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.05	0.00	0.0057	0.0147	-0.0240	0.0281	0.0483	0.0581
0.10	0.10	0.05	0.00	0.0834	0.1164	-0.0249	0.0255	0.0472	0.0587
0.50	0.50	0.05	0.00	0.4287	0.5798	-0.0224	0.0230	0.0397	0.0601
1.00	1.00	0.05	0.02	0.8035	1.2096	-0.0375	0.0361	0.0282	0.0678
2.00	2.00	0.05	0.19	1.2998	2.5254	-0.0539	0.0901	0.0208	0.0816

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.10	0.07	0.0021	0.0200	-0.0261	0.0284	0.0949	0.1068
0.10	0.10	0.10	0.00	0.0742	0.1240	-0.0266	0.0273	0.0931	0.1096
0.50	0.50	0.10	0.00	0.3943	0.6222	-0.0378	0.0368	0.0855	0.1145
1.00	1.00	0.10	0.00	0.6498	1.4226	-0.0787	0.0767	0.0709	0.1273
2.00	2.00	0.10	0.16	0.6298	3.5411	-0.1809	0.2071	0.0332	0.1487

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.50	0.38	0.1786	3.6458	-1.4082	-0.0916	0.2022	0.4872
0.10	0.10	0.50	0.45	0.2884	5.9131	-1.8006	-0.0905	0.1642	0.4878
0.50	0.50	0.50	0.73	0.5206	12.8933	-2.5414	-0.0774	0.1315	0.4913
1.00	1.00	0.50	0.83	0.7426	15.0891	-2.3302	-0.0353	0.1135	0.4705
2.00	2.00	0.50	0.88	0.2293	15.9057	-2.5807	0.0998	0.0985	0.4676

Estimator Percentiles (Sample Size 10,000)

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.00	0.22	0.0086	0.0116	-0.0084	0.0096	0.0017	0.0107
0.10	0.10	0.00	0.40	0.0950	0.1043	-0.0063	0.0096	0.0016	0.0107
0.50	0.50	0.00	0.41	0.4722	0.5110	-0.0042	0.0091	0.0026	0.0181
1.00	1.00	0.00	0.40	0.9229	1.0129	-0.0047	0.0117	0.0034	0.0210
2.00	2.00	0.00	0.48	1.6951	2.0381	-0.0040	0.0348	0.0040	0.0339

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.01	0.00	0.0087	0.0113	-0.0086	0.0088	0.0094	0.0145
0.10	0.10	0.01	0.00	0.0941	0.1052	-0.0066	0.0062	0.0060	0.0138
0.50	0.50	0.01	0.11	0.4749	0.5158	-0.0079	0.0090	0.0039	0.0197
1.00	1.00	0.01	0.38	0.9498	1.0192	-0.0045	0.0093	0.0065	0.0217
2.00	2.00	0.01	0.38	1.7639	2.0560	-0.0070	0.0299	0.0061	0.0341

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.05	0.00	0.0086	0.0115	-0.0088	0.0076	0.0492	0.0510
0.10	0.10	0.05	0.00	0.0941	0.1052	-0.0081	0.0091	0.0491	0.0516
0.50	0.50	0.05	0.00	0.4774	0.5248	-0.0074	0.0072	0.0468	0.0529
1.00	1.00	0.05	0.00	0.9174	1.0762	-0.0121	0.0144	0.0428	0.0564
2.00	2.00	0.05	0.00	1.7778	2.2651	-0.0378	0.0262	0.0311	0.0614

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.10	0.01	0.0064	0.0141	-0.0083	0.0087	0.0977	0.1022
0.10	0.10	0.10	0.00	0.0923	0.1049	-0.0089	0.0072	0.0976	0.1025
0.50	0.50	0.10	0.00	0.4607	0.5342	-0.0116	0.0104	0.0956	0.1038
1.00	1.00	0.10	0.00	0.8833	1.1157	-0.0228	0.0230	0.0912	0.1066
2.00	2.00	0.10	0.00	1.2665	2.5376	-0.0737	0.0788	0.0794	0.1196

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.50	0.46	0.0655	1.1732	-0.5468	-0.0465	0.4021	0.4922
0.10	0.10	0.50	0.45	0.1090	1.9154	-0.6737	-0.0007	0.3886	0.5070
0.50	0.50	0.50	0.64	0.6854	10.8585	-2.1349	-0.0796	0.2442	0.4877
1.00	1.00	0.50	0.68	0.5541	24.0956	-3.4825	0.0967	0.1621	0.4948
2.00	2.00	0.50	0.89	0.7043	56.3027	-5.2504	0.0571	0.0870	0.4651

Summary

Failure of this model is, as before, taken as the case where estimates are either meaningless (infinite or imaginary) or outside the accepted range (negative) for that parameter. We

find high failure rates for either end of the spectrum for added noise. With no noise the method will frequently fail and also for high levels. Unsurprisingly we find the best results where $\sigma < 1$ as the bimodal distribution makes estimation easier. This includes quite reasonable estimates for the level of the noise. In general the estimates are quite poor once the standard deviation of the noise is much above $\sigma = 0.1$. We find improved estimates when we look at samples of 10,000 rather than 1,000. In many cases the failure rate actually increases but when a result is obtained it is usually a better one.

6.3.3 Case 3: Five Parameters Asymmetric Moments (a, b, l_0, l_1, σ)

Estimator Percentiles (Sample Size 1,000) for $\sigma = 0$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.00	0.48	0.0051	0.0240	-0.0004	0.0017	0.0030	0.0330
0.01	0.10	0.00	0.67	0.0065	0.0173	-0.0001	0.0004	0.0015	0.0102
0.01	0.50	0.00	1.00	NA	NA	NA	NA	NA	NA
0.01	1.00	0.00	1.00	NA	NA	NA	NA	NA	NA
0.01	2.00	0.00	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.10	0.01	0.00	0.67	0.0563	0.1520	-0.0039	0.0020	0.0014	0.0099
0.10	0.10	0.00	0.51	0.0773	0.1300	-0.0006	0.0138	0.0086	0.0900
0.10	0.50	0.00	0.56	0.0801	0.1206	-0.0006	0.0025	0.0039	0.0212
0.10	1.00	0.00	0.58	0.0803	0.1208	-0.0005	0.0024	0.0037	0.0195
0.10	2.00	0.00	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.50	0.01	0.00	1.00	NA	NA	NA	NA	NA	NA
0.50	0.10	0.00	0.53	0.4042	0.6150	-0.0165	0.0098	0.0036	0.0207
0.50	0.50	0.00	0.54	0.4266	0.5741	-0.0014	0.0433	0.0237	0.1380
0.50	1.00	0.00	0.49	0.4298	0.5640	-0.0028	0.0117	0.0090	0.0462
0.50	2.00	0.00	0.45	0.4209	0.5754	-0.0025	0.0089	0.0065	0.0335

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
1.00	0.01	0.00	1.00	NA	NA	NA	NA	NA	NA
1.00	0.10	0.00	0.59	0.7946	1.1566	-0.0427	0.0419	0.0034	0.0200
1.00	0.50	0.00	0.47	0.8786	1.1397	-0.0227	0.0174	0.0083	0.0452
1.00	1.00	0.00	0.58	0.8585	1.1434	-0.0016	0.0788	0.0219	0.1534
1.00	2.00	0.00	0.45	0.8510	1.1273	-0.0052	0.0188	0.0095	0.0528

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
2.00	0.01	0.00	1.00	NA	NA	NA	NA	NA	NA
2.00	0.10	0.00	1.00	NA	NA	NA	NA	NA	NA
2.00	0.50	0.00	0.43	1.6674	2.4181	-0.0877	0.0693	0.0062	0.0341
2.00	1.00	0.00	0.48	1.6906	2.4009	-0.0552	0.0463	0.0100	0.0536
2.00	2.00	0.00	0.65	1.6856	2.4180	-0.0057	0.3431	0.0269	0.2098

Estimator Percentiles (Sample Size 1,000) for $\sigma = 0.01$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.01	0.24	0.0058	0.0250	-0.0015	0.0009	0.0055	0.0311
0.01	0.10	0.01	0.19	0.0074	0.0178	-0.0011	0.0001	0.0048	0.0135
0.01	0.50	0.01	1.00	NA	NA	NA	NA	NA	NA
0.01	1.00	0.01	1.00	NA	NA	NA	NA	NA	NA
0.01	2.00	0.01	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.10	0.01	0.01	0.20	0.0653	0.1659	-0.0067	0.0027	0.0043	0.0133
0.10	0.10	0.01	0.44	0.0761	0.1266	-0.0012	0.0120	0.0089	0.0885
0.10	0.50	0.01	0.37	0.0808	0.1222	-0.0017	0.0019	0.0042	0.0236
0.10	1.00	0.01	0.40	0.0835	0.1268	-0.0019	0.0014	0.0041	0.0212
0.10	2.00	0.01	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.50	0.01	0.01	1.00	NA	NA	NA	NA	NA	NA
0.50	0.10	0.01	0.37	0.4180	0.6100	-0.0172	0.0097	0.0037	0.0232
0.50	0.50	0.01	0.56	0.4304	0.5857	-0.0027	0.0484	0.0195	0.1428
0.50	1.00	0.01	0.42	0.4395	0.5719	-0.0044	0.0099	0.0085	0.0440
0.50	2.00	0.01	0.39	0.4268	0.5874	-0.0040	0.0074	0.0081	0.0354

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
1.00	0.01	0.01	1.00	NA	NA	NA	NA	NA	NA
1.00	0.10	0.01	0.41	0.8165	1.2001	-0.0434	0.0466	0.0043	0.0214
1.00	0.50	0.01	0.47	0.8970	1.1475	-0.0231	0.0188	0.0079	0.0465
1.00	1.00	0.01	0.59	0.8626	1.1640	-0.0039	0.0838	0.0311	0.1521
1.00	2.00	0.01	0.46	0.8510	1.1398	-0.0071	0.0197	0.0092	0.0528

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
2.00	0.01	0.01	1.00	NA	NA	NA	NA	NA	NA
2.00	0.10	0.01	1.00	NA	NA	NA	NA	NA	NA
2.00	0.50	0.01	0.41	1.6926	2.4255	-0.0844	0.0678	0.0059	0.0357
2.00	1.00	0.01	0.46	1.6993	2.4179	-0.0539	0.0483	0.0098	0.0545
2.00	2.00	0.01	0.67	1.6969	2.4207	-0.0089	0.3551	0.0376	0.2132

Estimator Percentiles (Sample Size 1,000) for $\sigma = 0.05$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.05	0.06	0.0257	0.0812	-0.0236	-0.0076	0.0390	0.0695
0.01	0.10	0.05	0.14	0.0278	0.0420	-0.0153	-0.0092	0.0471	0.0525
0.01	0.50	0.05	1.00	NA	NA	NA	NA	NA	NA
0.01	1.00	0.05	1.00	NA	NA	NA	NA	NA	NA
0.01	2.00	0.05	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.10	0.01	0.05	0.12	0.1650	0.4632	-0.0891	-0.0285	0.0471	0.0526
0.10	0.10	0.05	0.22	0.1058	0.1726	-0.0207	0.0004	0.0225	0.1051
0.10	0.50	0.05	0.00	0.1223	0.1706	-0.0194	-0.0106	0.0441	0.0555
0.10	1.00	0.05	0.17	0.1499	0.2177	-0.0226	-0.0134	0.0449	0.0546
0.10	2.00	0.05	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.50	0.01	0.05	1.00	NA	NA	NA	NA	NA	NA
0.50	0.10	0.05	0.00	0.5535	0.8831	-0.0751	-0.0215	0.0440	0.0552
0.50	0.50	0.05	0.50	0.4963	0.6715	-0.0279	0.0357	0.0312	0.1585
0.50	1.00	0.05	0.04	0.5194	0.6778	-0.0313	-0.0071	0.0256	0.0689
0.50	2.00	0.05	0.00	0.5761	0.7959	-0.0372	-0.0138	0.0342	0.0626

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
1.00	0.01	0.05	1.00	NA	NA	NA	NA	NA	NA
1.00	0.10	0.05	0.19	1.2555	2.0820	-0.1979	-0.0502	0.0447	0.0547
1.00	0.50	0.05	0.04	1.0150	1.3745	-0.0586	-0.0014	0.0220	0.0700
1.00	1.00	0.05	0.59	0.9686	1.3457	-0.0386	0.0636	0.0336	0.1600
1.00	2.00	0.05	0.11	1.0395	1.4235	-0.0459	-0.0016	0.0226	0.0776

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
2.00	0.01	0.05	1.00	NA	NA	NA	NA	NA	NA
2.00	0.10	0.05	1.00	NA	NA	NA	NA	NA	NA
2.00	0.50	0.05	0.00	2.1552	3.4219	-0.1962	0.0101	0.0343	0.0622
2.00	1.00	0.05	0.12	1.9786	2.9667	-0.1157	0.0206	0.0190	0.0770
2.00	2.00	0.05	0.67	1.8757	2.8801	-0.0527	0.3610	0.0420	0.2217

Estimator Percentiles (Sample Size 1,000) for $\sigma = 0.1$

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.10	0.07	0.0921	0.2889	-0.0882	-0.0363	0.0699	0.1407
0.01	0.10	0.10	0.13	0.1045	0.1663	-0.0563	-0.0421	0.0936	0.1059
0.01	0.50	0.10	1.00	NA	NA	NA	NA	NA	NA
0.01	1.00	0.10	1.00	NA	NA	NA	NA	NA	NA
0.01	2.00	0.10	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.10	0.01	0.10	0.15	0.4739	1.9193	-0.4511	-0.1205	0.0943	0.1064
0.10	0.10	0.10	0.26	0.1985	0.3080	-0.0763	-0.0203	0.0476	0.1694
0.10	0.50	0.10	0.00	0.2595	0.3645	-0.0689	-0.0511	0.0912	0.1080
0.10	1.00	0.10	0.18	0.4161	0.7401	-0.0895	-0.0633	0.0931	0.1076
0.10	2.00	0.10	1.00	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.50	0.01	0.10	1.00	NA	NA	NA	NA	NA	NA
0.50	0.10	0.10	0.00	1.0271	1.9353	-0.2850	-0.1284	0.0911	0.1089
0.50	0.50	0.10	0.54	0.6982	0.9583	-0.1051	-0.0074	0.0395	0.1853
0.50	1.00	0.10	0.00	0.7814	1.0556	-0.1029	-0.0547	0.0701	0.1249
0.50	2.00	0.10	0.00	1.0951	1.8073	-0.1300	-0.0795	0.0829	0.1153

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
1.00	0.01	0.10	1.00	NA	NA	NA	NA	NA	NA
1.00	0.10	0.10	0.15	3.1704	8.4312	-1.1891	-0.4042	0.0932	0.1076
1.00	0.50	0.10	0.00	1.4189	2.1058	-0.1789	-0.0665	0.0730	0.1227
1.00	1.00	0.10	0.69	1.3475	1.9506	-0.1436	0.3157	0.0339	0.2904
1.00	2.00	0.10	0.03	1.5446	2.4611	-0.1579	-0.0595	0.0525	0.1348

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
2.00	0.01	0.10	1.00	NA	NA	NA	NA	NA	NA
2.00	0.10	0.10	1.00	NA	NA	NA	NA	NA	NA
2.00	0.50	0.10	0.00	3.7122	8.3672	-0.7095	-0.1658	0.0835	0.1165
2.00	1.00	0.10	0.02	2.7434	5.0352	-0.3134	-0.0528	0.0578	0.1338
2.00	2.00	0.10	0.70	2.5734	4.7136	-0.2035	0.3906	0.0542	0.2414

Summary

This method can only be relied upon to function in certain areas of the parameter space. For strongly asymmetric cases $\left(\frac{a}{b} > 10\right)$ failure is common but also for the (exactly) symmetric case. As before we also find the method strongest when the parameters of the Beta distribution are between 0 and 1 and differ from each other.

Where the method works it gives better results than the simple Four Parameter Asymmetric model including fairly good estimates of the level of the noise. When, in the same circumstances, the method fails it is usually due to a solution indicating a negative level of noise. Correcting by taking the level of noise as zero in this case tends to give much the same results as in the Four Parameter Asymmetric model. As the noise levels rise the model shows all the same symptoms as the Four Parameter Asymmetric model of biased estimates but to a lesser degree. One difference is that when a reasonable estimate of the noise level is available then it would be possible to know the degree to which this bias had occurred and correct for it.

6.3.4 Case 5: Four Parameter Symmetric Moments

Estimator Percentiles (Sample Size 1,000)

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.00	0.11	0.0054	0.0140	-0.0257	0.0270	0.0110	0.0293
0.10	0.10	0.00	0.25	0.0830	0.1115	-0.0244	0.0277	0.0101	0.0306
0.50	0.50	0.00	0.41	0.4236	0.5422	-0.0183	0.0267	0.0079	0.0323
1.00	1.00	0.00	0.48	0.8034	1.0623	-0.0136	0.0385	0.0065	0.0407
2.00	2.00	0.00	0.00	0.2748	2.0440	-0.0008	0.2614	0.0130	0.1201

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.01	0.00	0.0058	0.0145	-0.0247	0.0244	0.0110	0.0329
0.10	0.10	0.01	0.10	0.0837	0.1141	-0.0225	0.0291	0.0101	0.0307
0.50	0.50	0.01	0.32	0.4237	0.5410	-0.0175	0.0270	0.0079	0.0338
1.00	1.00	0.01	0.44	0.8138	1.0554	-0.0110	0.0381	0.0066	0.0409
2.00	2.00	0.01	0.00	0.2988	2.0300	-0.0002	0.2585	0.0152	0.1201

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{I}_{0,0.05}$	$\hat{I}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.05	0.00	0.0053	0.0147	-0.0245	0.0281	0.0482	0.0592
0.10	0.10	0.05	0.00	0.0818	0.1168	-0.0231	0.0253	0.0468	0.0583
0.50	0.50	0.05	0.00	0.4283	0.5809	-0.0227	0.0247	0.0391	0.0605
1.00	1.00	0.05	0.02	0.7988	1.2010	-0.0366	0.0383	0.0272	0.0685
2.00	2.00	0.05	0.11	0.5942	2.4969	-0.0567	0.2089	0.0219	0.1191

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.10	0.12	0.0013	0.0196	-0.0263	0.0279	0.0953	0.1071
0.10	0.10	0.10	0.00	0.0725	0.1228	-0.0237	0.0286	0.0935	0.1108
0.50	0.50	0.10	0.01	0.3539	0.6139	-0.0324	0.0527	0.0863	0.1220
1.00	1.00	0.10	0.44	0.7611	1.4873	-0.0850	0.0496	0.0680	0.1139
2.00	2.00	0.10	0.71	1.2612	4.1031	-0.2303	0.0965	0.0272	0.1225

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.50	0.96	NA	NA	NA	NA	NA	NA
0.10	0.10	0.50	0.87	2.3784	8.8329	-2.2763	-0.9310	0.1087	0.3417
0.50	0.50	0.50	0.57	3.6282	46.9163	-4.7888	-1.0345	0.2010	0.3342
1.00	1.00	0.50	0.57	5.0290	96.3465	-6.7310	-1.1724	0.1970	0.3224
2.00	2.00	0.50	0.65	5.7918	168.9839	-8.2004	-1.1768	0.1965	0.3085

Estimator Percentiles (Sample Size 10,000)

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.00	0.18	0.0075	0.0100	-0.0076	0.0081	0.0111	0.0111
0.10	0.10	0.00	0.32	0.0940	0.1017	-0.0096	0.0092	0.0102	0.0102
0.50	0.50	0.00	0.39	0.4750	0.5059	-0.0051	0.0084	0.0079	0.0176
1.00	1.00	0.00	0.45	0.9442	1.0145	-0.0018	0.0122	0.0064	0.0230
2.00	2.00	0.00	0.00	0.3531	2.0281	-0.0015	0.2474	0.0050	0.1163

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.01	0.00	0.0086	0.0114	-0.0079	0.0084	0.0111	0.0111
0.10	0.10	0.01	0.00	0.0963	0.1042	-0.0067	0.0075	0.0102	0.0102
0.50	0.50	0.01	0.19	0.4780	0.5169	-0.0066	0.0084	0.0079	0.0177
1.00	1.00	0.01	0.40	0.9333	1.0178	-0.0038	0.0125	0.0064	0.0233
2.00	2.00	0.01	0.00	0.3890	2.0408	-0.0027	0.2439	0.0087	0.1159

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.05	0.01	0.0078	0.0113	-0.0094	0.0081	0.0483	0.0509
0.10	0.10	0.05	0.00	0.0944	0.1062	-0.0074	0.0082	0.0488	0.0512
0.50	0.50	0.05	0.00	0.4784	0.5215	-0.0074	0.0081	0.0468	0.0531
1.00	1.00	0.05	0.00	0.9304	1.0663	-0.0113	0.0134	0.0446	0.0564
2.00	2.00	0.05	0.00	1.7265	2.3629	-0.0375	0.0356	0.0263	0.0640

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.10	0.00	0.0028	0.0113	-0.0071	0.0089	0.0987	0.1032
0.10	0.10	0.10	0.00	0.0900	0.1056	-0.0076	0.0094	0.0985	0.1030
0.50	0.50	0.10	0.00	0.4313	0.5155	-0.0054	0.0213	0.0992	0.1092
1.00	1.00	0.10	0.56	0.7601	1.0188	-0.0029	0.0508	0.0964	0.1182
2.00	2.00	0.10	0.98	NA	NA	NA	NA	NA	NA

a	b	σ	Fail	$\hat{a}_{0.05}$	$\hat{a}_{0.95}$	$\hat{l}_{0,0.05}$	$\hat{l}_{0,0.95}$	$\hat{\sigma}_{0.05}$	$\hat{\sigma}_{0.95}$
0.01	0.01	0.50	1.00	NA	NA	NA	NA	NA	NA
0.10	0.10	0.50	1.00	NA	NA	NA	NA	NA	NA
0.50	0.50	0.50	0.86	8.2118	24.1012	-3.4489	-1.8569	0.1981	0.2745
1.00	1.00	0.50	0.52	15.6792	112.5247	-7.3728	-2.4966	0.1653	0.2680
2.00	2.00	0.50	0.65	18.1994	694.0079	-18.7541	-2.5049	0.1538	0.2809

Summary

In comparison with the symmetric method involving cumulants the result of this model are often very similar. Where both methods work well we find almost identical estimates. The

Moments method appear to be more influenced by the noise than the Cumulants method. The Cumulants method is more prone to failure in extreme cases but the estimates of the Moments method is these cases is rather suspect anyway.

6.4 Conclusions

All of the new methods do indeed appear to be capable of obtaining the correct solutions. There is no question that their ability to measure the level of the contaminating noise is useful as the existing method could not do this. These methods are also useful in that they may enable us to distinguish between a mixture of two normal distributions and that of a noisy Beta distribution.

All the methods work most effectively in the relatively narrow parameter range of $0.1 < a, b < 0.5$ where we have a clearly bimodal distribution (even with low levels of noise). Where we have a strongly asymmetric distribution the method will frequently find difficulty in distinguishing the second, lesser level. It is clear that having good knowledge of the levels of the process (or sample) is very important. When even a small amount of noise is added the estimation process is made much harder. If the levels of the process were able to be estimated more accurately inference would probably be possible for much higher levels of noise. This, for instance, would be the case where the sample represented percentages. The methods were also suprisingly effective at determining the level of noise obscuring the sample, even when all other estimates were hopelessly inaccurate.

None of the methods are robust enough to be effective in fitting the FM model to data unless the circumstances were ideal. By ideal we would require that the data was actually generated by a FM process, the noise levels were below $\sigma = 0.1$, the series was symmetric or close to symmetric and we had a long series to work with.

Chapter 7

Random Variate Generation

We have obtained an expression for the stationary distribution of the Filtered Markov (FM) process, and have found it to be Beta. In this chapter we show how the Filtered Markov model can be used for the generation of random Beta variates.

7.1 Random Variate Generation

There are many reasons why one might find it necessary to generate random numbers (or random variates). In a significant proportion of cases these random variates will be drawn from either the continuous or discrete Uniform distribution. There is also a real requirement for variates from other distributions, not least when undertaking simulations. Simulation is now a common method for solving scientific problems and may involve drawing random samples from any of the recognised distributions. The objective of drawing a random variate from a given distribution can usually be achieved in more than one way. A distribution may have a key property that gives rise to a generator. It is also possible that a relationship between the desired distribution and another, simpler, distribution may allow us to transform from the latter to the former. In general we would expect each distribution to have multiple methods available, displaying varying degrees of suitability for the purposes of efficient generation.

When we consider several different methods for generating a random variate there is not always a clear winner. In theory the best methods of all are exact methods. This is the case where we can invert the cumulative distribution function (CDF) to give a 1-1 relationship between a value in the support of the distribution and a probability measure of this value. For instance, if we represent a CDF by F and that of the random variable X (whose support is an interval) by F_X then

$$F_X(x) = \Pr(X \leq x)$$

Now $F_X(x)$ is a probability measure. It is a monotonically increasing function of x that maps the support of the random variable onto the set $[0, 1]$ such that each value of the probability measure of the random variable X identifies a point x uniquely. As a result we

can randomly select a value x from the distribution of X by randomly selecting a probability from $U[0,1]$ and then inverting the CDF of X.

$$\text{If } x \sim Dist(X) \text{ then } F_X(x) \sim U[0,1] \tag{7.1}$$

$$\text{So } x \sim Dist(F_X^{-1}(U[0,1])$$

This inversion of the CDF is not always (or even often) possible. Even where it is possible this method may still be far from efficient. If the inversion results in an infinite series or any form that requires numerical approximation the method may be impractical. The time taken to obtain a good approximation may be excessive or lengthy calculation may allow numerical errors to creep in. At the other extreme of the scale of elegance is the preparation of lengthy tables of pre-generated pseudo-random values. In order to obtain our random variate we would then choose one entry at random from this table. Between these two are many other methods, each concentrating on one feature or property of the distribution.

7.1.1 Choosing a Generator

Whenever a programmer requires the generation of values from a particular distribution, they must make a choice from the available methods. There are many different characteristics that should be taken into account to determine the most suitable method for the specific requirements of that application. In particular five attributes would be considered. They are: Statistical Reliability, Marginal Generation Time, Program Length, Set up Time, Memory Requirements. We shall consider each in turn.

- Statistical Reliability

Not all generators produce samples with precisely the target distribution. By Statistical Reliability we refer to a consideration of how closely the generated values adhere to the required distribution. Is the fit perfect or are they merely a close approximation? To establish this attribute of a generator it is necessary to demonstrate observations are, at least approximately, independent and drawn from the correct distribution. Testing for independence is relatively straightforward, while Chi-Square and Kolmogorov-Smirnov tests are frequently used for tests of fit. Stephens (1974) gives a good comparison of the methods available for this purpose.

- Marginal Generation Time

This is the demand in terms of processor time required to produce one random variate after all set up calculations or tasks have been performed. It will depend not only on

the complexity and length of the algorithm but also on the number and complexity of mathematical steps required. For instance, an addition and multiplication operation can be performed extremely quickly while exponential and logarithmic operations are comparatively slow. The marginal generation time might also depend upon the level of precision (statistical reliability) required by the user. Iterative methods, for instance, may often have the advantage of allowing the user to choose a level of precision to suit the current requirements, with possible consequences for performance.

- Program Portability/Length

Each method must be encoded into a specific function. This attribute is a measure of the ease with which this algorithm can be transferred from one machine to another. It would include a measure of the challenges involved in translating the code to suit your platform, preparing the new code and installing it. It gives a measure of the fixed investment by a programmer or developer, in terms of time and effort, for this algorithm to be made available for future use. This investment would only have to be made once for each type of platform as the coded algorithm could then be transported between similar machines.

- Set up Time

The time to prepare the algorithm for first use is covered under the previous attribute. The set-up time refers instead to the time required by the program itself when preparing to generate a sample of variates. This time may include the calculation of constants or preliminary values and will be required only once before generating a set of data but may have to be repeated if the values of the parameters of the generated variates change.

- Memory Requirements

This will depend upon the program length, memory occupied by constants, variables, arrays etc. It may also reflect the storage requirements of a pre-generated table of values. In practice the modern desktop machine is sufficiently powerful to make this a relatively minor consideration for the type of algorithms used for random variate generation.

Which of these considerations is deemed the most important depends completely on the application being considered. Rarely is a single algorithm strong in all areas. It is also not the case that the most commonly used distributions have the simplest algorithms. There are examples where the frequency of use ensures that portability and set-up time will be sacrificed for speed and reliability. When a user prepares a short program to conduct a particular experiment the opposite requirements may apply. They would prefer low initial investment (through low program length and complexity) even if this resulted in higher marginal generation times. A third example would be where a study was being made into the statistical behaviour of functions of random variables. In this case all other considerations would become inconsequential beside the requirement for a very high level of statistical reliability.

7.1.2 Kolmogorov-Smirnov Goodness of Fit Test

We have accepted that the output of a random variate generator may not always perfectly match the target distribution. In order to determine whether the approximation is sufficient for our purposes we may have to consider the 'goodness of fit' of a sample. By this we mean how closely the true distribution of a random variable corresponds to the distribution we have proposed that it is derived from. We do this by examining the empirical CDF (ECDF), constructed from a sample of the generated data, and making a comparison with its theoretical equivalent. When testing the values produced by a random number generator we have the advantage that we are hardly restricted by data collection. We can simply request as large a sample as we feel necessary, with relatively little time constraint, and then derive the ECDF. When samples are very large (of the order of millions) we can be sure of obtaining a very good estimate of the CDF and any statistical inconsistancies should become obvious.

In this case we are primarily concerned with the Beta distribution. We shall go on to show how the Filtered Markov model can be used to generate values closely approximating the Beta distribution. We shall term this the Stochastic Generator to distinguish it from other algorithms. We shall want to examine these values to determine how good this approximation is. The usual approach to this kind of problem is to apply a test of functional distance. Put simply, this looks at how closely the empirical CDF (constructed from a sample) resembles the true CDF of the Beta distribution. The most commonly used test of functional distance is the Kolmogorov-Smirnov test and we have chosen to use it to test the quality of the variates generated by the Stochastic Generator.

The Kolmogorov-Smirnov test (KS) is a non-paramateric test. The ECDF $\{E(n)\}$ of the sample distribution is constructed from the sample (arranged in ascending order) $\{x(1), x(2), ..., x(N)\}$ then taking the values $x(i)$ as estimates of the percentiles $(\frac{i}{N})$. An example of an ECDF as compared to a CDF in a continuous distribution is given in Figure 7.1. These estimates are then compared with the 'true' positions of the candidate CDF F in the interval around the estimate, $(x(i-1), x(i+1))$. The maximum absolute value of this distance for the whole sample is taken as the test statistic, D.

$$\begin{aligned} D^+ &= \sup_n |E(x(n)) - F(x(n))| \\ D^- &= \sup_n |E(x(n-1)) - F(x(n))| \ where \ E(x(0)) = 0 \end{aligned}$$

$$D = \max(D^+, D^-)$$

This test statistic can then be compared with a set of tables to determine the closeness of fit and whether to reject the null hypothesis $H_0 : E = F$.

The KS test is an extremely popular method for testing functional distance although there are others. Many of them share a general principle with KS but differ in the way of evaluating the test statistic. For instance the Kuiper test statistic is obtained by taking

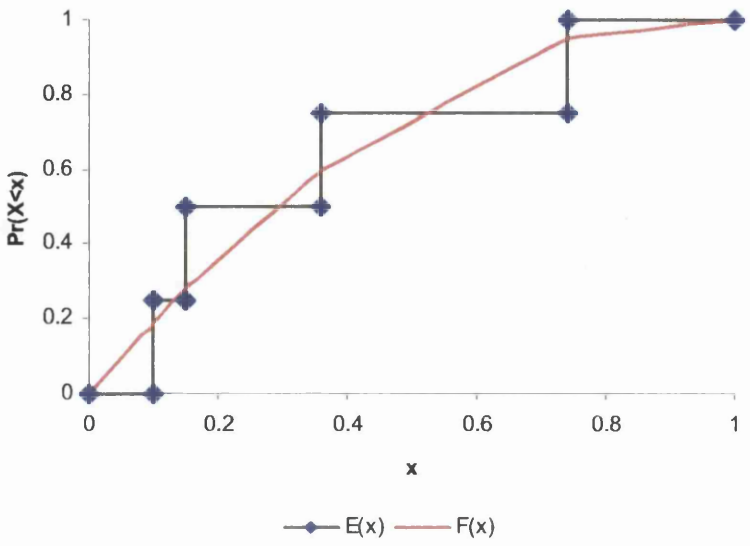

Figure 7.1: An illustration of the comparison between a theoretical CDF and an ECDF for a sample from a continuous variable.

the sum of D^+ and D^-.

Problems raised by the Beta Distribution

The usual method of applying the KS test is less ideal when considering the Beta distribution. When dealing with bounded functions with bounded (and closed) support it is effective but not all distributions have these characteristics. Unfortunately for us, the Beta distribution is an example of a distribution that does not. For low parameter values the Beta density function of the Beta distribution takes on the characteristic U shape. At each of the bounds of the x axis, at 0 and at 1, there is a pole. Although these points are not themselves included in the open range (0,1) from which Beta values are drawn this will affect the performance of KS. The generated values cluster close to the edges of the interval and through rounding errors, take on the values of these boundaries.

There are two ways that this problem affects our use of the test. The first is that there are values outside the range of the variable in a region which has a density value of 0. The second side effect of this rounding is to transform several, marginally different, values to the same value. This ensures that one large step in the CDF is matched against many small steps in the ECDF. So the two problems we have to deal with are:

Values outside the range

These don't really occur much until the parameter values are small. Matlab uses an algorithm called the Gamma Ratio method (see Section 7.2.1). When using this algorithm,

Parameters	Values Rounded(%)
0.1	1.25
0.05	7.84
0.01	34.53

Table 7.1: Failure rate for the Gamma Ratio algorithm for generating Beta variates.

which is employed in many embedded generators, the rounding is more frequently to 1 than to 0. By this time most embedded Beta generators are already failing regularly to produce values. When generating Gamma values with small parameters the values fall predominantly at the lower end of the range, close to zero. Matlab measures precision by giving the distance between 1 and the nearest smaller real number. This distance is only around 2^{-16}. By generating large numbers of Beta values using the Gamma ratio method we can estimate the frequency of these rounding errors, with the results given in Table 7.1. When rounding occurs it is usually to 1 rather than to 0, for the symmetric case.

A sample size of 100,000 using Johnk's Method (a Gamma variate generation algorithm) was used to produce these statistics. For parameter values smaller than 0.01 the Gamma Ratio method can not be relied upon to work adequately as the values routinely round to 0. Since this is the denominator of the generation equation this invalidates the method.

Duplication of Values

This duplication of values is closely linked with the previous problem. The most common reason that a duplication will occur is that both have rounded to one end of the scale. If this occurs occasionally it is unlikely to trigger a rejection of the distribution but as this number grows it becomes a problem. We have seen earlier that this rounding is a frequent occurrence for small parameter values.

For instance take the case with parameters of 0.1. We may find only around 1% of values rounding to 1. This is unlikely to trigger rejection. However for parameter values of 0.01, we may find a contribution of 0.3 to the test statistic. This will almost always result in rejection.

Adapting the Test.

With only a small adjustment we can avoid these pitfalls. The first, rather obvious, alteration is to allow for the existence of duplicate values. We can do this by simply combining several small steps into one large one. This will ensure that one comparison is made for each step taken by the ECDF rather than for each percentile. The second alteration is to avoid making any direct comparisons with the ends of the scale, 0 or 1. Since these have a CDF value of 0 in every case this will merely measure the proportion of duplicate values. By only comparing to values inside the range (0,1) we will avoid this addition to the test statistic. Both of these changes are in keeping with the principle of the KS test and merely

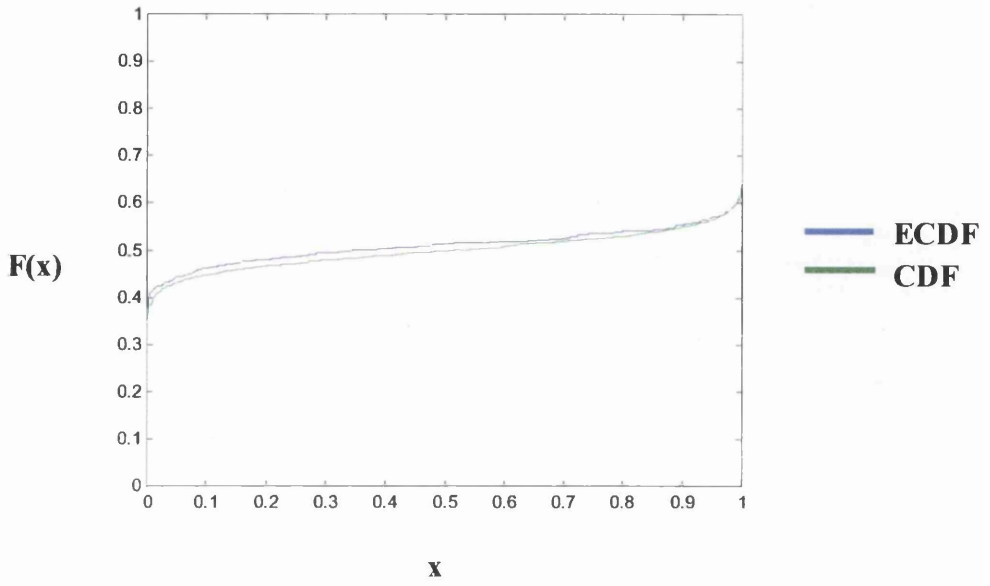

Figure 7.2: An example of the ECDF and CDF of a sample from the $Beta[0.05, 0.05]$ distribution.

ensure that we correctly compare the ECDF with the CDF. It should therefore not affect the null distribution of the test statistic or the effectiveness of the test as any excessive clustering of points at each end of the scale will affect the rest of the CDF.

In Example 7.2 we generate 1000 values from a Beta distribution with parameters $a, b = 0.05$. Applying the K-S test (at the 1% level of significance) without modification we obtain a test statistic of $D = 0.08$ which results in rejection of the Beta as the generating distribution. The modified K-S test results in a test statistic of $D = 0.0263$, within the range for acceptance. Figure 7.2 shows the ECDF and CDF of the data.

7.2 Beta Variate Generation

The Beta distribution is one of those with no dominant method of generation. There are many different methods available. They focus on different features of the distribution and none can claim to dominate in all areas. The primary reason for this is that the Beta distribution can take several, quite different, forms whereas the fast generation methods for random variate generation frequently use piecewise polynomial approximations to the distribution. For values of a,b<1 we have a U-shaped density function and for values of a,b>1 it has a unimodal density function. Each of the generation methods tends to be suitable for one case or the other. The PDFs of some of the main forms the Beta distribution can take are shown in Figures 7.3 to 7.6.

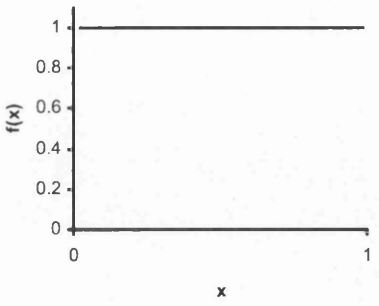

Figure 7.3: The PDF of a Beta[1,1] distribution.

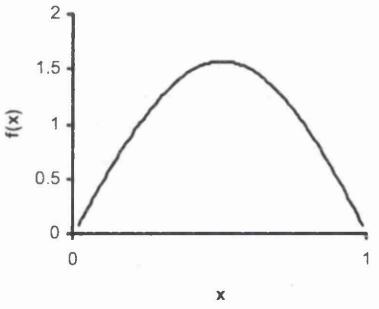

Figure 7.4: The PDF of a Beta[2,2] distribution.

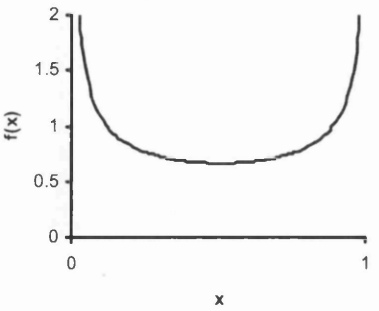

Figure 7.5: The PDF of a Beta[0.5,0.5] distribution.

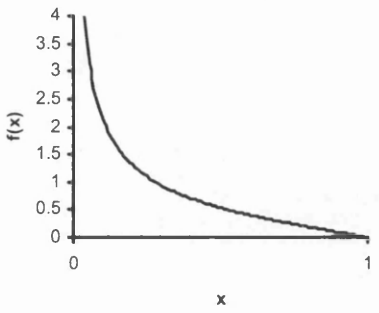

Figure 7.6: The PDF of a Beta[0.5,2] distribution.

7.2.1 Existing Generators

As we have already stated there is no dominant algorithm for generating Beta variates. This is not to say there is a paucity of algorithms. In fact the limitations of each of the available methods seems to have resulted in quite an assortment of different approaches being tried. We shall present here an overview of existing Beta generators and comment on their strengths and weaknesses.

Order Statistics

This method exploits a property of samples drawn from the standard Uniform distribution. The order statistics of this sample are Beta distributed

$$If\ X(i) \sim U[0,1]\ and\ \mathbf{X} = \{x(1), x(2), ..., x(N)\}$$

$$and\ the\ M^{th}\ order\ statistic\ of\ the\ sample\ \mathbf{X}\ is\ denoted\ by\ O_M(\mathbf{X})$$

$$then\ O_M(\mathbf{X}) \sim Beta(M, N)$$

There is, however, a significant problem with this method in that it can be used only to generate values from a Beta distribution with integer valued parameters. If the closest Beta distribution with integer parameters was a good subsitute then this fact would not matter so much. However, a brief study shows that in the parameter range that is most useful (and indeed used), namely when the parameters are small, the form of the distribution is highly dependent on their absolute (and relative) size. Another problem is that in order to generate Beta variates with high parameters rather large samples would have to be generated with the corresponding implications for generation time.

Gamma Method

This approach works by utilising the fact that a certain function of two Gamma variables is Beta distributed.

$$If\ G_a \sim Gamma(a)\ and\ G_b \sim Gamma(b)$$

$$then\ \frac{G_a}{G_a + G_b} \sim Beta(a, b)$$

This gives an easily implemented generator, suitable for all parameter values. However it is worth pointing out that different Gamma generation methods may be used for different parameter values. This method can be very competitive, as there are many good Gamma generators available although it is possible to improve on its performance. Among its potential weaknesses are that it requires an existing Gamma generator, or possibly more than one if the full range of parameters is to be utilised efficiently. This could greatly add to its set up time if a Gamma generator was not already available. There is another problem that we have already mentioned that should be treated as a failing of the application of the method rather than the method itself. It is that, due to limits on mathematical precision, the rounding down of values can result in errors. This only occurs at the lower end of the parameter range, with values of a,b below 0.01.

Johnk's Method

This is a method based on special properties of the beta density.

$$\begin{aligned} If\ R_1\ and\ R_2 &\sim U[0,1] \\ and\ R_1^{\frac{1}{a}} + R_2^{\frac{1}{b}} &\leqslant 1 \\ then\ \frac{R_1^{\frac{1}{a}}}{R_1^{\frac{1}{a}} + R_2^{\frac{1}{b}}} &\sim Beta(a, b) \end{aligned}$$

The expected time is not uniformly bounded in the parameters. It is also only appropriate when both parameters are less than 1. It can sometimes be slow, due to the two power transformations.

Standard Rejection Methods

A rejection method is a fundamentally different approach. It draws a random variable from a distribution approximating the required Beta distribution and then rejects those deemed unrepresentative These initial distributions are usually constructed in a piecewise fashion from polynomials. As a result of the slightly inelegant nature of the approach the algorithms are often a little impenetrable and abstract. These methods in general can be very effective and can often provide a competitive approach. Details of many of the possible rejection methods for the Beta distribution can be found in Devroye (1986). Each method

uses a slightly different set of distributions to form the proposed value and has its own rules for rejecting unsuitable candidates. The two most competitive versions are detailed below.

Cheng's Log-Logistic Method The Log-Logistic method of Cheng (1978) recognises that if W has a Beta-Prime distribution with PDF

$$\begin{aligned} f_W(x) &= \frac{x^{a-1}(1+x)^{-(a+b)}}{B[a,b]} \\ and\ X_{a,b} &= \frac{W}{1+W} \end{aligned} \tag{7.2}$$

then $X_{a,b}$ has the $Beta(a,b)$ distribution. The problem is then one of sampling from (7.2). One way to do this is use envelope rejection with a log-logistic target distribution. This is one of the best performing algorithms as it is robust and can be used for all values of a and b, something that cannot be said for many of the alternatives. Its performance is weakest when one parameter is small but it is easily implemented and requires little set-up time.

Atkinson's Switching Algorithms This approach introduced in Atkinson (1979) is not so much one method as a family of methods. In fact three different cases are included, each requiring a slight variation in the algorithm. The three cases, which assume $a \leqslant b$ and cover all possible parameter values are:

$$\begin{aligned} (i) \quad & a \leqslant b \leqslant 1 \\ (ii) \quad & a \leqslant 1\ and\ b > 1 \\ (iii) \quad & a > 1\ and\ b > 1 \end{aligned}$$

Although this method is not uniformly the fastest it does give the best overall performance over the entire parameter range. This is particularly true when a large sample is drawn without resetting the parameter values, due to a relatively high set-up time. The cost for this is a long and complex routine, and a method that is hardly transparent and should therefore be tested carefully after coding.

Strip Table Methods

A strip table method is favourite when speed is the primary requirement. The preparation and set up time requirements are prohibitive unless the code is to be heavily used. The basic approach relies on keeping a numerical representation of the distribution function stored on the machine to allow us to invert the cumulative distribution function numerically. Using this we can transform a probability measure, sampled from the Uniform distribution into a value from the target distribution. This approach is far more suitable for a distribution such at the Normal, where a standard distribution can be shifted and scaled to fit different parameters. The Beta is particularly unsuitable in this respect as each pair of parameters

describes a different shape of the PDF and CDF.

7.2.2 Summary

In general several methods have something to recommend them. The choice does depend upon the requirements of the user. If they are unwilling to invest a lot of time then the Gamma Ratio method is suitable as it is likely that Gamma generators may already be available. If they do not mind coding short programs, and a and/or b vary frequently, one of the rejection methods can be used. The method of Cheng is very robust with competitive performance. Overall Atkinsons's switching algorithms give the best performance if the decision is made primarily on Marginal Generation Time. Both of these two rejection methods can be used over the entire parameter range and the choice will come down to whether speed is valued more than program length or set-up time. If values are only required with parameters both less than one it may also be worth considering Johnk's method.

7.3 The Stochastic Generator

So far we have been looking at the different approaches currently available for generating Beta variates. The relevance of this to the stochastic process we have constructed relates to its stationary distribution. The next question we must consider is whether using the stochastic process offers a competitive algorithm for generating Beta variates.

7.3.1 The Stationary Distribution

The conditional and unconditional stationary densities are Beta distributed. The parameters of these Beta distributions are linked to the switching intensities (a, b) and time scaling factor (τ) of the process.

The unconditional PDF:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

The PDF condtional on the current state:

$$\begin{aligned} f_0(x) &= \frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b+1)} x^{a-1}(1-x)^{b} \\ f_1(x) &= \frac{\Gamma(a+b+1)}{\Gamma(a+1)\Gamma(b)} x^{a}(1-x)^{b-1} \end{aligned}$$

Recall that a, b are the switching intensities and τ is the time scaling factor (or switching rate parameter) and

$$a = \alpha.\tau \; and \; b = \beta.\tau$$

One of the consequences of this is that we immediately have expressions for the moments of the process. There is another consequence of this property. If the process has been running for a sufficient period of time and is stopped the value of the process is a random Beta variate. This gives an entirely new method of generating random numbers with a Beta distribution.

7.3.2 Convergence of the Process

We know that the stationary distribution is perfectly Beta. This gives us confidence that the output of the generator, if left sufficiently long, will closely resemble the Beta distribution. This does not guarantee us a competitive generator. Only if we can reach a good approximation quickly can we hope to compete with the existing methods. Two ways we can measure this convergence are by looking directly at the declining influence of the initial value of the process and the covariance (or correlation) of the process.

Influence of the Initial Point

We know that the standard process $X(t)$, with levels 0 and 1, follows only two rules, one for each of the two regimes of the driving signal $S(t)$.

$$\begin{aligned} X(dt) &= x(0)\exp\left(-\tfrac{dt}{\tau}\right) && if\ S(t) = 0\ for\ all\ 0 < t < dt \\ X(dt) &= 1 - (1 - x(0))\exp\left(-\tfrac{dt}{\tau}\right) && if\ S(t) = 1\ for\ all\ 0 < t < dt \end{aligned}$$

We can examine the result of using the same Markov process $S(t)$ with two different starting positions $x(0)$ and $y(0)$ (where $x(0) < y(0)$) to give us processes $X(t)$ and $Y(t)$. At time $t = 0$ the difference is $y(0) - x(0)$. After a period dt the distance between the two process is

$$\begin{aligned} Y(dt) - X(dt) &= (y(0) - x(0))\exp\left(-\tfrac{dt}{\tau}\right) && if\ S(t) = 0\ for\ all\ 0 < t < dt \\ Y(dt) - X(dt) &= (y(0) - x(0))\exp\left(-\tfrac{dt}{\tau}\right) && if\ S(t) = 1\ for\ all\ 0 < t < dt \end{aligned}$$

We can see that the length of the interval $[X(t), Y(t)]$ is completely predictable from the amount of time that has elapsed. We can therefore say that for a standard process with levels l_0 and l_1, after time t the envelope within which the process must be will be only $(l_1 - l_0)\exp\left(-\frac{t}{\tau}\right)$ wide. As a rough guide this will mean that convergence of the process will become practically complete during the interval $5 < \frac{t}{\tau} < 10$.

Correlation

Another way in which we can measure the convergence is by examining the correlation coefficient of the process. For very small time intervals dt this will be high declining to almost zero quite rapidly. We can obtain an expression for the Correlation of the process at two successive points in the same process.

On page 57 we introduced an expression for the Correlation coefficient of a Filtered

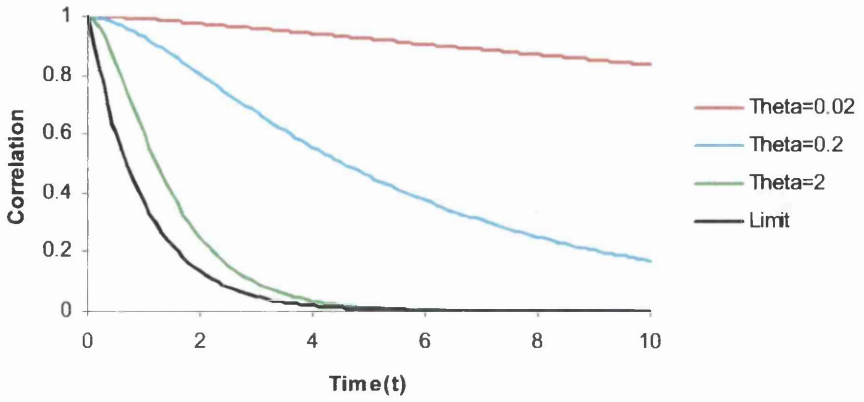

Figure 7.7: The correlation coefficient for two observations from the same Filtered Markov process, time t apart.

Markov process. The expression we obtained was dependent upon the switching parameters and the time scaling factor.

$$Corr[X(t), X(t+k)] = \frac{\theta\rho^k - \rho^{\theta k}}{\theta - 1} \tag{7.3}$$

We can display the structure of the correlation for different levels of θ when τ is taken as 1. This is shown below in Figure 7.7. The line marked as 'Limit' represents the limiting case as $\theta \to \infty$. We can see that for the higher values of θ we have very fast convergence but quite slow convergence for low θ. It is also worth presenting this information in another way. If we consider the value of the correlation after the amount of time that has elapsed is equal to the expected time until another switch we obtain another perspective. If we take the case of a symmetric process ($a = b$) and for each θ look at the correlation after time $k = \frac{2}{\theta}$ then we obtain the results shown in Figure 7.8. Note that the expression for the correlation (given in (7.3)) is not defined for $\theta = 1$, and we must use L'Hôpital's rule to obtain the value of the correlation at this point. This picture shows a slightly different side to the convergence. We see that for processes with smaller parameter values ($\theta < 2$) the correlation should become insignificant within only a small number of switches.

7.3.3 The Stochastic Generator

As we have seen earlier, the stationary distribution of the FM process is Beta. In theory, at least, it is possible to generate values from the Beta distribution using a long enough realisation of the process. That it is feasible does not guarantee that it is desirable though and a thorough consideration of the strengths and weaknesses is required.

The principle behind the method is very simple and easily converted to a programming language. The Beta distribution has (potentially) four parameters, We have two shape

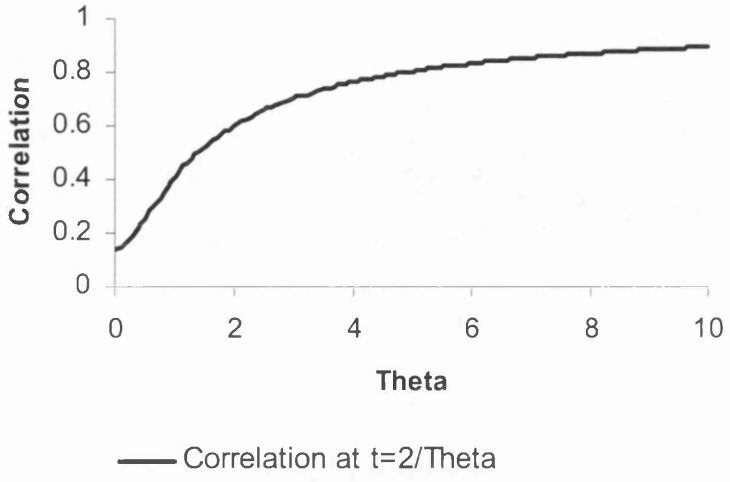

Figure 7.8: The correlation coefficient for two observations from the same Filtered Markov process, seperated by the expected length until the next regime switch.

parameters, a and b, and scale and shift parameters (c and d) to allow the distribution to be generalised. Of these we shall consider a and b only as the scale and shift can be applied to a standard Beta distribution in a linear transformation. So if we require a value X where $X \sim c.Beta(a, b) + d$ then we simply sample from a process that dwells in states 0 and 1 for periods that are exponentially distributed with alternating means $\frac{1}{a}$ and $\frac{1}{b}$ respectively. A simple linear transformation gives us our required distribution.

The Sampling Process

When the process has been running sufficiently long we can stop it at any point and determine its value. When generating observations from a process we would normally move from one observation point t_n to the next t_{n+1} but when generating random values these observations are unnecessary. This can be simply demonstrated. If we examine the distribution of the process just before a switchpoint (from regime 0 to regime 1) we find simply the stationary distribution conditional on being in regime 0. This is due to the 'lack of memory' property of Poisson processes. From this we can conclude that we can move from one switchpoint to the next, randomly generating the intervals between them rather than concern ourselves with regular observations. At each switchpoint it will be necessary to evaluate the level of the process, which depends upon the level of the process at the previous switchpoint and the time that has elapsed since then. In doing this we are actually sampling from a discrete time Markov chain. We know that the conditional distribution for each regime will be correct but we also require that distribution between regimes is correct in order for the stationary distribution of this chain to be Beta.

As far as determining our algorithm we have two different approaches we can take. When starting the process to obtain a single Beta variate we need to determine the point at which we interrupt the process. This could be done either by specifying the exact time that must elapse before we stop the process or to specify the number of switches we require the process to make before taking an observation from a switchpoint.

- Fixing Finish Time
  This has one clear advantage. The influence of the starting state on the final distribution is easy to measure. The disadvantages, however are significant. The complexity of the technique is much increased with a requirement for a continual check on amount of time that has elapsed. These conditional statements are very time consuming for a processor. As the parameter values become very large the length of time spent in each regime becomes shorter requiring many more switches (and hence iterations) to reach the finish time. This renders the marginal generation time unbounded as the switching rate tends to infinity. It also does not guarantee that any switches will take place during the time interval leaving the generated value determined precisely by the starting point. In order to avoid the worst of these problems the finishing time could be determined by a function of the parameter values, ensuring a more practical choice. As long as the initial regime and level of the chain are chosen from the stationary distribution of the process then the final value of the chain should also have the correct distribution.

- Fixing Switch Number
  Using this method has several clear advantages. First, the process becomes trivially simple. The initial point is determined and then a number of values are drawn from the exponential distribution for the time spent in each period and the final value is determined. Secondly, since the length of a switch is very long for low parameter values, the run-in time is long enough for very convincing convergence to have taken place. The only problem is at the other end of the scale. For large parameter values ($\theta > 1$), intervals between switches are very short, and very little movement of the process (and hence convergence) will take place leaving the process very dependent upon the initial distribution. The choice of how to initialise the chain is slightly more complex than in the first case. If we choose an odd number of switches then we will always find the final regime of the chain different to the initial regime. It is important to choose the initial regime from a distribution such that the final regime of the chain is distributed correctly.

The choice of approach will depend upon the purpose for which the generator is intended. When considering performance the second approach does result in a simpler routine. We shall mostly concern ourselves with the second approach as performance will be crucial in any comparisons made with other algorithms.

Initial Conditions

The process will never become independent of the initial value $x(0)$. However it does become, quite quickly, effectively independent of it for many choices of parameters. Depending on the method we choose for sampling we may or may not have control on the degree of independence. As such it is important to study the importance, or otherwise, of the initial value. If the initial value is a random variable then we will sidestep many of the problems associated with fixed starting values. There are several considerations when choosing the distribution of the starting value value.

- Distribution
  The closer we can get to the required Beta distribution when choosing the initial value the less convergence is required. When working with high parameter values, for instance, the Beta distribution can be closely approximated by other distributions (the Normal distribution for instance) with appropriately chosen mean and variance. A suitable choice will give a good starting point for the process to converge quickly.

- Extreme Values
  The process cannot ever actually reach the upper and lower bounds given by the two levels 0 and 1. The longer it spends without switching the closer it can get to these bounds, and it is important to ensure the generator distribution has sufficiently complete coverage of the required range. One quick and easy way of determining the initial position is to choose the mean value of the process. This guarantees fast convergence in most cases but for fast switching may not allow the full range [0,1] to be generated.

- Speed
  When looking to maximise the performance of the generator every little fraction counts. Choosing a Normally distributed initial point may take many times longer to generate than a fixed starting point (such as the mean value).

We should not forget how varied the forms of the Beta distribution are. It is likely that different initial distributions would suit different ranges of the parameter values. If we are able to choose one initial distribution to suit all parameter values it will probably be due to the decreased importance of the starting point for many of these sub-groups rather than its consistent suitability.

Choice of Initial Distribution

In terms of maximising performance, fixing the number of switches seems to offer the most promise. In order to further optimise the performance we need to select the best distribution for our intial point. We shall consider several different options and consider the advantages

and drawbacks of each. The parameters of the Beta distribution we intend to generate are a and b which will have levels 0 and 1.

- Bernoulli.
  The initial regime of the Filtered Markov process $S(0)$ is chosen from a Bernoulli distribution where the probabilities are taken from the parameters of the required Beta. The intial value $x(0)$ is then taken as the level of the opposing regime i.e.

$$\begin{aligned} S(0) &\sim Bernoulli\left(\frac{a}{a+b}\right) \\ X(0) &= 1 - S(0) \end{aligned}$$

  In choosing $S(0)$ we ensure that the choice of regime is in keeping with the stationary distribution of the Filterered Markov process. In choosing this value of $X(0)$ we ensure the greatest possible range of values for the process after the first switch encouraging fast convergence. This choice of starting conditions is very fast and suitable for lower switching intensities.

- Uniform.
  The regime of the process $S(0)$ is also taken from a Bernoulli distribution with appropriate probabilities but the initial value is Uniformly distributed between 0 and 1 i.e.

$$\begin{aligned} S(0) &\sim Bernoulli\left(\frac{a}{a+b}\right) \\ X(0) &\sim U[0,1] \end{aligned}$$

  This is also quite fast and more suitable for the intermediate values of the switching intensities where the Beta distribution may be relatively evenly distributed across its support.

- Normal.
  The initial regime is chosen from a Bernoulli distribution as before but the level of the process is taken from a Normal distribution with the same mean and variance as the appropriate conditional Beta distribution. i.e.

$$S(0) \sim Bernoulli\left(\frac{a}{a+b}\right)$$

$$If\ S(0) = 0 \quad \mu = \left(\frac{a}{a+b+1}\right) \text{ and } \sigma^2 = \left(\frac{a(b+1)}{(a+b+1)^2(a+b+2)}\right) \tag{7.4}$$

$$If\ S(0) = 1 \quad \mu = \left(\frac{a+1}{a+b+1}\right) \text{ and } \sigma^2 = \left(\frac{(a+1)b}{(a+b+1)^2(a+b+2)}\right) \tag{7.5}$$

$$X^*(0) \sim Normal(\mu, \sigma^2) \tag{7.6}$$

But the Normal distribution has infinite support while the Beta is bounded so we take

$$X(0) = \begin{cases} X(0) = 0 & if\ X^*(0) \leqslant 0 \\ X(0) = X^*(0) & if\ 0 < X^*(0) < 1 \\ X(0) = 1 & if\ X^*(0) \geqslant 1 \end{cases}$$

This gives us an initial distribution that is already very close to the Beta distribution for higher switching intensities. This is not the case for lower values and it is by far the most time consuming choice of initial distribution.

In order to test these three choices for the initial conditions we shall resort to simulations. We shall generate many thousands of values using this process, obtain from these an empirical cumulative density function (ECDF) and then test this against the Beta distribution using the Kolmogorov-Smirnov test of functional distance. Each column in the tables below will represent a different model specification. The model specification will begin with the choice of starting distribution Bern, Uni and Norm to represent Bernoulli, Uniform and Normal respectively. The number following this will give the number of switches that will be added after the initial value is chosen. Uni2, for instance, will be generated using a Uniform initial point and then two intervals culminating in switchpoints with the value at the second switch point taken as the generated Beta variate.

Table 7.2 gives the functional distance (as defined in the KS test) between an ECDF produced using a random sample of 10,000 variates and the theorectical CDF for symmetric Beta distributions. Table 7.3 gives the mean rejection rate by the KS test (at a 1% level of significance) of the ECDF from the Beta CDF over 1000 samples of 1000 from symmetric Beta distributions. Tables 7.4 and 7.5 are identical except that they focus on asymmetric Beta distributions. For comparison we also include the exising Matlab generator (which uses the Gamma Ratio method) for comparison and the figures are given in the column headed 'Beta'. It should be noted that the first two rows are absent from the Beta column. This is due to the failure of the Gamma Ratio method (as used by Matlab) to reliably generate values for these parameter values due to rounding errors.

For the Bernoulli and Uniform prior we find that the process in unable to converge quickly enough for higher a and b. This is due to the very short expected length of the intervals between switches. The Normal distribution does not have such a problem as it is already a good approximation to the Beta distribution with high parameters. The generator is far less sensitive to choice of prior for small a and b, due to the rapid convergence. We can also see even over 4 switches the rejection rate for the Normal starting distribution remains very low. The only exceptions to this are for the highly asymmetric cases such as $a = 10, b = 10,000$. To be on the safe side we should probably take 6 switches instead of 4, after which the rejection rate is indistinguishable from the existing generator.

Parameters		Models						
a	b	Bern2	Bern4	Uni2	Uni4	Norm2	Norm4	Beta
0.001	0.001	0.0234	0.0237	0.0232	0.0233	0.0232	0.0242	-
0.005	0.005	0.0254	0.0252	0.0249	0.0253	0.0255	0.0252	-
0.01	0.01	0.0261	0.0259	0.0263	0.0257	0.0267	0.0250	0.0260
0.05	0.05	0.0270	0.0275	0.0275	0.0269	0.0276	0.0268	0.0269
0.1	0.1	0.0272	0.0270	0.0274	0.0274	0.0264	0.0268	0.0274
0.5	0.5	0.0417	0.0275	0.0342	0.0276	0.0268	0.0267	0.0275
1	1	0.0890	0.0345	0.0327	0.0279	0.0272	0.0273	0.0272
2	2	0.1892	0.0826	0.0370	0.0270	0.0280	0.0274	0.0276
5	5	0.3811	0.2730	0.1425	0.0889	0.0277	0.0282	0.0273
10	10	0.4852	0.4428	0.2454	0.2053	0.0268	0.0267	0.0278
50	50	0.5130	0.5125	0.3883	0.3832	0.0274	0.0276	0.0273
100	100	0.5131	0.5121	0.4204	0.4192	0.0273	0.0287	0.0275
1000	1000	0.5127	0.5126	0.4792	0.4803	0.0278	0.0280	0.0268

Table 7.2: A comparison of the K-S Goodness of Fit test statistics of the ECDF of a sample from the Stochastic Beta Generator for symmetric Beta distributions.

Parameters		Models						
a	b	Bern2	Bern4	Uni2	Uni4	Norm2	Norm4	Beta
0.001	0.001	10	7	3	9	9	5	-
0.005	0.005	9	6	10	9	10	11	-
0.01	0.01	10	11	5	8	4	4	10
0.05	0.05	9	11	11	11	17	5	6
0.1	0.1	15	12	10	8	6	2	14
0.5	0.5	157	12	45	10	3	6	6
1	1	1000	41	24	11	11	6	8
2	2	1000	1000	61	12	5	4	5
5	5	1000	1000	1000	1000	5	5	14
10	10	1000	1000	1000	1000	14	14	15
50	50	1000	1000	1000	1000	10	16	15
100	100	1000	1000	1000	1000	12	5	3
1000	1000	1000	1000	1000	1000	16	10	5

Table 7.3: A comparison of the rejection rate of the K-S Goodness of Fit test of the ECDF of a sample from the Stochastic Beta Generator for symmetric Beta distributions.

Parameters		Models						
a	b	Bern2	Bern4	Uni2	Uni4	Norm2	Norm4	Beta
0.01	**0.1**	0.0276	0.0275	0.0273	0.0276	0.0277	0.0271	0.0274
0.01	**1**	0.0275	0.0271	0.0281	0.0272	0.0273	0.0274	0.0271
0.01	**10**	0.0271	0.0270	0.0341	0.0276	0.0274	0.0271	0.0274
0.01	**100**	0.0272	0.0276	0.0486	0.0274	0.0276	0.0269	0.0273
0.1	**1**	0.0279	0.0277	0.0495	0.0275	0.0280	0.0271	0.0276
0.1	**10**	0.0296	0.0277	0.1569	0.0342	0.0300	0.0273	0.0274
0.1	**100**	0.0277	0.0275	0.2971	0.0693	0.0310	0.0273	0.0273
1	**10**	0.2412	0.0970	0.4451	0.2594	0.0401	0.0291	0.0272
1	**100**	0.2827	0.1251	0.8315	0.6735	0.0486	0.0306	0.0269
1	**1000**	0.2887	0.1292	0.9644	0.9036	0.0503	0.0311	0.0274
10	**100**	0.8884	0.8559	0.8195	0.8097	0.0400	0.0351	0.0273
10	**1000**	0.9707	0.9397	0.9765	0.9750	0.0445	0.0386	0.0276

Table 7.4: A comparison of the K-S Goodness of Fit test statistics of the ecdf of a sample from the Stochastic Beta Generator for asymmetric Beta distributions.

Parameters		Models						
a	b	Bern2	Bern4	Uni2	Uni4	Norm2	Norm4	Beta
0.01	**0.1**	13	17	3	12	12	10	9
0.01	**1**	10	7	21	10	7	11	7
0.01	**10**	6	8	56	6	9	5	10
0.01	**100**	8	6	357	11	6	11	8
0.1	**1**	15	11	427	13	17	6	11
0.1	**10**	29	10	1000	58	30	9	10
0.1	**100**	30	15	1000	953	32	6	12
1	**10**	1000	999	1000	1000	193	25	6
1	**100**	1000	1000	1000	1000	395	37	8
1	**1000**	1000	1000	1000	1000	452	52	12
10	**100**	1000	1000	1000	1000	176	84	13
10	**1000**	1000	1000	1000	1000	278	139	12

Table 7.5: A comparison of the rejection rate by the K-S Goodness of Fit test of the ecdf of a sample from the Stochastic Beta Generator for asymmetric Beta distributions.

7.3.4 Comparison of Performance

The generation of random Beta variates from this Stochastic process has much to recommend it. The method is transparently simple, easy to code and relatively quick. This is not the only reason it should be considered as a serious alternative to the existing models. Some of the methods applied, or the coding of them may occasionally fail to succesfully return Beta values when the parameters fall into the extremes of the possible range. There is definitely a case for arguing that one, reliable, method could replace the current patchwork of generators that are currently employed by most Statistical software.

Statistical Reliability

We already know that the Stochastic method is reliable in the sense that it will always return a value, no matter how extreme the parameters. In assessing the suitability of the generator we are more concerned with statistical reliability. In the case of the stochastic process we know that the stationary distribution will match the desired Beta distribution but need to be sure that the output of the generator is close to that stationary distribution. We can judge this convergence by examining the KS test statistic for different parameter values and different numbers of steps of the generator. In the chosen version of the generator we use a Normal prior, and we can see from Table 7.2 that the generator has converged after only 2 steps for the symmetric case, even for extreme parameters. The asymmetric case is dealt with in Table 7.4, and we see slower convergence here especially when both parameters are larger than 1. For the most problematic cases we plot the mean KS test statistic versus the number of steps used by the generator to see how this convergence develops. The results are given in Figure 7.9 and show the mean KS test statistic over 1000 experiments, each consisting of a sample of 1000 values. The dotted lines represent the rejection levels for the KS test, at the 1%,5% and 10% levels. As we can see in even the most difficult cases the generator converges to an acceptable level by 6 steps and (effectively) fully converges after about 10 steps. The evaluation of the marginal generation time is done using the Normal prior with 6 steps. This should be sufficient for all but the most demanding scenarios and is probably a conservative choice when dealing with the usual range of parameters.

Marginal Generation Time

The currently used methods vary in their performance. There is not just variation between the different methods but variation for different parameter values. Some are fast but work only for limited sets of the parameter space and some are fast in some areas and slow is others. The measured speeds are obviously dependent upon the capabilities of the processor used to generate the values and should not be taken as absolute. They do, however give a good indication of the relative performance of the different methods. With the exception of a change in the way mathematical functions are evaluated by processors this relative performance should be unlikely to change dramatically.

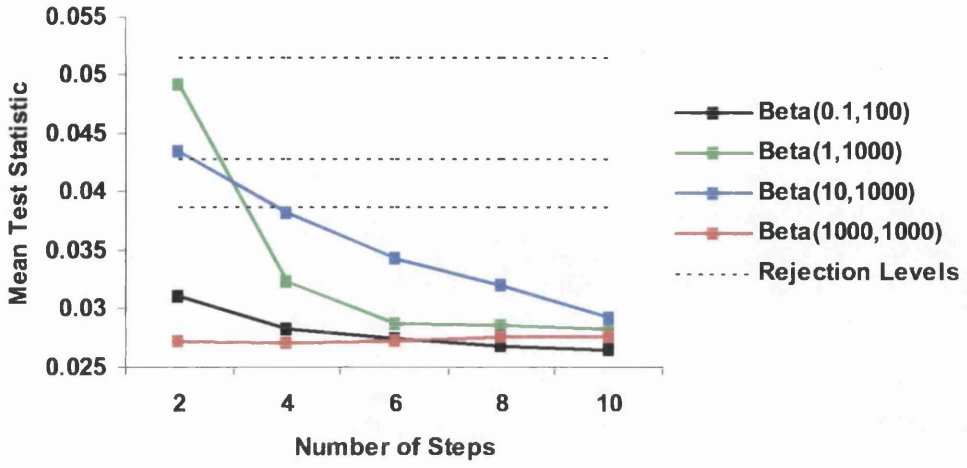

Figure 7.9: Convergence plots for the Stochastic Generator for different Beta distributions. The number of steps of the generator is plotted against the mean KS test statistic. The dotted lines represent the rejection level of the KS test at (from top down) 1%, 5% and 10%.

In order to fully test the alternative algorithms available we shall perform our own simulations. To obtain these benchmark timings the methods were coded in as minimal a form as possible, with all non-essential commands removed. A selection of the methods have been tested and their performance given below (as μs per variate) so that a comparison can be made with the performance of the Stochastic method. In practice none of these methods can acheive these speeds. The actual marginal generation time is small compared to the time taken to check the command syntax and parameter set and then select the correct generator. This is why each method should be measured in its minimum form. It is, however, worth bearing in mind that the Stochastic generator works for all parameter values. This would save time when compared against the suite of generators (such as Atkinson's Switching method) but not individual ones. Even without this factor it is clear to see that the Stochastic generator performs competitively, especially in certain areas of the parameter space.

Marginal Generation Times for Beta Generation Algorithms (μs)

Gamma Ratio (Varying Parameters)					
a			b		
	0.01	0.1	1	10	100
0.01	141.3	144.2	192.3	185.2	188.3
0.1		144.2	194.2	185.3	187.2
1			241.3	233.3	235.3
10				226.3	229.4
100					228.4

Cheng's Log-Logistic (Varying Parameters)					
a			b		
	0.01	0.1	1	10	100
0.01	231.4	317.4	333.5	335.5	337.5
0.1		224.3	284.4	290.4	292.4
1			181.3	202.3	206.3
10				187.2	190.3
100					188.3

Atkinson's Switching (Varying Parameters)					
a			b		
	0.01	0.1	1	10	100
0.01	143.2	126.2	107.1	107.1	109.2
0.1		134.2	106.1	107.1	109.2
1			94.2	108.2	109.1
10				161.2	197.3
100					165.2

Gamma Ratio (Fixed Parameters)					
a			b		
	0.01	0.1	1	10	100
0.01	136.2	137.2	135.2	177.3	179.2
0.1		140.2	137.2	178.3	179.3
1			132.2	176.2	177.3
10				217.4	218.4
100					220.3

Cheng's Log-Logistic (Fixed Parameters)					
a			b		
	0.01	**0.1**	**1**	**10**	**100**
0.01	106.1	194.3	211.3	209.3	211.3
0.1		96.1	159.2	170.3	169.2
1			53.1	78.1	81.2
10				60.1	64.1
100					61.1

Atkinson's Switching (Fixed Parameters)					
a			b		
	0.01	**0.1**	**1**	**10**	**100**
0.01	77.1	59.1	40.1	40	40.1
0.1		64.1	38.1	39	39.1
1			37.1	38.1	37.1
10				46.1	78.1
100					50

Stochastic Method					
a			b		
	0.01	**0.1**	**1**	**10**	**100**
0.01	99.1	100.1	99.1	101.2	99.1
0.1		99.2	99.2	99.1	100.1
1			180.3	179.2	179.2
10				180.2	179.3
100					180.2

Program Portability/Length

This may be a relatively minor consideration for a serious developer, but some of these methods have much more complex structure than others. When a method is required that is fast to code and easy to understand then some weight may be given to complexity. In Table 7.6 we make a comparison of the minimum number of commands required to code each of the leading Beta generation algorithms. Atkinson's suite of generators suffers under this perspective. The algorithm consists of three different generators with the most appropriate one selected for each parameter set. The rejection methods are also less transparent that some other methods. Again we can see that the Stochastic method performs well when compared with the best of the existing generators.

Algorithm	Lines of Code (approx.)
Atkinsons's Switching Family	80
Cheng's Log-Logistic Method	60
Stochastic Method	20

Table 7.6: A comparison of the minimum number of commands required to code each of the leading Beta generation algorithms.

7.4 Conclusion

Since a low Marginal Generation time is probably the most important single property a generator can have it is ultimately the determining factor in the importance of a generator. As such we must accept that the Stochastic Generator cannot comprehensively outperform Atkinson's algorithm across the whole parameter range. It does offer better performance where $a, b < 1$ even using 6 switching intervals, but it suffers from the disadvantage that no performance improvement is gained when drawing large samples with the same parameter values. In this lower range it is likely that a more careful choice of initial distribution may result in even lower generation times. However if speed is not the sole consideration the Stochastic Generator must be worthy of consideration.

Chapter 8

The Ladder Models

In an earlier chapter we outlined the use of discrete time, discrete level models for use as approximations to the Filtered Markov (FM) process. In this chapter we propose a class of these models for this purpose and also for use as a simple but effective method for extending the Markov switching models to incorporate gradual switching mechanics.

8.1 Another look at Gradual Switching

Given the great difficulties in fitting the FM model to noisy time series data it has been necessary to construct a simpler, more robust, model. So we go back to a more general, and mechanically simple, model that has the two-state case as its simplest form. We retain the idea of a two regime process but allow the process to take intermediate levels as it adjusts from one regime to the other. After the gradual transition has been completed to either regime the model will remain at a stable level until the next switch occurs. We call these two levels the stable levels or end levels and all other levels intermediary. We link the direction of movement through the levels to the current regime and restrict the process to remain within the finite level structure. This allows us the Markov Regime Switching behaviour but allows much more complex transitions between the upper and lower ends of the range of levels. This is stated more formally below.

8.2 A Class of Gradual Switching Models

In order to explore the possibility of practical gradual switching models we introduce a class of discrete time, discrete level models. As with Hamilton there is a Markov chain controlling the Mean level of the series with a Gaussian error term added to introduce the observable series.

$$y(n) = g(n) + \epsilon(n)$$

Where the mean level of the series at time n is given by $g(n)$ and $\epsilon(n) \sim N(0, \sigma^2)$

The model for determining the mean level consists of two chains, with discrete levels and in discrete time. The first chain $S(n)$ will consist of two-regimes that determine which of two movement patterns the series exhibits. The second chain $L(n)$ controls the mean value of the series. The first chain $S(n)$ can sometimes be influenced by the second chain $L(n)$ and $L(n)$ will always be influenced by $S(n)$. Together the chain of combined states $(S(n), L(n))$ form a Markov chain, as the distribution in the future state depends solely on its current state. When we need to refer to the state the process occupies $(S(n), L(n))$ we will need to distinguish between $(0, k)$ and $(1, k)$ (which are the case of the same level with different regimes). We will do this by indexing the first as dk (for down regime, level k) and uk (for up regime, level k).

The mean value of the process at a given point n is at one of a 'ladder' of $N_{(L)}$ levels.

$$\{l_0, l_1, ..., l_{N_{(L)}-1}\} \text{ where } l_u < l_v \text{ if } u < v$$

Such that

$$E[Y(n)] = l_u \; if \; L(n) = u$$

The chain $L(n)$ is influenced by $S(n)$ which controls the direction of movement on the ladder

$$\begin{aligned} \text{if } S(n) &= 1 \text{ and } L(n) = u \text{ then } L(n+1) \geq u \\ \text{if } S(n) &= 0 \text{ and } L(n) = u \text{ then } L(n+1) \leq u \end{aligned}$$

Since the ladder can move only up while $S(n) = 1$ once the ladder reaches the top of the ladder $(L(n) = N_{(L)} - 1)$ no further change in the level of the chain is possible until the direction of movement changes as a result of a change in regime of S.

Example 1:

In its simplest form $S(n)$ could be a (two regime) Markov chain with transition probabilities a and b (for states 0 and 1 respectively) and $N_{(L)} = 2$. Then we have regimes 0 and 1 with associated levels $l_0 = 0, l_1 = 1$ and the transition matrix

$$\begin{bmatrix} \Pr[S(n) = 0 | S(n-1) = 0] & \Pr[S(n) = 1 | S(n-1) = 0] \\ \Pr[S(n) = 0 | S(n-1) = 1] & \Pr[S(n) = 1 | S(n-1) = 1] \end{bmatrix} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

While working with these models we have found that the transition matrices are simpler to work with when we interpret the switching to occur immediately after, rather than before, an observation. The result of this is that the first observation in a new regime will represent the end of an interval in the new regime rather than the beginning, and should be accompanied by movement in the level structure. This will result in only one level being available for each regime of $S(n)$, leading to

$$L(n) = S(n)$$

This is of course a simple two-state Markov chain. The process can be specified in a much more general way.

8.3 The General Ladder Model

As before we define the set of levels of the process in as general way as possible with levels

$$\{l_0, l_1, ..., l_{N_{(L)}-1}\} \; where \; l_u \leqslant l_v \; if \; u < v$$

The current regime of the controlling Markov process is given by $S(n)$ and the level of the series is controlled by $L(n)$. The transition probabilities for the Markov chain $(S(n), L(n))$ are given by

$$p_{ij}(u, v) = \Pr(S(n+1) = j \;\&\; L(n+1) = v \mid S(n) = i \;\&\; L(n) = u)$$

This defines a general multi-level function driven by the Markov process $(S(n), L(n))$. In order to restrict it to give us our General Ladder Model we require that the level of the chain $L(n)$ can move only in the direction indicated by the regime $S(n)$

$$\begin{aligned} p_{00}(u, v) &= 0 \text{ if } v > u \\ p_{11}(u, v) &= 0 \text{ if } v < u \end{aligned}$$

We also require that the stable levels points of the ladder act as boundaries for the level chain. Only a switch of regime will free the process from the 'stable' states $(0, 0)$ or $(1, N_{(L)} - 1)$.

$$\begin{aligned} p_{00}(0, 0) &= 1 - \sum_{v=1}^{N_{(L)}-1} p_{01}(0, v) \\ p_{11}(N_{(L)} - 1, N_{(L)} - 1) &= 1 - \sum_{v=0}^{N_{(L)}-2} p_{10}(N_{(L)} - 1, v) \end{aligned}$$

This gives us the general model. Not many restrictions are placed on it at this stage to allow it to encompass several clearly different cases. All we have really done at this stage is to define a chain of levels that switches between two regimes, one of growth and the other of decay. While in the growth regime $(S(n) = 1)$ the level may only increase up to a limit and vice versa for the decay regime. Within this definition several distinct sub-groups have emerged as being practically useful. Each one has slightly different practical benefits.

8.4 The Ladder Model L[$N_{(L)}$,$N_{(L)}$]

This takes the general model and restricts it further. First we make the regime chain $S(n)$ and its switching probabilities independent of the level chain $L(n)$.

For all u we have

$$\begin{aligned}
\sum_{v=0}^{N_{(L)}-1} p_{01}(u,v) &= a \\
\sum_{v=0}^{N_{(L)}-1} p_{00}(u,v) &= 1-a \\
\sum_{v=0}^{N_{(L)}-1} p_{10}(u,v) &= b \\
\sum_{v=0}^{N_{(L)}-1} p_{11}(u,v) &= 1-b
\end{aligned}$$

We then make the movement of the level chain deterministic and dependent only on the regime of the regime chain.

$$p_{00}(u,v) = \begin{cases} 1-a & if\ u=v=0 \\ 1-a & if\ u>0\ and\ v=u-1 \\ 0 & otherwise \end{cases}$$

$$p_{11}(u,v) = \begin{cases} 1-b & if\ u=v=N_{(L)}-1 \\ 1-b & if\ u<N_{(L)}-1\ and\ v=u+1 \\ 0 & otherwise \end{cases}$$

$$p_{10}(u,v) = \begin{cases} b & if\ v=u-1 \\ 0 & otherwise \end{cases}$$

$$p_{01}(u,v) = \begin{cases} a & if\ v=u+1 \\ 0 & otherwise \end{cases}$$

We can observe that it is now impossible for the level process to inhabit states u_0 and $d_{N(L)-1}$. These correspond to the level 0 while in regime 1 and to level $N_{(L)}-1$ while in regime 0. This is a deliberate measure to ensure that the process is easier to work with when inverting matrices and obtaining likelihoods. We declare also that the value of the levels are equally spaced and independent of state

$$l_k = l_0 + \left(\frac{k}{N}\right).(l_{N_{(L)}-1} - l_0)$$

This process, for $N_{(L)}=4$, can be represented by the diagram shown in Figure 8.1.

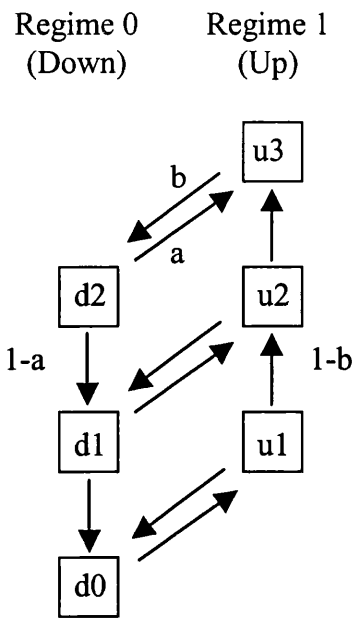

Figure 8.1: A diagram representation of the movement pattern of the Ladder Model, $L[4,4]$

This process, with $l_0 = 0$ and $l_3 = 1$, would have the following set of combined states

$$\textit{The states are } \{d0, d1, d2, u1, u2, u3\} \textit{ with levels } \{0, \tfrac{1}{3}, \tfrac{2}{3}, \tfrac{1}{3}, \tfrac{2}{3}, 1\}$$

The transition matrix between states would be

$$T = \begin{bmatrix} (1-a) & . & . & a & . & . \\ (1-a) & . & . & . & a & . \\ . & (1-a) & . & . & . & a \\ b & . & . & . & (1-b) & . \\ . & b & . & . & . & (1-b) \\ . & . & b & . & . & (1-b) \end{bmatrix}$$

Where a dot symbolises 0 probability and if the distribution at time n is given by f_n then $f_{n+1} = f_n.T$

8.5 The Slide Model S[$N_{(L)}$,$N_{(L)}$]

There are times when the Simple Ladder Model is not suitable. It has a tendency (documented in Chapter 9) to overfit. By overfitting we mean the model returning very high estimates of the regime switching parameters suggesting very frequent switches of regime. In these circumstances we may prefer to utilise a version of the Ladder Model termed the Slide Model. This is identical to the Simple Ladder Model is many respects except that we

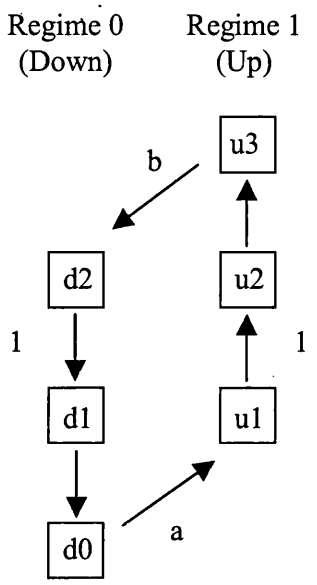

Figure 8.2: A diagram representation of the movement pattern of the Slide Model, $S[4,4]$

introduce one significant restriction. There is far more difference between the stable end levels (l_0 and $l_{N_{(L)}-1}$) and the intermediary levels (l_1 to $l_{N_{(L)}-2}$) that represent the process gradually adjusting from one stable level to the other. In the Slide Model we demand that while the level chain is inhabiting the intermediary levels no change is permitted in the regime chain. So the switching probabilities become

$$\begin{aligned} p_{00}(0,0) &= 1-a \\ p_{01}(0,1) &= a \\ p_{11}(N_{(L)}-1, N_{(L)}-1) &= 1-b \\ p_{10}(N_{(L)}-1, N_{(L)}-2) &= b \end{aligned}$$

The movement rules become

$$\begin{aligned} p_{00}(u,v) &= 1 \quad if\ u>0\ and\ v=u-1 \\ p_{11}(u,v) &= 1 \quad if\ u<N_{(L)}-1\ and\ v=u+1 \end{aligned}$$

The flow chart for the Slide Model $S[4,4]$ can be seen in Figure 8.2 and below we have the transition matrix

$$T = \begin{bmatrix} (1-a) & . & . & a & . & . \\ 1 & . & . & . & . & . \\ . & 1 & . & . & . & . \\ . & . & . & . & 1 & . \\ . & . & . & . & . & 1 \\ . & . & b & . & . & (1-b) \end{bmatrix}$$

There are certain consequences of these added restrictions. One of these consequences is that the regime chain is now dependent upon the level chain and is therefore no longer Markov. Together with the level chain the combined state $(S(n), L(n))$ is still Markov and this is what we model. In order to know the future distribution of the regime chain we require information on the length of time since the last switch.

8.6 The Unravelled Models

We have already mentioned that fact that overfitting can be a major problem for gradual switching models. The introduction of the Slide model was one attempt to deal with this. This prevents the model from trying to fit short deviations from the stable end levels. But sometimes even this approach cannot restrain the model sufficiently when facing significant noise. It is quite easy, however, to take the number of switches the driving process makes using the following adaptation and make it a parameter of the model. It is worth emphasising at this point that this structure is not exactly a model. It is probably better described as a 'method', a method for fitting a Ladder or Slide model when the existing algorithm struggles. It must be established straight away that once we have introduced these adaptions the chains we are working with are no longer Markov.

First we examine an Unravelled Slide model $US[N_{(L)}, N_{(L)}, c]$. We retain $L(n)$ unchanged from the Slide model. We modify $S(n)$ so it consists of $c+1$ levels, where c is the maximum number of switches we will allow. Each of which can be visited only once.

As before we define the transition probabilities using p where

$$p_{ij}(u,v) = \Pr[S(n+1) = j \ \& \ L(n+1) = v \mid S(n) = i \ \& \ L(n) = u]$$

For this model we do not allow a return to regimes visited earlier

$$p_{ij}(u,v) = 0 \qquad if \ j < i$$

Nor do we allow the regime chain to take two switches simultaneously

$$p_{ij}(u,v) = 0 \qquad if \ j > i+1$$

Nor may it exceed the prescribed number of switches ·

$$p_{ij}(u,v) = 0 \qquad if\ j > c+1$$

The final regime becomes an absorbing regime. We then treat the alternate regimes as equivalent. After deciding (for instance) that s_0 represents the growth regime then $\{s(0), s(2), s(4), ..\}$. will all be taken to represent growth regimes with identical effects on the level chain. If we retain this assumption that the first regime represents growth then we have the following transition probabilities

$$p_{00}(u,v) = \{p_{ij}(u,v) : j = i\ and\ i\ is\ even\} = \begin{cases} 1-a & if\ v = u = 0 \\ 1-a & if\ u > 0\ and\ v = u-1 \\ 0 & otherwise \end{cases}$$

$$p_{11}(u,v) = \{p_{ij}(u,v) : j = i\ and\ i\ is\ odd\} = \begin{cases} 1-b & if\ v = u = N_{(L)} - 1 \\ 1-b & if\ u < N_{(L)} - 1\ and\ v = u+1 \\ 0 & otherwise \end{cases}$$

$$p_{01}(u,v) = \{p_{i,i+1}(u,v) : i\ is\ even\} = \begin{cases} a & if\ v = u+1 \\ 0 & otherwise \end{cases}$$

$$p_{12}(u,v) = \{p_{i,i+1}(u,v) : i\ is\ odd\} = \begin{cases} b & if\ v = u-1 \\ 0 & otherwise \end{cases}$$

Let us say we want a Slide process with three levels. We also want to limit the regime chain to make only three switches during the length of the series. We do this by replacing the regime chain we have used before with a regime sequence. If the process is to start in $(0,1)$ say we know that the states $(S(n), L(1))$ visited, in order, will be

$$\underset{start}{(0,1)} \Rightarrow (0,0) \overset{switch1}{\Rightarrow} (1,1) \Rightarrow (1,2) \overset{switch2}{\Rightarrow} (2,1) \Rightarrow (2,0) \overset{switch3}{\Rightarrow} (3,1) \Rightarrow \underset{end}{(3,2)}$$

We know that the Unravelled Slide model will visit each and every one of these states and arrive eventually in $(S(n), L(n)) = (3,2)$. Rather than represent this using the original regime chain we can use a continuous sequence in which each state is visited once and once only. This sequence of states can be indexed

$$\{d0, d1, u2, u3, d4, d5, u6, u7\}$$

Where states separated by 4 steps are identical in terms of properties. So

$$d0, d4 \textit{ are identical}$$
$$d1, d5 \textit{ are identical}$$
$$d2, d6 \textit{ are identical}$$
$$\textit{and so on}$$

When developing a fitting algorithm for use with a computer program it is easier to combine the regime and level chains into one combined state $C(n)$ with states $\{0, 1, ...\}$ equal to combined states $\{d0, d1, ...\}$. The transition matrix for $C(n)$ (which we shall call **C**) is larger than those we have worked with before, seeing as we now have $c.N_{(L)}$ states. The following matrix corresponds to the 4-regime case given above. A dot represents zero probability of transition.

$$C = \begin{bmatrix} . & 1 & . & . & . & . & . & . \\ . & 1-a & a & . & . & . & . & . \\ . & . & . & 1 & . & . & . & . \\ . & . & . & 1-b & b & . & . & . \\ . & . & . & . & . & 1 & . & . \\ . & . & . & . & . & 1-a & a & . \\ . & . & . & . & . & . & . & 1 \\ . & . & . & . & . & . & . & 1 \end{bmatrix}$$

The design of this model is solely to allow us more control of fitting the Slide model. When estimating the likelihood of a solution we ignore all contributions that derive from outside the range of combined states. All that this allows us to do is limit the search for a maximum likelihood solution to the set of regime patterns requiring no more than a certain number of switches c. Once we have these figures for different values of c we can make comparisons of likelihood between different values of c to find the most probable number of switches that will have taken place.

It is also possible to work with an Unravelled Ladder Model. This is the same as the Unravelled Slide Model but uses similar rules for switching regimes and levels as found in the Ladder model (adapted for the restriction of regimes of course). In practice the Ladder Model becomes more useful (relative to the Slide Model) once the constraint of fixed switching is implemented. Under these conditions the unnecessary switching that troubles the Ladder model is dissuaded by virtue of the fixed switch numbers.

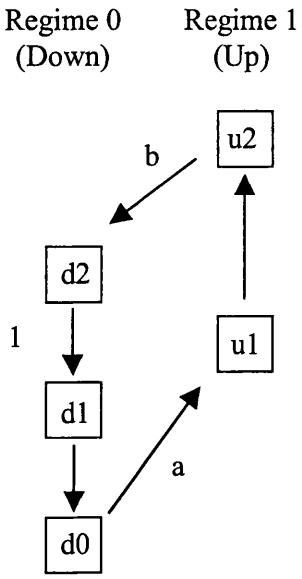

Figure 8.3: A diagram representation of the movement pattern of the Asymmetric Slide Model, $S[3,4]$

8.7 The Asymmetric Models

In an attempt to allow greater flexibility and allow varying transition rates for the different regimes we introduce the Asymmetric Slide and Ladder models. To introduce this asymmetry to the Slide model is fairly straightforward. This is simply the Slide model as we have introduced it but with a different number of levels in the two regimes. We dispense with number of levels $N_{(L)}$ and replace it with $N_{(U)}$ and $N_{(D)}$, which are the number of levels in the up regime and down regime respectively. Figure 8.3 shows a diagrammatical representation of the $S[3,4]$ model. By varying the number of levels in each state we can control the transition rate without removing the deterministic nature of the movement. The transition matrix is equally simple, constructed almost identically to the symmetric case.

Things are not so simple when we start to consider the Asymmetric Ladder model, termed $L[N_{(U)}, N_{(D)}]$. In the symmetric case every level in both regimes could only make a transition to two other levels of the process. Each regime used the same level structure so there was no difficulty in establishing which transitions are possible. When working with an asymmetric Ladder we retain the same rules for movement when no switching occurs but are faced with the problem that the levels do not have equivalents in the opposing regime. In Figure 8.4 we demonstrate one possible solution to the problem for $L[3,4]$. We have had to make a decision as to what transitions are possible from state $d1$. It must be noted that several different (sensible) options may exist and each will lead to its own solution when fitting a model. The greater the number of levels on each side the greater the number of possible variations. We have chosen one particular method for determining the rules of

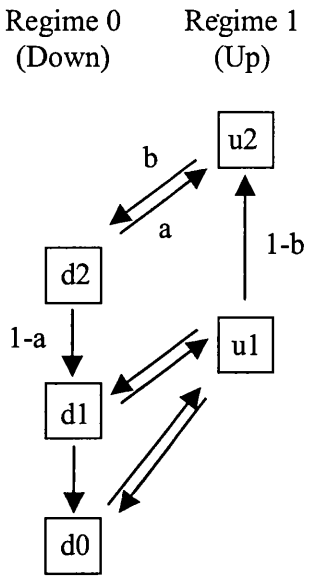

Figure 8.4: A diagram representation of the movement pattern of the Asymmetric Ladder Model, $L[3,4]$

movement which is probably as justifiable as any.

8.8 The Line Fit Model

In some of the Case Studies in Chapter 9 we find ourselves questioning the validity of the results produced by the Bayesian filters. In order to test what correspondence there is between the MLE and the imputed regime sequences suggested by the filters (for the MLE) we shall introduce another model. In many respects it is probably more similar to the Unravelled model than the earlier Ladder models. It is really a form of parameter augmentation, in that we include the regimes occupied during the time series as parameters. Due to the problems optimising a likelihood function with such a large number of parameters we only use it when we have a very good idea of the most credible regime history. We then take the imputed historical sequence of regimes suggested by the filter and maximise the likelihood, making only slight alterations to this sequence of regimes. For an asymmetric Ladder model we term this the Line Fit model and denote it $LF[N_{(U)}, N_{(D)}]$.

For the Bayesian filter we obtain two pieces of output. The likelihood is obtained using

$$f(\mathbf{y}_N|\Theta, \Pi)$$

where $\mathbf{y}_N$ is the full set of N observations, Θ are the parameters of the model and Π are

the initial conditions. We also obtain the inferred probabilities

$$\Pr[S(n) = i \mid \mathbf{y}_N, \Theta, \Pi]$$

Due to the way these probabilities are obtained by the filter there is no guarantee that the sequence suggested by these probabilities is even possible. There is another way to obtain a maximum likelihood estimate of the parameters, one we touched upon when first introducing the filters in Section 2.2. We can expand our parameter set Θ to include the regimes occupied by the signal $\{s(n)\}$. We define this new parameter set

$$\Sigma = [\Theta, \mathbf{s_N}] \qquad \text{where } \mathbf{s_N} \text{ is a regime sequence } \{s(1), s(2), ..., s(N)\}$$

Our likelihood function for this parameter set is

$$f_{\Sigma}(\mathbf{y}_N|\Sigma, \Pi)$$

We shall call the values of Σ obtained by maximising this likelihood distribution the 'sequence MLE' or MLE_{Σ}. The reason we chose not to use this version of the likelihood function is the very large number of additional parameters it introduces. This makes the likelihood function very difficult to maximise. To combat this we shall only introduce this approach when we already have obtained the sequence of probabilities from the basic filter. In this case we ensure that our initial proposal for $\mathbf{s}_N$ is likely to be close to the best sequence.

The process of obtaining this likelihood is much simpler than for the basic filter requiring only the following calculation (where there is no autoregressive noise).

$$\begin{aligned} f_{\Sigma}(\mathbf{y}_N|\Sigma, \Pi) &= \Pr[S(1) = s(1) \mid \Pi].f(y(1)|S(1) = s(1)) \\ +\sum_{n=2}^{N} \Pr[S(n) &= s(n) \mid S(n-1) = s(n-1)].f(y(n)|S(n) = s(n)) \end{aligned}$$

Of course the other advantage of obtaining MLE_{Σ} is that we can plot the sequence obtained from this against the original data.

8.9 Variations of the Ladder Models

Many different variations on this Ladder theme are possible. In this Section we shall detail some of the ways we can generalise the model further.

8.9.1 Level Structure.

We have generally considered only the case where the levels of the ladder are equally spaced and independent of the current state. This is not a necessary requirement. We could take

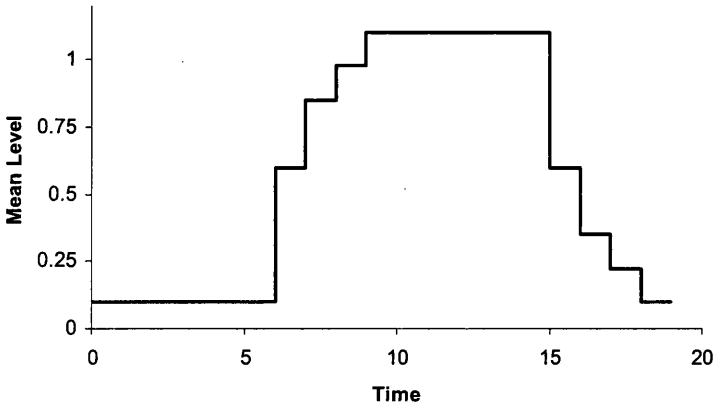

Figure 8.5: A representation of a regime transition for a $S[4,4]$ with a variable level structure.

a level structure where each level was a fixed proportion of the level before

For instance

$$\{u0, u1, u2, u3, u4, d4, d3, d2, d1, d0\} = \{0, \tfrac{1}{2}, \tfrac{3}{4}, \tfrac{7}{8}, 1, 1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, 0\}$$

Which would look a little something like Figure 8.5. If enough levels are used and the transition probabilities are chosen appropriately we could use this as a discrete time, discrete level approximation to the FM model.

8.9.2 Switching Behaviour.

The Asymmetric models are not the only way we can vary the transition times between regimes. Another way we can achieve this is to vary the rate of transition between the major levels. In the previously stated models movement must occur when the process is not in one of the stable end states. A Ladder model incorporating variable transition rates would allow (but not demand) this movement by introducing two new parameters p and q,

which represent the probability of moving when no switching occurs.

$$p_{00}(u,v) = \begin{cases} 1-a & if\ v=u=0 \\ (1-a)p & if\ u>0\ and\ v=u-1 \\ (1-a)(1-p) & if\ u>0\ and\ v=u \\ 0 & otherwise \end{cases}$$

$$p_{11}(u,v) = \begin{cases} 1-b & if\ v=u=N_{(L)} \\ (1-b)q & if\ u<N_{(L)}\ and\ v=u+1 \\ (1-b)(1-q) & if\ u<N_{(L)}\ and\ v=u \\ 0 & otherwise \end{cases}$$

$$p_{01}(u,v) = \begin{cases} a.p & if\ u<N_{(L)}\ and\ v=u+1 \\ a(1-p) & If\ v=u \\ 0 & otherwise \end{cases}$$

$$p_{10}(u,v) = \begin{cases} b.q & if\ u>0\ and\ v=u-1 \\ b(1-q) & if\ v=u \\ 0 & otherwise \end{cases}$$

This would allow more flexibility in modelling the switch since the switches do not need to be of fixed length. It also allows us to retain the variable transition rates of the asymmetric model without any of the problems in determing the transition patterns. Of course there is no limit to the number of parameters that could be added. In theory you could work with movement and switching probabilities unique to each level of the process although this process would be hard to fit without including other constraints due to the large number of degrees of freedom.

8.10 Summary

In this Chapter we have introduced a class of discrete time, discrete level Markov chains for the purpose of extending the two-regime Markov switching model to incorporate gradual switching between the regimes. Many variations are possible on the examples we mentioned including for instance Asymmetric Unravelled models. The models are not too complex in their structure as we have already found that the earlier FM process was rather difficult to fit to data where noise was present due to the great flexibility of the structure. They should still, however be capable of capturing evidence of gradual switching mechanics, if it exists, for real time series data.

Fitting the Ladder Models is fairly straightforward. We simply adapt the smoother developed by Hamilton and outlined earlier, which obtains a Maximum Likelihood measurement of the data set. As we have given examples of the use of the Filters in Section 2.2.1 and again in Section 5.3 we shall not repeat ourselves. Obtaining the MLE will then be a simple process of optimisation using method discussed in Chapter 3.

Chapter 9

Fitting the Model to Real Data

In earlier chapters we have specified a range of gradual switching models. Of these the Ladder models have proved to be the most stable, and so in this chapter we shall attempt to apply these models to real time series in order to model gradual switching behaviour.

9.1 Identifying Gradual Switching

As we have observed earlier, the field of non-linear time series models is very broad. We cannot expect one model to span the whole field and instead should judiciously select an appropriate model for a given set of data. But what characteristics will lead us to consider a gradual switching model? There are certain properties that we would expect from any Markov Switching model and these should be apparent, but for this particular class of models to be a prime candidate we require something more.

The key feature of any switching model is that it allows more than one type of behaviour. Whenever it appears that we can divide the data set into shorter series, each displaying different behaviour, then some kind of switching may be required. If many of the subsets appear very similar to each other then it is possible that a regime switching model would be appropriate. An example of this can be seen in Figure 9.1. The data, taken from Nelson (1973), show the civilian unemployment rate in the United States between 1890 and 1974. It is quite apparent that the usual level of the series for the observation period is around 5%, except for two intervals. The first of these intervals corresponds with an economic depression which started with the 'Panic of 1893' and did not fully abate until 1897. The second exceptional interval matches the 'Great Depression' of 1930-1939. In this example it seems quite sensible to consider a regime switching model, not only from structural considerations but also from visual inspection of the data. The simplest form of regime switching we could consider in this case would link the mean level of the process to the regime occupied.

If this kind of model reflected the trend of the process rather than the mean level we might describe the process as having a Markov Trend. In this case the graph would be

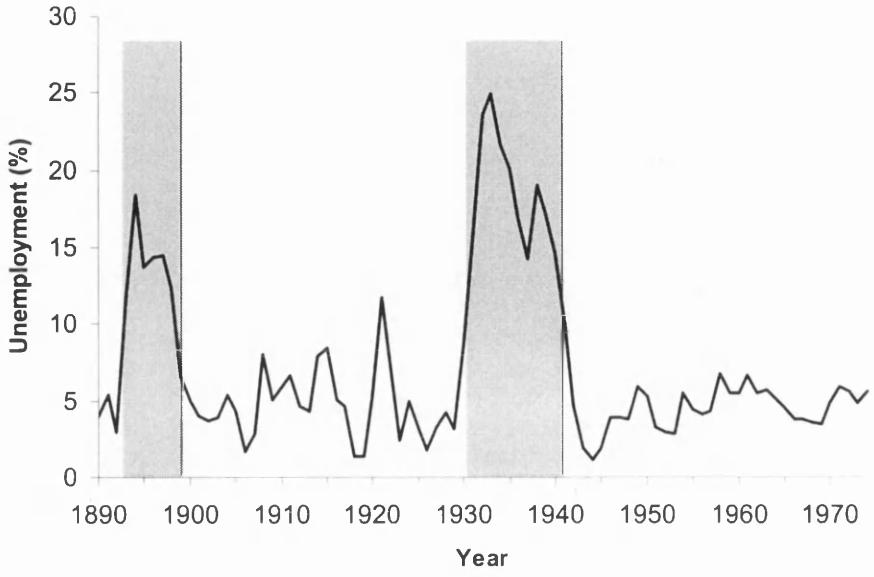

Figure 9.1: The rate of Civilian Unemployment in the US between 1890 and 1974 with possible abnormal regimes marked.

characterised by different gradients in different sections. This zig-zagging behaviour can be seen in another data set taken from Nelson (1973). The data, displayed in Figure 9.2, shows the monthly figures for automobile registrations from 1947 to 1968. Superimposed on the graph are straight lines, representing the approximate gradient. These lines are estimated visually and serve only to illustrate the apparent tendency for the series to alternate between two linear growth rates, one positive and the other negative. If the lengths of these sojourns are predictable for each regime then the series might be more suited to a seasonal trend model. If, as in this example, the sojourn lengths are variable then a Markov switching model may be more appropriate. Whether or not this pattern of growth and decay represents any meaningful structural component is unclear, but it shows that a regime switching model may indeed be a useful tool in modelling the growth rate of this particular time series.

We now have seen examples of the kind of behaviour that would lead us to consider the family of Markov switching models. They may have Markov level or Markov trend and may have complex autoregressive processes to model the residuals. In this research we have presented various models incorporating Gradual Switching dynamics. As there is limited research in this field there is not a wealth of examples to call upon. If we consider how this gradual switching would affect the case of a Markov trend pattern we should expect to see the same characteristic regime pattern of the previous example but with evidence that the gradient changes slowly rather than instantaneously. One such data set that may serve as useful evidence in that used by Makridakis et al. (1998) It shows the real daily

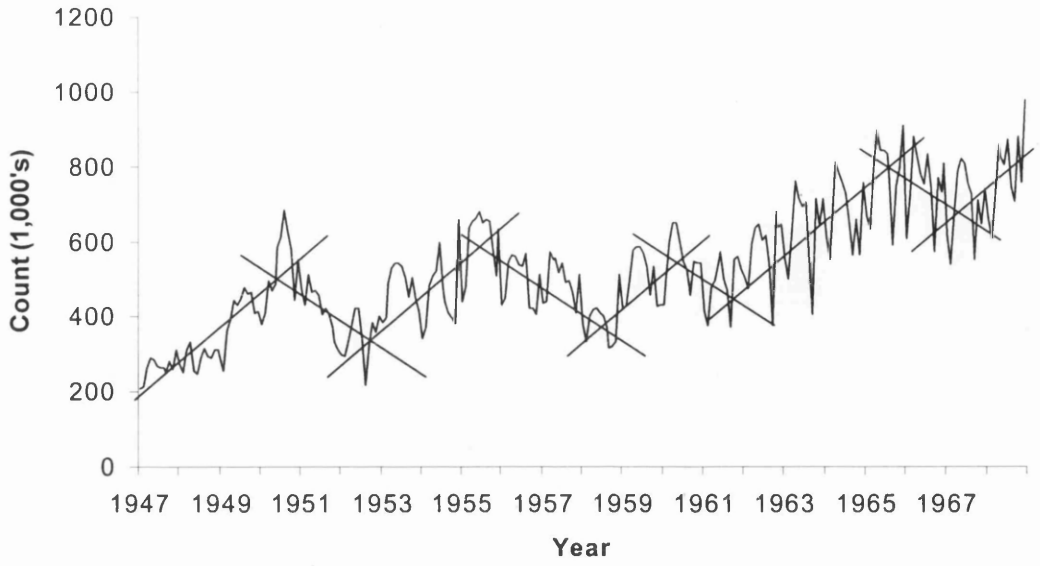

Figure 9.2: The number of car registrations per month in the United States between 1947 and 1968. Superimposed on the data are two linear growth regimes.

wages in the U.K. between 1264 and 1994, a period spanning over 700 years. Several things are immediately apparent in this time series, shown in Figure 9.3. First, we see a clear separation of the period into regimes. It is also clear that these regimes cannot be said to have constant gradient. There is a way that we could use the switching model to incorporate this kind of behaviour. That is if we allow the process to change gradually from one state to another. While the straight lines in the graph represent possible gradients for the positive growth regime the fitted line is actually a realisation of a gradual switching model, fitted by eye only, to demonstrate it's capacity to model these dynamics. There is no suggestion that this kind of model is the only option when trying to work with a time series like this, only that is provides another possible approach. In order to defend against the accusation of overfitting it should be noted that the fitted line only inhabits six growth states (three positive and three negative) during the whole 735 values in this times series.

So far the modelling has been by eye only, for illustrative purposes. The next step is to select several data sets that may, or may not, display gradual switching. To each of these we shall apply various Ladder and Slide models in an attempt to identify gradual switching mechanics and then model it.

The data sets we have selected are:

1. US Unemployment Figures.
2. The Diameter of Ladies Skirts.
3. UK Real Wages.

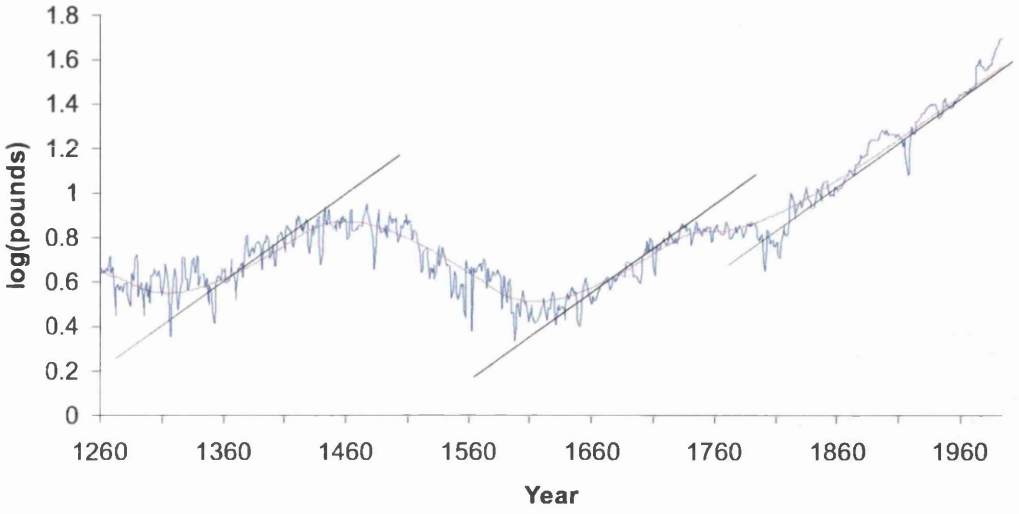

Figure 9.3: The level of the real daily wage in the UK over a period of around 700 years. Superimposed on the data are straight lines, representing a possible linear growth regime, and a sample realisation of a gradual switching model.

4. US postwar GNP (to which Hamilton fitted the original Markov Switching model)

9.2 Case 1: U.S. Unemployment

We have already introduced this data set as one that is ideal for demonstrating regime changes. Given the ideal nature of the data set we would expect ideal results. Fortunately the data (and method) do not disappoint us.

9.2.1 A Visual Inspection

On examining the data, shown in Figure 9.1, we can see clear evidence of regime change behaviour. There is the suggestion that the two deviations from the normal regime may be at different levels. As we want an easy introduction to the use of these models we transform the data to our advantage. In this case the transformation consists of taking logs to bring them closer without making major, and unnatural adjustments to the data. This ensures that we do not have to consider using any more than two regimes.

9.2.2 The LadderModel

As one would expect we will start at the beginning, with the first model we specified, the Ladder model. In specifying the model we make certain decisions regarding the exact form of the transition matrix and level structure. These assumptions are given below:

Ladder Model	Log-Likelihood
$L[2,2]$	12.4023
L[3,3]	**15.0410**
$L[4,4]$	14.5895
$L[5,5]$	13.7007
$L[6,6]$	12.4378
$L[7,7]$	11.6724

Table 9.1: The Maximised Log-Likelihood of the US Unemployment data when modelled using the Ladder model with different number of levels

- Movement between the levels within a regime is considered as compulsory when not at the upper or lower end of the level structure ($p = q = 1$). This is the original Ladder model we specified.
- The regime indicated by $S(n)$ is presumed to be the regime occupied during the interval $t \in (n-1, n]$. As a result if the regime of the signal changes the level of the signal will also change.
- In order to find a good starting distribution for the filter we use an approach suggested by Hamilton. We consider the series in reverse, starting with some vague prior distribution for the regimes occupied in the final few steps. We then apply the filter to the reversed series using our chosen parameter estimates. The purpose of this is to recover a convincing distribution of regimes for the first few values of the series. These distributions are then used as our prior distribution when we come to run the filter forwards. These distributions are almost completely independent of our chosen prior and tend to be quite good, arising as they do from the data itself. It also reduces the number of parameters that need to be estimated.

We vary the number of levels in the Ladder, $L[N_{(L)}, N_{(L)}]$, for both regimes and obtain log-likelihood values that are shown in Table 9.1. The maximum likelihood (ML) occurs for a ladder that has 3 levels. This is not exactly a lengthy ladder but this should not draw attention from the fact that the models with greater numbers of steps (4 or 5) receive more support than the basic two-regime Markov model, which is included in the table as $L[2,2]$. If we perform an identical analysis using the Slide model with the same conditions we obtain almost identical results. The results of this are given in Table 9.2. Plotting the log-likelihood values for both models together in Figure 9.4 we see clear evidence of some kind of gradual switch

We shall take a ladder of length 3 as the optimum, as suggested by the likelihood. The parameter estimates obtained for $L[3,3]$ are given in Table 9.3. We can now move on to look at the inferred distribution of states when the parameter values correspond the those of the MLE.

Slide Model	Log-Likelihood
$S[2,2]$	12.4077
S[3,3]	**15.0378**
$S[4,4]$	14.9827
$S[5,5]$	14.4197
$S[6,6]$	14.2181
$S[7,7]$	13.0513

Table 9.2: The Maximised Log-Likelihood of the US Unemployment data when modelled using the Slide model with different number of levels

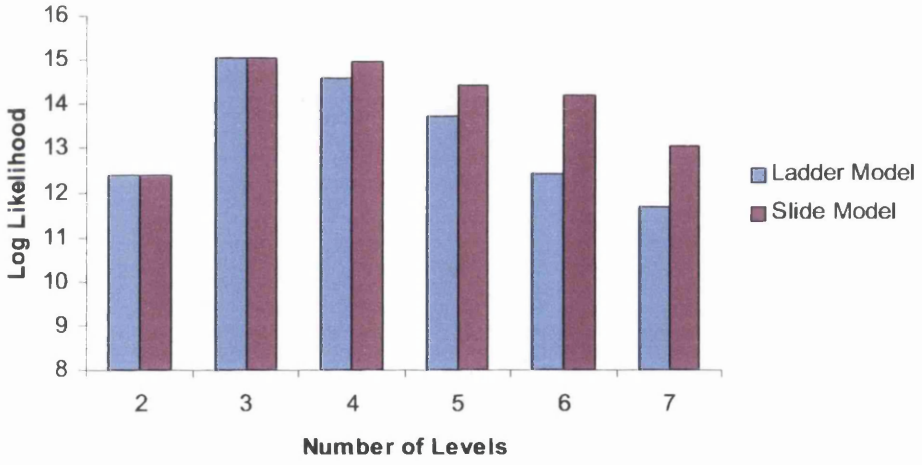

Figure 9.4: A comparison of the Maximised Log-Likelihood for the US Unemployment data for the Slide and Ladder Models with different numbers of levels.

Parameter	a	b	l_0	l_2	σ
Value	0.0269	0.1294	0.6216	1.2356	0.1720

Table 9.3: The Maximum Likelihood estimates of the parameters of the L[3,3] model for the US Unemployment series.

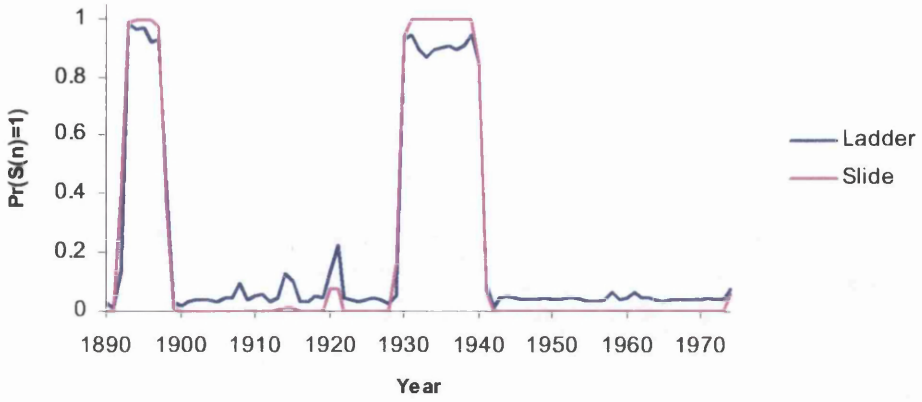

Figure 9.5: A comparison of the inferred probability the signal is in the upper regime ($S(n) = 1$) for the Slide and Ladder models.

The two lines shown in Figure 9.5 represent the inferred probability of being in the upper (atypical) regime for the Ladder and Slide models. It is clear that the greater flexibility of the ladder model leads to (marginally) less clarity in the distinction between regimes. For instance, a couple of possible aborted switches could be indicated around 1915 and 1920. By an aborted switch we mean a regime switch occuring while the process still occupies an intermediary level. Finally we can obtain the sequence (and corresponding parameter estimates) that maximise the sequence likelihood using $LF[3, 3]$. We take the results of the basic filter and then use a hill climbing algorithm to find MLE_Σ. We display the results of this in Figure 9.6.

9.2.3 Conclusion

With this data we found exactly what we were looking for; a short and friendly data set to allow us to trial the models. Furthermore there is some evidence of gradual switching between the two regimes. Both of the models (Slide and Ladder) suggest a three level chain. This would mean that the transition from one of the stable end levels to the other takes 2 months to complete.

9.3 Case 2: Skirt Diameter

Our second case study data set is an unusual one, taken from Roberts (1992). The data points are measurements (in mms) of the diameter of ladies skirts (at the hem) over a 100 year period. It has been chosen as it is short, unusual and demonstrates some of the difficulties of working with gradual switching models. The subject of this time series is hardly one that would spring to mind when considering a Markov Switching model. Its

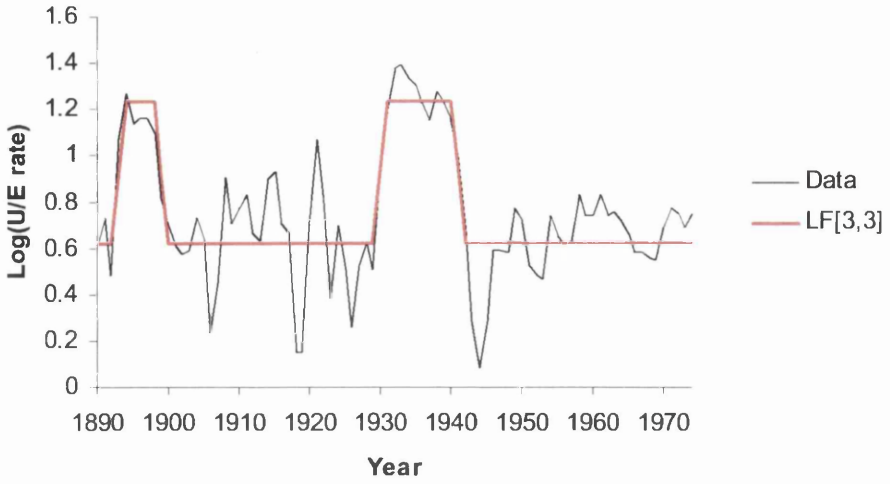

Figure 9.6: A comparison of the time series data with the Sequence Maximum Likelihood path.

inclusion here is to demonstrate evidence supporting the existence of gradual switching in a wide variety of sources. In the first case study we modelled the level of the observed series rather than the growth rate. In this case we shall be modelling the growth rate. The two key pieces of 'evidence' that suggest we may be able to model the data are:

1. The apparent similarity of the (magnitude of the) gradient during uninterrupted periods of increase or decrease.

2. The 'rounded' nature of the transition from the rising to falling states.

With the decision made to apply a gradual switching model we choose the simplest form of the method to apply. In general this will be the Ladder model. When we look at the data (see Figure 9.7) we see evidence of two different gradients, one positive and one negative. The curving nature of the intervening sections show evidence of a gradual switch between these two regimes. As such we shall model for the existence of Markov trend rather than Markov level by taking the first difference of the data. A plot of the differenced data is given in Figure 9.8.

9.3.1 The Ladder Model

We begin by fitting the Ladder model for different numbers of levels. We make the usual restriction requiring a change in level to accompany a change in regime and obtain our prior from the data by reversing the filter. Unfortunately when we apply the model to the data we obtain multiple solutions dependent upon our choice of starting point for the maximistion algorithm. These multiple solutions are documented in Table 9.4. When we examine these

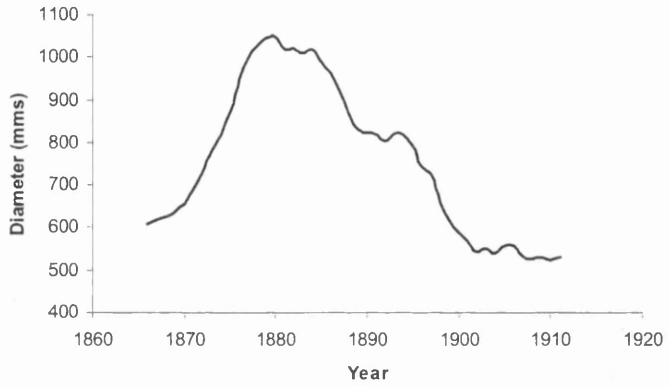

Figure 9.7: Time series data displaying the diameter of ladies skirts (at the hem) over a period of nearly 50 years

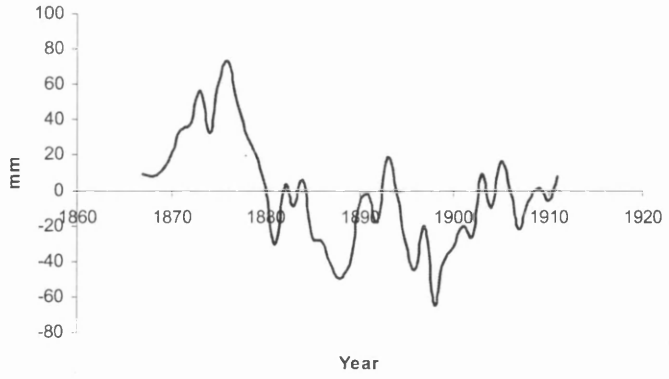

Figure 9.8: Time series data displaying the first difference of the diameter of ladies skirts (at the hem) over a period of nearly 50 years

Model	Solution	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_{N_{(L)}-1}$	$\hat{\sigma}$
L[2, 2]	1	-201.7133	0.0000	0.0725	20.2169	-16.3193	34.2788
L[2, 2]	2	-211.7424	0.8281	0.9586	24.0596	-22.2704	30.4297
L[3, 3]	1	-202.0927	0.0369	0.1242	19.0928	-15.2822	45.3334
L[3, 3]	2	-206.4237	0.7487	1.0000	21.5052	-23.3298	38.2186
L[4, 4]	1	-200.7379	0.0002	0.0888	19.2300	-17.1531	38.8454
L[4, 4]	2	-200.0680	0.5434	0.6240	12.4394	-36.2837	49.3603
L[5, 5]	1	-199.5216	0.0001	0.0886	18.8292	-17.2097	42.2922
L[5, 5]	2	-195.2566	0.5458	0.5354	10.4527	-44.0679	51.2415

Table 9.4: The Maximised Log-Likelihood of the Skirt Length data when modelled using the Ladder model with different number of levels

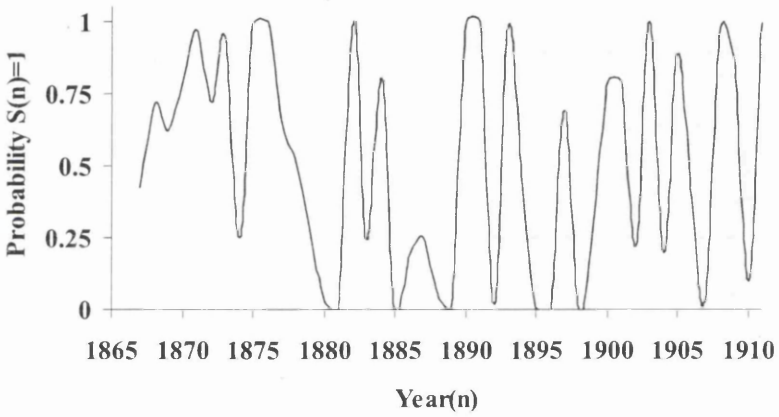

Figure 9.9: The inferred probabilities of the signal being in the upper regime ($S(n) = 1$) for the L[6,6] model.

solutions we find those with higher likelihood do not fit in with our interpretation of the time series as one characterised by regime switching.

The multiple solutions arise from using different starting positions for the hill climbing algorithm. There are solutions (with a close to 0) that represent the cases where the fitted model changes regime only once and those (with $a, b > 0.5$) that represent a frenzied case of 'over-switching'. This is obvious from looking at the inferred state distribution for the $L[6, 6]$ model, given in Figure 9.9. The filter is usually clear as to which regime the process occupies but allows far too many switches to allow the pattern of sojourns to fit happily into our idea our of regime switching solution.

9.3.2 The Over-Switching Phenomena

The important questions are 'Why does this happen?' and 'How can we prevent it?'. It appears that, when faced with a high level of noise relative to the difference in levels, we

find that the process fits to the noise rather than any likely regime pattern we identified by eye. This results in very high regime switching probabilities and a process that does not exhibit long sojourns in one regime. This does not really tally with the idea of a regime switching model. It is also the case that as so many more switches take place there is also a much greater number of credible regime patterns which explains the plethora of local maxima. Overfitting seems to be a major problem when working with gradual switching models and measures need to be made to overcome it. One obvious solution would be to introduce prior distributions for the various parameter values. We shall try to avoid doing this as we feel that doing so would only lead us to a solution for that particular set of priors.

9.3.3 The Slide Model

Given that we are unlikely to obtain any satisfactory results using the Ladder model perhaps we should try to apply the Slide model. This model takes a more extreme approach to preventing over-fitting. That approach is in demanding that once a change in regime starts, the gradual switch runs its course and no further switching is allowed until the level of the chain has reached the other end of the ladder. As before we have compulsory movement, switches accompanied by movement and priors obtained from an inverted process. As we apply the Slide Model we find the hill climber much clearer about the MLE for a given ladder length. The problem of multiple local maxima appears to have receded somewhat but it may have been replaced by another problem. We can see from the results in Table 9.5 that the log-likelihood is rising but a glance at the MLEs suggests something is wrong. By the time we reach $S[10, 10]$ we find the switching intensities are very close to 0 and 1, meaning that the process never remains in the upper state and never leaves the lower state. There is little point in continuing to try higher numbers of levels of the signal. The overall conclusion drawn after examining the inferred distributions of $s(n)$ suggest that only two switches are involved, one during the series and one shortly before it starts. Despite this the solution obtained from maximising the sequence likelihood for only two switches is disappointing, as we can see in Figure 9.10. Many of the features that the model is completely failing to capture could be modelled more effectively using incomplete switches but the Slide model is structured to exclude these. We have already tried the Ladder model and found other factors rule it out. It is therefore necessary to turn to the Unravelled model to constrain the Ladder to prevent over-fitting.

9.3.4 The Unravelled Ladder

The final step in our journey is to see if constraining the maximum number of switches used by the Ladder model to some 'reasonable' number will assist us in fitting a Ladder mode to this data. As far as possible we would hoped to avoid making such subjective distinctions and allow the model to fit itself but there appears no alternative in this case (if we wish to progress with these models). The results of fitting this model are given in Table

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_{N_{(L)}-1}$	$\hat{\sigma}$
S[2, 2]	-202.2190	0.0000	0.0701	20.4443	-16.0515	33.3300
S[3, 3]	-201.7149	0.0000	0.0719	20.2170	-16.3251	34.2659
S[4, 4]	-201.5368	0.0000	0.0773	20.1649	-16.4261	34.9702
S[5, 5]	-200.4452	0.0001	0.1115	19.2120	-17.1851	38.8432
S[6, 6]	-199.0826	0.0000	0.1290	18.8355	-17.2116	42.2086
S[7, 7]	-198.0626	0.0000	0.1592	18.6009	-17.2790	45.4674
S[8, 8]	-197.3313	0.0000	0.2132	18.5042	-17.3066	49.0717
S[9, 9]	-196.8990	0.0000	0.3354	18.6316	-17.2007	53.4108
S[10, 10]	-196.2009	0.0000	1.0000	18.6854	-17.1097	60.5842

Table 9.5: The Maximised Log-Likelihood of the Skirt Length data when modelled using the Slide model with different number of levels

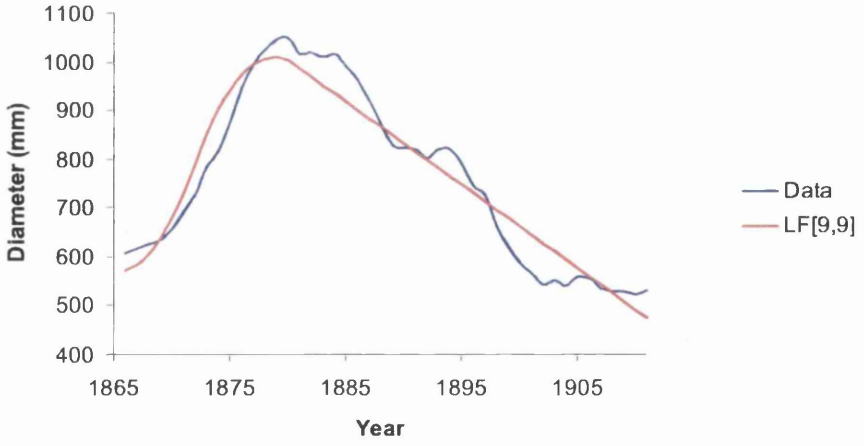

Figure 9.10: The series corresponding to the SMLE sequence using the two-switch pattern suggested by the Slide model.

No. of Switches	No. of Levels 6	7	8	9	10	11	12
1	-199.62	**-199.58**	-199.80	-200.05	-200.36	-200.87	
2	-200.71	-200.47	-200.38	**-200.36**	-200.46	-200.69	-201.05
3	-200.91	-200.73	-200.69	**-200.66**	-200.69	-200.85	
4		-200.74	-200.37	-200.05	**-199.96**	-200.11	
5			-200.83	-200.03	-198.99	-198.26	**-198.24**
6					-195.16	**-194.50**	-195.02
7					-196.75	**-196.12**	-196.59
8				-195.27	**-194.90**	195.25	
9				-196.62	**-196.31**	196.71	

Table 9.6: The Maximised Log-Likelihood of the Skirt Length data when using Unravelled Ladder models with different number of levels and switches.

Table 9.6. There is plenty of evidence of some kind of gradual switch in the series data. For most different number of switches the model suggests a Ladder Length of between 9 and 12 steps. The maximum likelihood is recorded for 6 switches, combined with 11 levels of the ladder. Figures 9.11 and 9.12 show the log-likelihood and likelihood respectively. What is clear when plotting the log-likelihood becomes even more striking when plotting the likelihood. Given that 6 switches (during the observed series) gives the best performance closely followed by 8 we shall examine the sequence likelihood in each case. The two Unravelled Ladder models we shall persist with are $UL[11, 11, 6]$ and $UL[10, 10, 8]$, that is, 6 switches of a ladder with 11 levels and 8 switches of a ladder with 10 levels.

9.3.5 Determining Switch Points

Before we can begin working with the sequence likelihood we should try to identify the likely positions for switches between regimes. This will help us avoid the problems associated with local maxima in the likelihood function. In order to do this we shall examine the inferred regime for each point of the time series suggested by the Basic Filter. The inferred probability of being in regime 1 (the upper regime) is shown in Figure 9.13 for the six and eight switch cases.

The two lines match very closely with the eight switch case trying to model the slight feature in a smooth looking graph around 1881. For six switch points we estimate the switches to occur at {1877,1889,1894,1899,1906,1908} while for the eight switch case we would estimate {1877,1882,1885,1889,1894,1899,1906,1908}. It is worth noting that when we differenced this series we took the difference between a point and the preceding one, so these switches may indeed have occurred a year earlier than this.

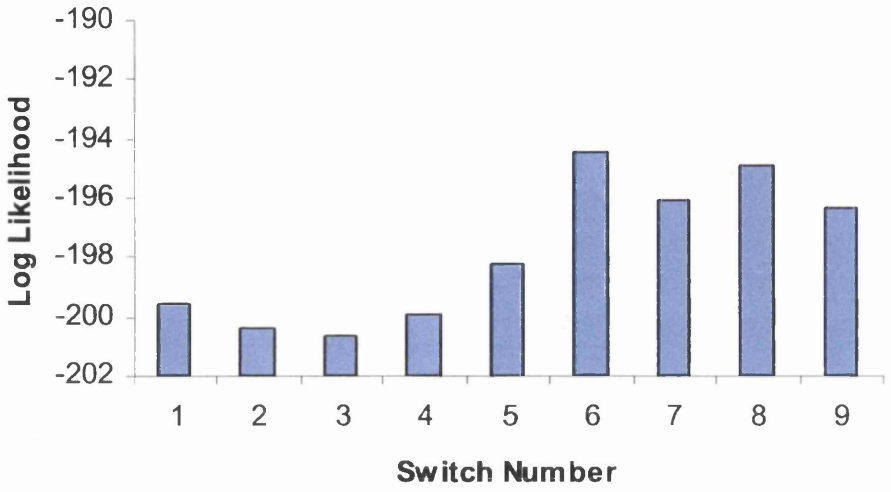

Figure 9.11: The maximised Log-Likelihood for the Skirt data using the Unravelled Ladder model for the optimum ladder length for each number of switches (within the observed series)

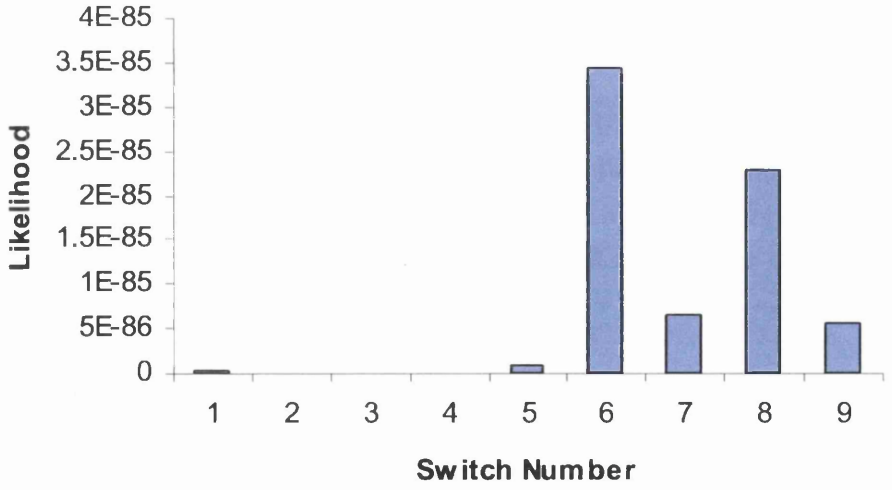

Figure 9.12: The maximised Likelihood for the Skirt data using the Unravelled Ladder model for the optimum ladder length for each number of switches (within the observed series)

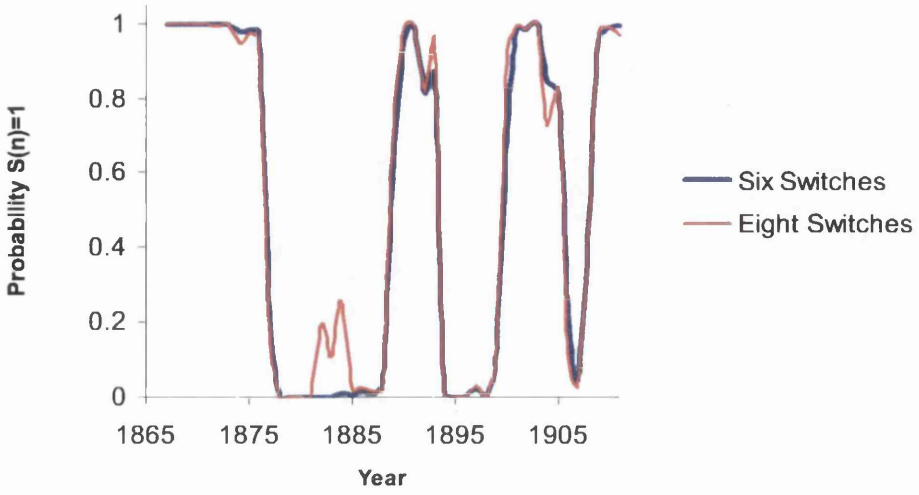

Figure 9.13: A comparison of the inferred probability that the regime of the driving signal is in the upper regime ($S(n) = 1$) between $UL[11, 11, 6]$ and $UL[10, 10, 8]$.

Parameter	a	b	l_0	$l_{N_{(L)}-1}$	σ
UL[11, 11, 6]	0.12	0.15	49.99	-42.46	12.98
UL[10, 10, 8]	0.15	0.21	52.29	-45.65	12.34

Table 9.7: The maximum likelihood parameter estimates for the Skirt Length data using the Unravelled Ladder model.

9.3.6 Imputed Sequence

We can use these parameter estimates and approximate switch points as initial values for maximising the sequence likelihood. We shall then plot two lines against the data for Unravelled models utilising both 6 and 8 switches.. These series will correspond to those suggested by the MLE and the sequence MLE. In Figure 9.14 we show the two series for the six switch case and in Figure 9.15 we show the two series for eight switches.

9.3.7 Conclusions

We can obtain a reasonably good fit for the data using as few as 6 switches between two regimes, one representing falling skirt diameters and the other rising skirt diameters. The parameter estimates in each case (using the Line Fit model) can be found in Table 9.7. For this data set there is no suggestion that the choice of model has any significant structural meaning. It may be that there is a precise relationship between cycles of fashion and that simple mathematical rules are followed but this does not need to be the case. This data set was chosen to demonstrate the versatility of this kind of model and test the fitting algorithms and the use of the Unravelled ladder.

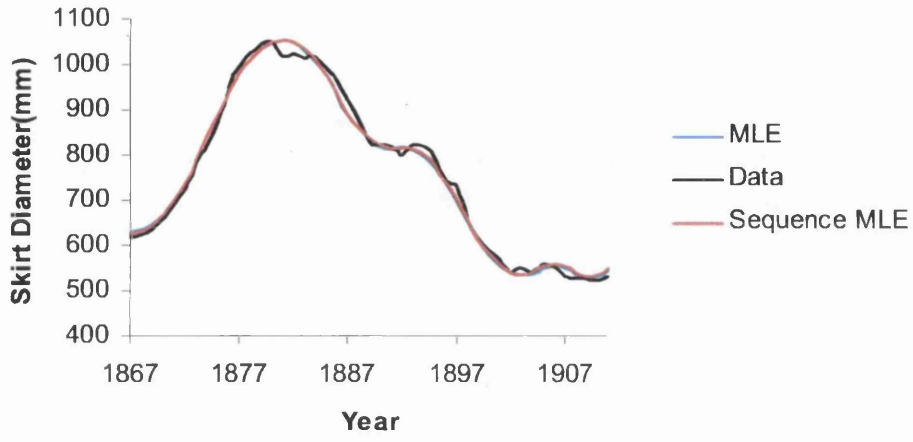

Figure 9.14: A comparison of the series suggested by the MLE and the sequence MLE for a $UL[11, 11, 6]$ model.

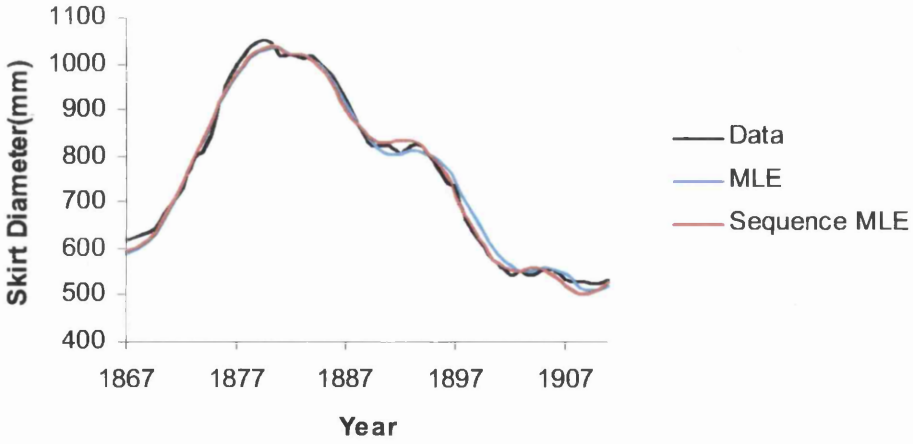

Figure 9.15: A comparison of the series suggested by the MLE and the sequence MLE for a $UL[10, 10, 8]$ model.

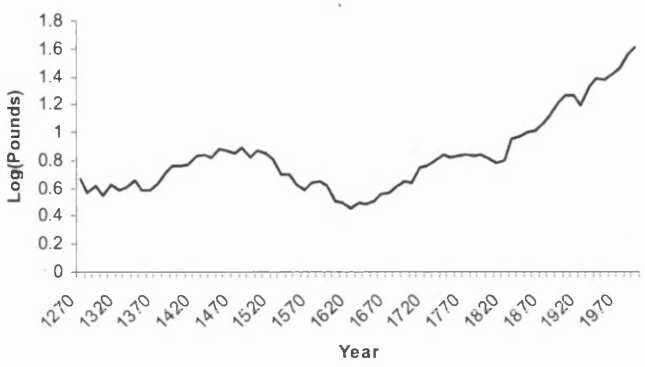

Figure 9.16: An annual figure for the level of UK Real wages, measured in Ln(pounds) between 1260 and 1994.

9.4 Case 3: UK Real Wages

For our third case study we have chosen a time series representing the real daily wages in the U.K. over much of the last millenia. The source of the data is Makridakis *et al.* (1998). We do not work with the series itself which consists of 735 annual measurements covering the period 1260 to 1994. As a result of a visible non-linearity in the series we will takes logarithms of the data and then difference to allow us to work with the growth rate. Due to the size of the data set and the considerable noise to gradient ratio we shall work with a series of 10 year averages (not a 10 year moving average), a plot of which can be seen in Figure 9.16. The order in which these transformations were made was to first construct the 10 year averages, then take logarithms and difference. It is this differenced series we shall analyse. We can see alternating periods of growth and decay of the real daily wage. It is hoped that the model will identify this apparent cycle and also confirm that the change between these two periods is gradual rather than abrupt.

9.4.1 The Ladder Model

We first apply the Ladder model to the data series and examine the results. The maximised likelihood values for the Ladder model with different number of levels can be found in Table 9.8, and is presented visually in Figure 9.17. It is clear that the maximum value occurs for a ladder length of 13. If we look at the sequence of inferred distributions for 'up' or 'down' state we find the probable number of switches. These distributions (shown in Figure 9.18) suggest two full switches (around 1380 and 1500) and several more incomplete switches (shortly before the series starts at 1270 and around 1720). It is also possible to detect the slight suggestion of a switch around 1880 but it seems unlikely that this will be very influential.

Ladder Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_{N_{(L)}-1}$	$\hat{\sigma}$
L[6, 6]	111.79	0.1303	0.0544	-0.022	0.0315	0.048
L[7, 7]	111.90	0.1354	0.0567	-0.024	0.032	0.048
L[8, 8]	111.98	0.1379	0.0575	-0.026	0.0324	0.0479
L[9, 9]	112.04	0.1382	0.0572	-0.027	0.0328	0.0478
L[10, 10]	112.07	0.1345	0.0549	-0.029	0.0331	0.0478
L[11, 11]	112.10	0.1284	0.0513	-0.032	0.0332	0.0477
L[12, 12]	112.16	0.1067	0.0434	-0.036	0.0332	0.0475
L[13, 13]	112.21	0.0933	0.0394	-0.038	0.0337	0.0474
L[14, 14]	112.17	0.0867	0.0379	-0.04	0.0342	0.0474
L[15, 15]	112.06	0.0862	0.0379	-0.042	0.0343	0.0474
L[16, 16]	111.91	0.0897	0.0387	-0.046	0.0343	0.0475
L[17, 17]	111.81	0.0947	0.0388	-0.057	0.0336	0.0475
L[18, 18]	111.78	0.0973	0.039	-0.066	0.0333	0.0475
L[19, 19]	111.78	0.0993	0.0393	-0.073	0.0333	0.0475
L[20, 20]	111.79	0.098	0.039	-0.079	0.0333	0.0475

Table 9.8: The Maximised Log-Likelihood of the UK Wages data when using Ladder models with different number of levels and switches.

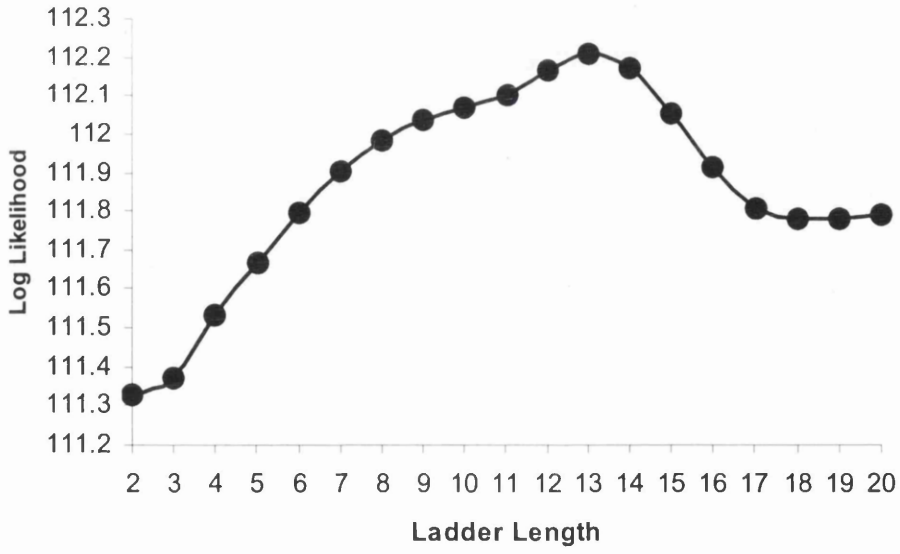

Figure 9.17: A plot of the Maximised Log-Likelihood for the Ladder Model with different numbers of levels.

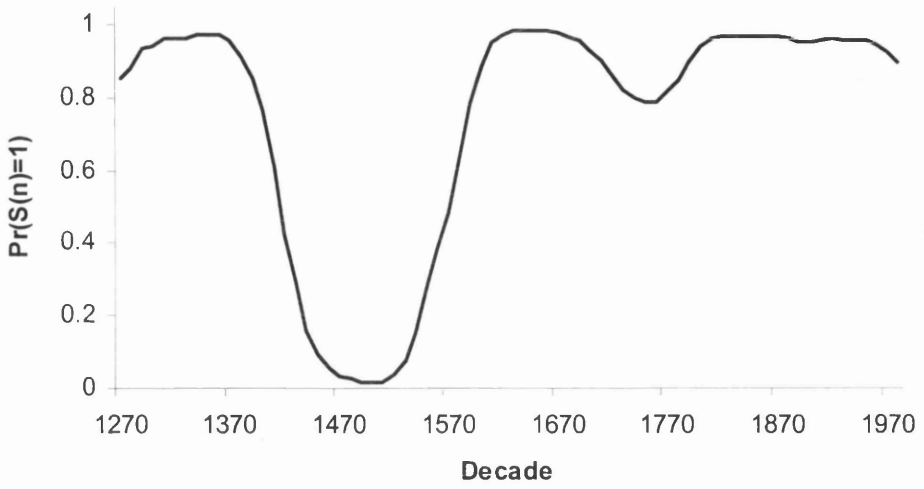

Figure 9.18: The inferred probability that the signal is in the upper regime ($S(n) = 1$).

Case (3 switch)	Log-Likelihood	$\hat{l}_0$	$\hat{l}_{12}$	$\hat{\sigma}$
MLE	88.0117	-0.0379	0.0337	0.0474
Sequence MLE	103.6385	-0.0338	0.0296	0.0559

Table 9.9: Some parameter estimates for the UK Wages data (with 3 switches with the series) found using by maximising log-likelihood and sequence log-likelihood.

9.4.2 Imputing a Regime Sequence

Interpreting the results of our application of the Ladder model we shall assume the time series can be modelled using 5 switches, 4 within the range of the time series itself and one shortly before it starts. We shall use the estimates obtained from fitting the Ladder model as a starting point for finding the sequence MLE (for 3 and 5 switches). Some of the parameter estimates given by the maximising the likelihood and the sequence likelihood, when trying to obtain the best sequence, are shown in Tables 9.9 and 9.10. A comparison of the best sequences found using MLE and sequence MLE are shown in Figures 9.19 and 9.20.

Case (5 switch)	Log-Likelihood	$\hat{l}_0$	$\hat{l}_{12}$	$\hat{\sigma}$
MLE	119.8076	0.0337	-0.0379	0.0474
Sequence MLE	125.8620	0.0377	-0.0389	0.0422

Table 9.10: Some parameter estimates for the UK Wages data (with 5 switches with the series) found using by maximising log-likelihood and sequence log-likelihood.

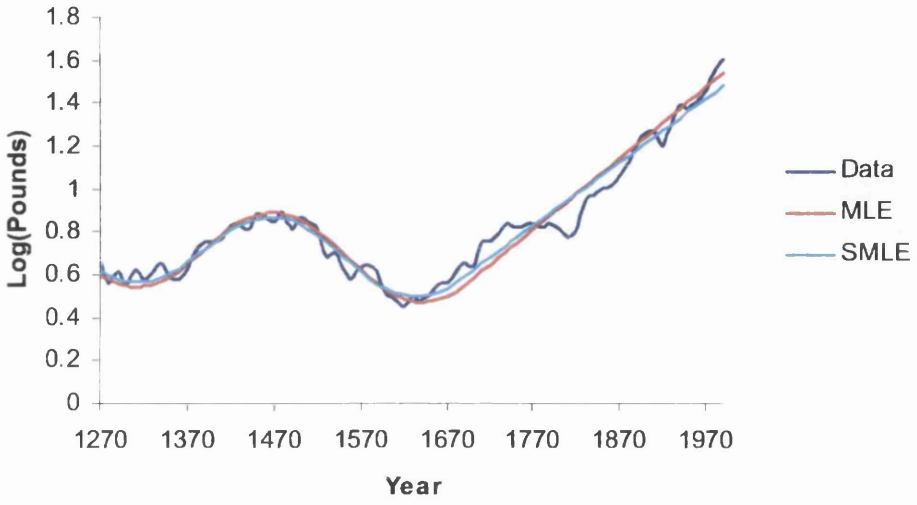

Figure 9.19: A comparison of the best sequence (using 3 switches) found by maximum likelihood (MLE) and sequence maximum likelihood (SMLE).

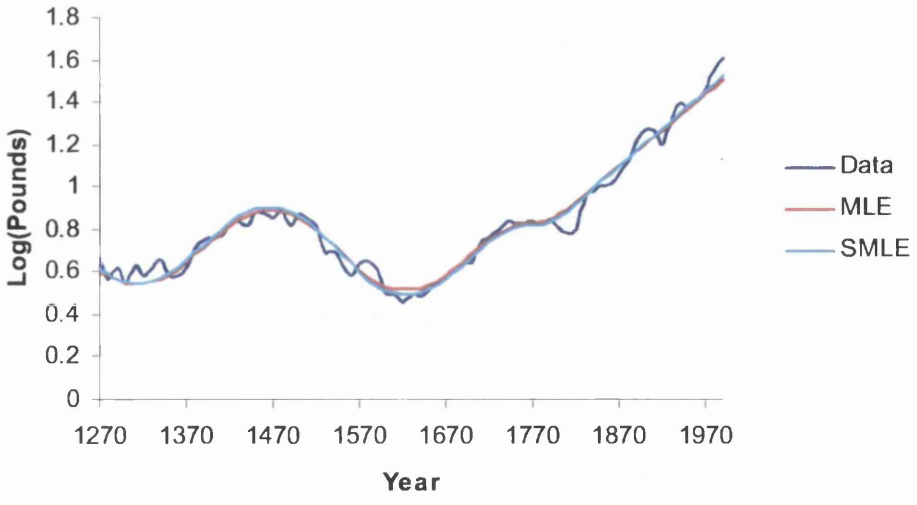

Figure 9.20: A comparison of the best sequence (using 5 switches) found by maximum likelihood (MLE) and sequence maximum likelihood (SMLE).

9.4.3 Conclusions

Each of the two cases gives a pretty reasonable fit with all the major (large scale) features included. Given that the number of switches is fairly low it is probably worth the extra effort to work with the 5 switch case. The maximum likelihood estimates of the parameters of the $L[13, 13]$ model can be found in Table 9.8.

Note: The model suggests growth periods of 1250 to 1400, 1560 to 1700 and 1760 to the present day. The periods of decay are 1400 to 1560 and 1700 to 1760.

9.5 Case 4: US Postwar GNP

The fourth case study uses a familiar times series. It is the growth rate of the difference of the log of US Gross National Product between the years 1952 and 1984. It is included for two reasons. First, it is the data set that Hamilton used when he introduced his Markov switching model and is a worthy candidate for reasons of continuity. Secondly, it will be interesting whether the inference shows up any evidence in support of gradual switching. The series was downloaded directly from Hamilton's website to ensure we were using the same figures.

9.5.1 Hamilton's Markov Switching Model

The original analysis of this data set, undertaken by Hamilton (1989), uncovered an apparent sequence of rising and falling regimes of GNP. His results can be expressed in two parts, as the parameter estimates he obtained and also the inferred distribution of the process between these two states. The clarity of this separation of the series into apparent regimes will have encouraged the development of this class of models and strengthened the assumption that there was an underlying Markov signal. Figure 9.21 shows us the probability of being in the falling regime. It is worthy of comment that the algorithm appears to be confident that the process is either in one regime or the other. The three lines correspond to the simple smoother at the time (using only current or previous observations), four quarters later and the full sample smoother. There is very little difference between the second and third versions. We should not read too much into this, as the clear regime pattern suggested can be misleading. The filter constructs the likelihood from taking each step and examining it separately from the others. As a result the sequence suggested by the inferred distributions does not necessarily represent anything like the single best solution. Instead it represents something more like a distribution drawn from all solutions, weighted by their likelihood. This becomes clear if we try to obtain the sequence maximum likelihood solution.

9.5.2 The Pointer Filter

It is clear from the inferred distribution that the pattern of the hidden driving signal consists of long periods of growth, interrupted by short periods of recession. Given the lengths of

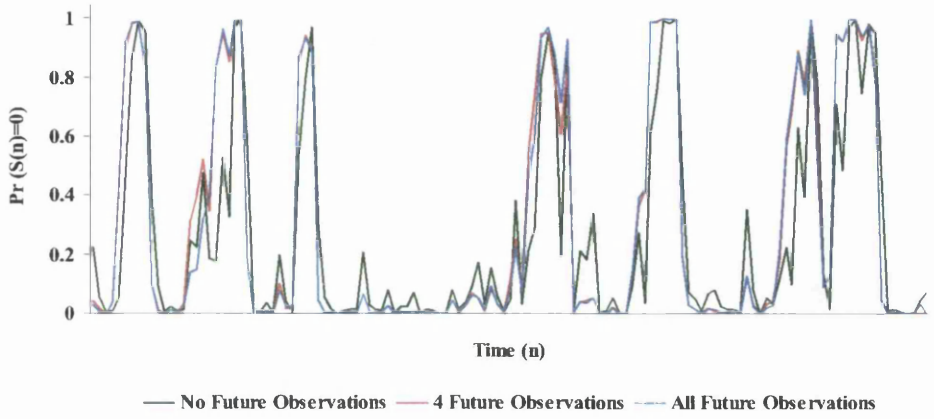

Figure 9.21: A comparison of the different inference obtained from the filter about the current regime ($Pr[S(n) = 0]$) at the time, 4 quarters later and using the whole sample.

time involved it is unlikely that more than one or two switches will occur during the recent history of the process (the last 4 quarters). In this instance we could consider using the Pointer Filter to fit the Hamilton Markov Switching model and compare its results with those of the Basic Filter. We apply the method detailed in Section 2.3, requiring the process to remember only the last two switchpoints. On top of the Markov growth pattern we add the the same AR(4) noise used in the original analysis.

The algorithm optimises robustly as did the original, leaving us with one clear maximum. This is given below, in Table 9.11, along with the original parameter estimates. These results are very close to the original values. Even more striking is the similarity in the inferred distribution between levels, shown in Figure 9.22. Again we find almost total agreement with the original findings. The saving in terms of performance is slight in this case and therefore in no way surplants the original algorithm in such a straightforward case. When dealing with more complex model dynamics, such as gradual switching, and well behaved data it can help prevent the order of the distribution matrices becoming unmanageable. After successfully applying this algorithm to fitting the same model as Hamilton the next step is to introduce gradual switching mechanics. We can do this quite simply by transforming the state history to include a gradual, rather than instant, switch. First we attempt to fit some versions of the model with asymmetric switching periods. We choose to apply a basic model with gradual downward switch, immediate upward switch and no autoregressive noise added. We can see in Table 9.12 the results of these initial experiments. On the plus side we find much the same kind of solutions as we found in the original Hamilton model. We do, however, find little to indicate the relevance of these models. The noise levels do not fall significanly as we extend the switching period and remain at the same levels as when we work with the simple two-state model.

From here it would have been a natural progression to have explored this a little further

Parameter	Hamilton	Vector
$\hat{\mathbf{a}}$	0.2450	0.2755
$\hat{\mathbf{b}}$	0.0951	0.1052
$\hat{\mathbf{l}}_0$	-0.3577	-0.4092
$\hat{\mathbf{l}}_1$	1.16	1.1565
$\hat{\sigma}$	0.7690	0.7618
$\hat{\phi}_1$	0.014	0.0208
$\hat{\phi}_2$	-0.058	-0.0741
$\hat{\phi}_3$	-0.247	-0.2436
$\hat{\phi}_4$	-0.213	-0.1805

Table 9.11: A comparison of the parameter estimates obtained by maximising likelihood for Hamiltons Markov Switching model using both Hamiltons filter and the Pointer Filter.

by adding autoregressive noise to the gradual switching process. When adding autoregressive noise to Hamilton's model we are faced with distribution matrices that expand rapidly as the order of the autoregression grows. This can discourages us from working with high order autoregressions, as the time required to obtain solutions also grows. This problem is considerably worse when considering gradual switching models. While a process with two-regimes of 1 level each and an autoregression of order r requires matrices of size 2^{r+1}, a two-regime, n level gradual switching process could require matrices of size $2^{r+1}n^{r+1}$. It is not unusual to work with gradual switching models with up to 10 levels. The result of this is that finding a solution is a much more time consuming process. The introduction of a lengthy autoregression also makes much a much higher dimensioned parameter space with more complex structure (and more local maxima). This compounds the problem by requiring much more thorough exploration of the parameter space. These are the problems faced when combining gradual switching models and autoregressive noise, and why we try instead to work with Pointer filter (which reduces the size of the required matrices) or the Line Fit model (which introduces the regime history as parameters to be estimated). Due to the limitiations of the Line Fit model, we use it only to refine an approximate solution we already have. As a result we turn first to the Pointer filter. The introduction of gradual switching behaviour was not itself problematic. It proved rather difficult to obtain the maximum likelihood values. The likelihood space seemed rather complex and the algorithm had a tendency to behave rather unpredictably. After much experimentation a decision was taken to abandon this approach. One of the major factors in making this decision was the time taken to obtain each solution. Instead of persisting with this line, the focus shifted to models that can be applied much faster.

9.5.3 The Ladder Model

Rather optimistically (given our previous experience) we shall begin by applying the Ladder model. While the evidence we have (Hamilton's parameter estimates) suggests the ratio

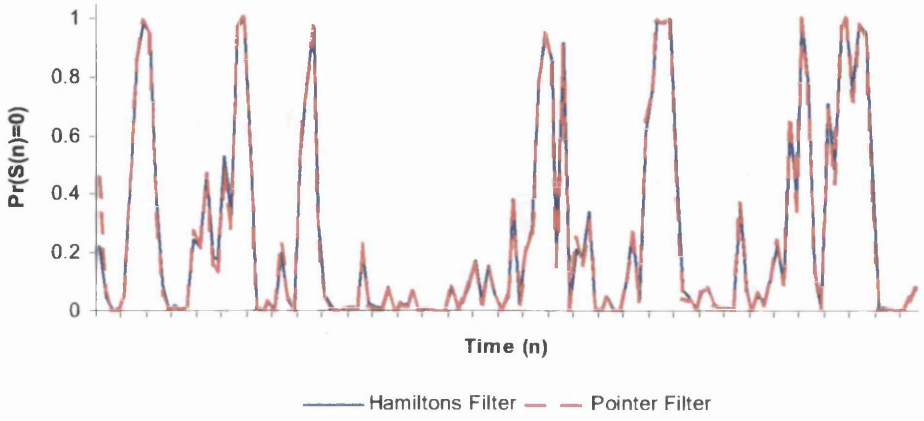

Figure 9.22: A comparison between the inferred probability of being in recession ($S(n) = 0$) using both Hamilton's filter and the Pointer filter.

Model	Log-Likelihood	$\hat{\mathbf{a}}$	$\hat{\mathbf{b}}$	$\hat{\mathbf{l}}_0^{(\mathbf{U})}$, $\hat{\mathbf{l}}_0^{(\mathbf{D})}$	$\hat{\mathbf{l}}_{N_{(U)}-1}^{(\mathbf{U})}$, $\hat{\mathbf{l}}_{N_{(D)}-1}^{(\mathbf{D})}$	$\hat{\sigma}$
$\mathbf{L[2,3]}$	-185.4000	0.2775	0.1025	-0.4077	1.1470	0.8108
$\mathbf{L[2,4]}$	-185.2377	0.3301	0.1032	-0.4507	1.1290	0.8166
$\mathbf{L[2,5]}$	-185.2146	0.3326	0.1021	-0.4536	1.1163	0.8186
$\mathbf{L[2,6]}$	-185.1858	0.3422	0.0959	-0.5038	1.0909	0.8209
$\mathbf{L[2,7]}$	-185.0708	0.3431	0.0959	-0.5081	1.0897	0.8208

Table 9.12: The maximised Log-likelihood and MLE for gradual switching Ladder models fitted using the Pointer Filter.

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
L[2, 2]	-184.9760	0.3202	0.0973	-0.4938	1.0914	0.8232
L[3, 3]	-183.1656	0.4292	0.1478	-0.6897	1.1672	0.8033
L[4, 4]	-183.3183	0.4437	0.1733	-0.8547	1.2137	0.8078
L[5, 5]	-185.1080	0.4338	0.1825	-0.9898	1.2293	0.8282

Table 9.13: The maximised Log-likelihoods and MLEs for US GNP using different Ladder models.

between noise levels and difference in levels is below 1, we should still expect difficulty in applying the higher order ladders. The results of these early experiments are given in Table 9.13. Here we face a familiar problem. Is the increase in likelihood due to better fitting of the regime switching behaviour or due to modelling of the fluctuations of the noise. The answer to this becomes clear as the order of the ladder increases. As the order increases the values of the switching parameters rises. For the $L[5, 5]$ model the mean sojourn time for the upper and lower levels are 5.48 and 2.31 respectively, not enough for the process to reliably complete the switch from one end of the scale to the other. It is clear the addition of the extra levels is not merely slowing down the rate of transition between the ends of the ladder. As the switching parameters approach 0.5 it allows the model to behave as a bounded random walk up and down a discrete scale with a finite number of levels. As with previous data if we want solutions reflecting regime switching behaviour then we must move beyond the Ladder model. The next, natural, step to take is to employ the Slide model.

9.5.4 The Slide Model

The usual approach we take when faced with this problem is to switch to using the Slide Model. This has the advantage that the switching process is unstoppable and however convincing the initial evidence of a switch, an inappropriate switch of regime will punish the likelihood severely. It does forfeit a property the Ladder model possesses, which is the ability to model a partial (incomplete) transition. This is the case where the Ladder switches regime before the level of the signal has reached one of the stable end-levels. The results of fitting the Slide model when the model is symmetric (the number of levels in each regime is the same) are given in Table 9.14. We have also included some cases where the level numbers are different for different regimes. These are the Asymmetric Slide models (see Table 9.15).

We would like to think that this reduction in likelihood heralded an improved model fit, but sadly this may not be the case. The likelihood score is constructed partly from the Markov switching probabilities and partly from the residuals. By adjusting the transition matrix to obtain a Slide model we introduce several deterministic movements. These carry a probability of 1, which subtracts nothing from the likelihood score. It is questionable whether the likelihood value obtained from models with two different orders are comparable. As a result the likelihood may be a good method for determining the optimum parameter

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
$S[2,2]$	-184.9760	0.3202	0.0973	-0.4934	1.0914	0.8233
$S[3,3]$	-182.6793	0.4218	0.0848	-0.7293	1.0982	0.8266
$S[4,4]$	-180.5120	0.7516	0.0801	-1.0175	1.1040	0.8266
$S[5,5]$	-180.9961	0.9999	0.0916	-0.8803	1.1562	0.8448

Table 9.14: The maximised log-likelihoods and MLEs for US GNP using different symmetric Slide models.

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
$S[2,2]$	-184.9760	0.3199	0.0974	-0.4928	1.0915	0.8232
$S[2,3]$	-183.6453	0.3345	0.0908	-0.5544	1.1048	0.8268
$S[2,4]$	-181.2549	0.4204	0.0839	-0.7712	1.1038	0.8068
$S[2,5]$	-180.1126	0.4952	0.0816	-0.8447	1.1094	0.8088
$S[2,6]$	-179.1744	0.5893	0.0847	-0.8673	1.1296	0.8100
$S[2,7]$	-178.0377	0.7288	0.0875	-0.9143	1.1537	0.8051
$S[2,8]$	-177.2544	0.9146	0.0924	-0.9154	1.1805	0.8052
$S[2,9]$	-177.1244	0.9999	0.1009	-0.8445	1.2187	0.8081
$S[3,4]$	-179.9285	0.5345	0.0809	-0.9309	1.1128	0.8038
$S[3,5]$	-178.5958	0.6680	0.0834	-0.9858	1.1322	0.8034
$S[3,6]$	-177.9186	0.7580	0.8940	-0.9382	1.1627	0.8077
$S[4,5]$	-179.2435	0.9999	0.0832	-1.0582	1.1288	0.8245

Table 9.15: The maximised log-likelihoods and MLEs for US GNP using different Asymmetric Slide models.

values for particular specification of the Slide model, but not so good for measuring one version against another. In this case a fairer judgement may be made by examining the standard deviation of the noise. The best results are obtained for an asymmetric switching process with a slower transition down that up. We find the best model with $S[3,5]$ levels up and down respectively although the $S[3,4]$, $S[2,7]$ and $S[2,8]$ cases display almost the same level of noise. The variance of the added noise represents only around 95% of that in the simple two-level model.

9.5.5 Evidence of Gradual Switching

The question remains as to whether this slight improvement in the fit of the models represents a significant enough gain to be heralded as evidence of gradual switching. In order to draw some conclusions from this we shall need to determine some way to perform a meaningful comparison of the two-regime model with and without gradual switching. One way this can be done is to construct a model that can display both types of behaviour. In order to do this we will have to take at face value the results of the inference and assume that the series really does display long, uninterrupted periods of growth. The model can be defined

as follows:

We have a process with $N_{(L)}$ levels, $N_{(U)}$ of them in the upper regime and $N_{(D)}$ of them in the lower regime. Apart from this the transition matrix of the process is constructed identically to the Slide model. The following example is for the case $N_{(U)} = 3, N_{(D)} = 5$.

We have combined states $\{d0, d1, u1, u2, u3, u4\}$. Note that combined states $d2$ and $u0$ are ignored, as before. The combined states have corresponding levels $\{l_0^{(D)}, l_1^{(D)}, l_1^{(U)}, l_2^{(U)}, l_3^{(U)}, l_4^{(U)}\}$. Note also that as we are dealing with an asymmetric level structure while $l_0^{(D)} = l_0^{(U)}$ and $l_{N(D)-1}^{(D)} = l_{N(U)-1}^{(U)}$, $l_k^{(D)}$ is not necessarily the same level as $l_k^{(U)}$ as it was in the symmetric case.

$$T = \begin{bmatrix} (1-a) & . & . & . & a & . \\ 1 & . & . & . & . & . \\ . & 1 & . & . & . & . \\ . & . & 1 & . & . & . \\ . & . & . & . & . & 1 \\ . & . & . & b & . & (1-b) \end{bmatrix}$$

The significant difference is in the level structure. We (slightly) generalise by adding another parameter.

$$\begin{aligned} l_k^{(D)} &= l_0^{(D)} + \gamma \left(\frac{k}{N_{(D)} - 1} \right) . (l_{N(D)-1}^{(D)} - l_0^{(D)}) \\ l_k^{(U)} &= l_{N(U)-1}^{(U)} + \gamma \left(\frac{N_{(U)} - 1 - k}{N_{(U)} - 1} \right) . (l_0^{(U)} - l_{N(U)-1}^{(U)}) \end{aligned}$$

When $\gamma = 1$ we have an asymmetric Slide model $S[N_{(U)}, N_{(D)}]$.

When $\gamma = 0$ we have a two-regime, two-level model, although it will not be permitted to switch as freely as the $S[2, 2]$ model.

We shall denote this model $S[\mathbf{N}_{(U)}, \mathbf{N}_{(D)}, \gamma]$. The reliability of the conclusions we draw from this will be very dependent upon the capacity of this construction to recreate the results of the original model.

The results of applying this model can be found in two separate tables. When we set $\gamma = 1$ we obtain the same results as for the simple Slide model, shown in Table 9.15. The results for the case $\gamma = 0$ are found in Table 9.16. In the first we mimic a Slide model and in the second we mimic the two-regime model but with modified level structure so as to make the Markov transition probabilities comparable with the Slide model. We can see, for $\gamma = 0$, that despite the algorithm gaining some likelihood as a result of the deterministic movement, this is balanced by the deteriorating fit due to the greater restrictions. For $\gamma = 1$ we see falling levels of noise rather than rising ones as the number of levels increases

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
S[2, 2, 0]	-184.9760	0.3198	0.0974	-0.4924	1.0918	0.8231
S[2, 3, 0]	-184.9206	0.2970	0.0979	-0.4879	1.0831	0.8305
S[2, 4, 0]	-184.7260	0.3122	0.1036	-0.5367	1.0634	0.8346
S[2, 5, 0]	-184.7867	0.3341	0.1019	-0.6416	1.0326	0.8383
S[2, 6, 0]	-185.1177	0.3570	0.0995	-0.7663	1.0054	0.8391
S[2, 7, 0]	-185.3373	0.1639	0.1254	-0.1658	1.1401	0.8696
S[3, 5, 0]	-183.7392	0.4374	0.0845	-0.6924	1.0222	0.8375

Table 9.16: The maximised Log-likelihoods and MLEs for the modified Slide model (mimicking the two-regime Slide).

Model	Log-Likelihood ($\gamma = 0$)	Log-Likelihood ($\gamma = 1$)	Likelihood Ratio
S[2, 4]	-184.7260	-181.2549	3.4711
S[2, 6]	-185.3373	-178.0377	7.2996
S[3, 5]	-183.7392	-178.5958	5.1434

Table 9.17: A comparison of the maximised Log-Likelihood for the two versions of the modified Slide model.

and this is reflected in the higher likelihood values. To allow some comparison to be made between the two cases we look at the difference in likelihoods (see Table 9.17). These models $S[\mathbf{N}_{(U)}, \mathbf{N}_{(D)}, 0]$ and $S[\mathbf{N}_{(U)}, \mathbf{N}_{(D)}, 1]$.are not nested and so we must be cautious in our interpretation of the likelihood ratio statistic. What conclusion we are able to draw from this is open to debate but it is useful as an indication that there does appear to be some evidence of gradual switching. In order to make any robust conclusion it will be necessary to respecify the model so the two cases are nested.

We shall continue with broadly the same direction of attack but should consider whether we have treated the two-state case too harshly. To keep the model simple, when $\gamma = 0$ the transition does not occur between the two-regimes until the stable end level in the regime is reached. It may be fairer to allow this transition to occur at any point within the level structure of a regime. To do this we must replace γ with two parameters.

To keep the model reasonably practical we limit ourselves to the asymmetric case $S[2, N_{(D)}]$ with only two levels in the upper regime. The two new parameters we introduce control the position of the switch within the level structure and the degree to which this switch is instantaneous. We can now specify the level structure more formally:

$0 \leqslant g \leqslant N_{(D)}$ is the threshold parameter, controlling the position of the switch

$0 \leqslant h \leqslant 1$ is the contrast parameter, controlling the suddeness of the switch

We have states $\left\{d_0, d_1, ..., d_{N_{(D)}-1}, u_1\right\}$ with associated values $\left\{l_0^{(D)}, l_1^{(D)}, ..., l_1^{(U)}\right\}$

For each value in the lower regime d_k we transform in one of two ways.

$$If \quad k < g \quad then \quad l_k^{(D)} = l_0^{(D)} + \left(\frac{k}{N_{(D)}-1}\right)^{\left(\frac{1}{h}\right)} (l_{N_{(D)}-1}^{(D)} - l_0^{(D)})$$

$$If \quad k \geqslant g \quad then \quad l_k^{(D)} = l_0^{(D)} + \left[1 - \left(\frac{N_{(D)}-1-k}{N_{(D)}-1}\right)^{\left(\frac{1}{h}\right)}\right] (l_{N_{(D)}-1}^{(D)} - l_0^{(D)})$$

We denote this model by $\mathbf{S}[2, N_{(D)}, g, h]$. So for the case where $h = 1$ we have the Asymmetric Slide model while for $g = N_{(D)}$ and $h = 0$ we have the same two-regime, two-level model ($\mathbf{S}[2, N_{(D)}, 0]$) we have just been working with. One advantage we now have is that we are able to clearly see that the models $\mathbf{S}[2, N_{(D)}, g, 0]$ (which represent approximations of the two-regime, two-level model) are a special case of the general model. This gives us the opportunity to apply the Likelihood Ratio test to the results.

Running the optimisation routine for these two cases ($\mathbf{S}[2, N(D), g, 1]$ and $\mathbf{S}[2, N_{(D)}, N_{(D)}, 0]$) gives the same results as before. Now we can make the small adjustment of introducing these two parameters as variables and optimise again. The likelihood for this test model is not always easy to maximise, but the relative simplicity of the routine and model structure allows repeated attempts with different starting positions. Robust maxima are found and they always appear to have a contrast value close to 1 (see Table 9.18). This value of h corresponds to the unmodified Slide model. Given that $h = 1$ gives the higher likelihood we can take the unmodified Slide model to represent all versions of the Slide model. Now we need to find the best value of g when $h = 0$. These are tested in Table 9.19 and optimum values found for the transition between regimes in the level structure. We can now take the best values in Table 9.19 to represent the maximum likelihood for the model with a contrast value of $h = 0$. A comparison should then be made against model with $h = 1$, as these represent the Slide model The results of this are given in Table 9.20.

The test statistics we obtain for the Likelihood Ratio test have significance values of 0.045 and 0.004 respectively. These are significant at the 5% level and allow us to conclude that gradual switching is present, given certain conditions. These conditions concern the quality of the approximation of the two-regime, two-level model to the original. We can see the MLE's in most cases remain pretty close to the original values, although this cannot be said of the $S[2, 7]$ case. To test this further we can examine the inferred distribution of regimes produced by all three cases in Table 9.20. The results are displayed in Figure 9.23.

We would be hard pressed to seperate the four graphs on display and given the acceptable similarity betweeen the MLEs there is no reason to reject any of the models as a poor representation of the two-level model, despite the additional restrictions they impose.

Model	Log-Like'd	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}, \hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}, \hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$	$\hat{g}$	$\hat{h}$
S[2, 4, g, h]	-181.2547	0.4217	0.0836	-0.7752	1.1027	0.8068	3	0.9999
S[2, 5, g, h]	-180.1125	0.4972	0.0814	-0.8502	1.1088	0.8088	1	0.9999
S[2, 6, g, h]	-179.1635	0.5899	0.0848	-0.8594	1.1278	0.8098	2	0.9592
S[2, 7, g, h]	-178.0374	0.7343	0.0874	-0.9142	1.1527	0.8049	2	0.9981

Table 9.18: The maximised Log-likelihoods and MLEs for the modified Slide model (with contrast and offset parameters).

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}, \hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}, \hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
S[2, 4, 2, 0]	-183.9989	0.3756	0.0803	-0.5453	1.0621	0.8339
S[2, 4, 3, 0]	-183.2638	0.4734	0.0772	-0.4573	1.0894	0.8327
S[2, 7, 2, 0]	-184.6420	0.3932	0.0948	-0.5971	1.0275	0.8504
S[2, 7, 3, 0]	-182.5161	0.5529	0.0909	-0.5630	1.0565	0.8347
S[2, 7, 4, 0]	-182.1011	0.6633	0.0837	-0.4440	1.0825	0.8405
S[2, 7, 5, 0]	-182.2061	0.8242	0.0801	-0.3238	1.1087	0.8501
S[2, 7, 6, 0]	-183.3437	0.7163	0.0880	-0.1465	1.1872	0.8506
S[3, 5, 2, 0]	-182.6593	0.5095	0.0857	-0.5154	1.0682	0.8356
S[3, 5, 3, 0]	-182.4385	0.6263	0.0770	-0.4258	1.0866	0.8417
S[3, 5, 4, 0]	-182.6934	0.9999	0.0693	-0.3600	1.0854	0.8566

Table 9.19: The maximised Log-likelihoods and MLEs for the modified Slide model (with contrast and offset parameters).

Model	Log-Likelih'd ($h = 0$)	Log-Likelih'd ($h = 1$)	Test Statistic	d.o.f.	Significance (p=)
S[2, 4, 3, h]	-183.2638	-181.2549	4.0178	1	0.0450
S[2, 7, 4, h]	-182.1011	-178.0377	8.1268	1	0.0044

Table 9.20: A comparison of the maximised Log-Likelihood of the two versions of the modified Slide model (offset parameter is optimised). The test statistic of the Likelihood Ratio test is given.

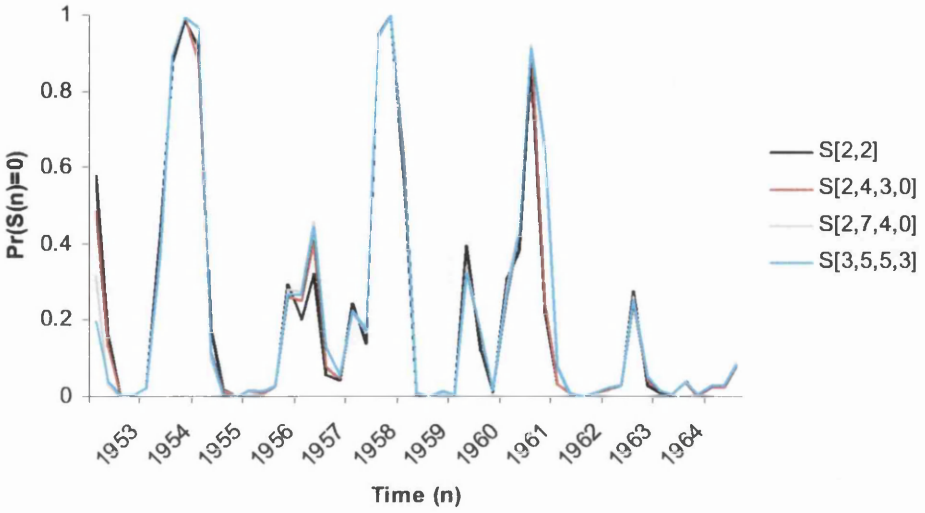

Figure 9.23: The inferred probabilites of the US GNP being in recession ($S(n) = 0$) for different versions of the modified Slide model. Only a short section of the series is shown (1952 to 1964).

9.5.6 Unravelled Models

Although we have managed to collect evidence for the existence of gradual switching in the original dataset, we have by no means conclusively proved it. From here we can progress in two ways. First, we can continue applying these models and consider the results. Secondly, we can continue to seek out more evidence for their suitability. In practice we can do these two things simultaneously by moving onto one of the other variations of the general class of models, namely the Unravelled models detailed in Section 8.6.

The first step in the process of fitting Unravelled models is to determine the probable number of switches. The likelihood will not optimise well if we allow this to vary during the process so we perform the optimisation separately for each possible number of switches. We would usually estimate the number of switches by eye and then vary around this (as far as time permits) as less obvious switching behaviour could easily be overlooked. In practice we would not want a model that presumed switching every few points, as this would be deviating from the original intention of working with this kind of model. The converse, of a model that demonstrated many fewer switches than anticipated might still be of great interest in identifying a less systematic (or reversible) regime change.

On examining the distribution (see Figure 9.24) between regimes for the $L[2,2]$ model, based on full sample inference, we find several clear periods where there is a high probability the process is in the lower regime. It is possible, indeed likely, that each of the most probable paths that contribute to the bulk of the likelihood do not exploit every one of these regions. For clarity we have labelled these regions A through G. The inference appears to show

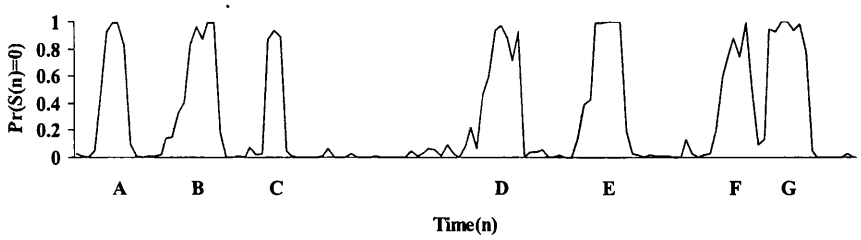

Figure 9.24: The inferred probability (using the Full Sample smoother) of the signal being in the recession state ($S(n) = 0$). Letters A to G represent the 7 likely sojourns in this regime suggested by the data.

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_1$	$\hat{\sigma}$
UL[2, 2, 10]	-187.7633	0.3209	0.0433	-0.9648	0.9473	0.8633
UL[2, 2, 11]	-190.7098	0.2667	0.0535	-0.7054	0.9728	0.8763
UL[2, 2, 12]	-187.0398	0.3145	0.0536	-0.8190	0.9822	0.8513
UL[2, 2, 13]	-189.5046	0.2542	0.0651	-0.5790	1.0198	0.8595
UL[2, 2, 14]	-186.6543	0.3072	0.0647	-0.6949	1.0181	0.8406
UL[2, 2, 15]	-188.6710	0.2417	0.0784	-0.4631	1.0706	0.8447
UL[2, 2, 16]	**-186.5183**	**0.3101**	**0.0760**	**-0.6133**	**1.0467**	**0.8322**
UL[2, 2, 17]	-188.2137	0.2483	0.0911	-0.4071	1.1026	0.8350
UL[2, 2, 18]	-186.5654	0.3202	0.0875	-0.5580	1.0690	0.8260
UL[2, 2, 19]	-188.0146	0.2614	0.1035	-0.3735	1.1251	0.8279
UL[2, 2, 20]	-186.7340	0.3322	0.0991	-0.5148	1.0882	0.8207

Table 9.21: The maximised Log-Likelihood for the Unravelled Ladder model UL[2,2,-] for different numbers of switches

quite firmly that the process begins and ends in the same (upper) state and so we would expect an even number of switches. If a realisation of the process utilised every one of these regions then we would expect 14 switches. So we apply the Unravelled Ladder $UL[2, 2, c]$ model with between 10 and 20 switches. The results of this are displayed in Table 9.21.

As expected there is a clear advantage to the versions of the model that require an even number of switches. We also find that the maximum likelihood value occurs for 16 switches, rather than 14. If we consider the other measurement of fit we find the estimated noise level to continue to fall for higher numbers of switches. This demonstrates the difficulty of working with these models. The algorithm is very eager to add the possibility of extra switches in order to mop up some of the residual noise, regardless of whether any of the likely regime sequences we could construct would implement them.

Unravelled Slide Model

We shall first consider what happens when we use the Unravelled Asymmetric Slide model. We assume the number of switches is in the range of 12 to 18, that the number of levels

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$	Min(σ)
US[2, 2, 16]	-186.518	0.3101	0.0760	-0.6133	1.0467	0.8322	No
US[2, 3, 14]	-185.065	0.3323	0.0680	-0.7142	1.0509	0.8339	No
US[2, 4, 14]	-182.358	0.4389	0.0693	-0.9173	1.0676	0.8079	No
US[2, 5, 14]	-181.068	0.5340	0.0722	-0.9512	1.0872	0.8084	Yes
US[2, 6, 14]	-180.139	0.6266	0.0762	-0.9336	1.1131	0.8108	Yes
US[2, 7, 14]	-178.879	0.7788	0.0805	-0.9506	1.1431	0.8057	Yes
US[2, 8, 14]	-178.051	0.9584	0.0856	-0.9274	1.1725	0.8053	Yes
US[2, 9, 14]	-177.874	0.9999	0.0933	-0.8369	1.2127	0.8099	Yes
US[2, 10, 14]	-179.417	0.9999	0.1029	-0.7290	1.2481	0.8201	Yes
US[3, 4, 14]	-180.917	0.5511	0.0719	-1.0144	1.0931	0.8038	Yes
US[3, 5, 14]	-179.646	0.6966	0.0753	-1.0436	1.1164	0.8046	Yes

Table 9.22: The maximised Log-Likelihood for different versions of the Unravelled Slide model (The number of switches is fixed at 14).

Number of	Model				
Switches(c)	US[2, 2, c]	US[2, 3, c]	US[2, 4, c]	US[2, 5, c]	US[3, 4, c]
12	-187.0398	-185.579	-183.178	-181.787	-181.715
14	-186.6543	-185.065	-182.358	-181.068	-180.917
16	-186.5183	-185.124	-182.775	-181.944	-181.755
18	-186.5654	-185.458	-183.479	-183.232	-183.009
20	-186.7340	-185.937	-184.311	-184.574	-184.255

Table 9.23: The optimum switch number found by a comparison of the maximised Log-Likelihood for different versions of the Unravelled Slide model

in the down regime is between 1 and 9, and the number of levels in the upper regime is between 1 and 3.

We first choose a triple of these constant parameters $(N_{(U)}, N_{(D)}, c)$ and apply the model $US[N_{(U)}, N_{(D)}, c]$, and then by measuring the log-likelihood, optimise with respect to the parameters (a, b, l_0, l_1, σ). Although we use likelihood to measure the suitability of a particular model, specified by its constant parameters, it is misleading to attempt to make direct comparisons between models with different parameters using likelihood. The deterministic structure of parts of the transition matrix encourage longer and longer transition times. A summary of the results is given in Table 9.22.

On comparing these models we find two things. Initially we notice that the findings of the simple unconstrained Slide model are borne out, with the lowest level of noise reported by the $US[2, 8, 14]$ and $US[3, 4, 14]$ models, in the $N_{(U)} = 2$ and $N_{(U)} = 3$ categories respectively. We are also pleased to find clear and stable maxima in the likelihood function. We can see that this seems largely independent of the model chosen by comparing the performance of each model across different switch numbers in Table 9.23.

Presenting this in graph form makes the results easier to comprehend (See Figure 9.25).

Each line represents a different model, as defined by the ladder lengths u and v, applied

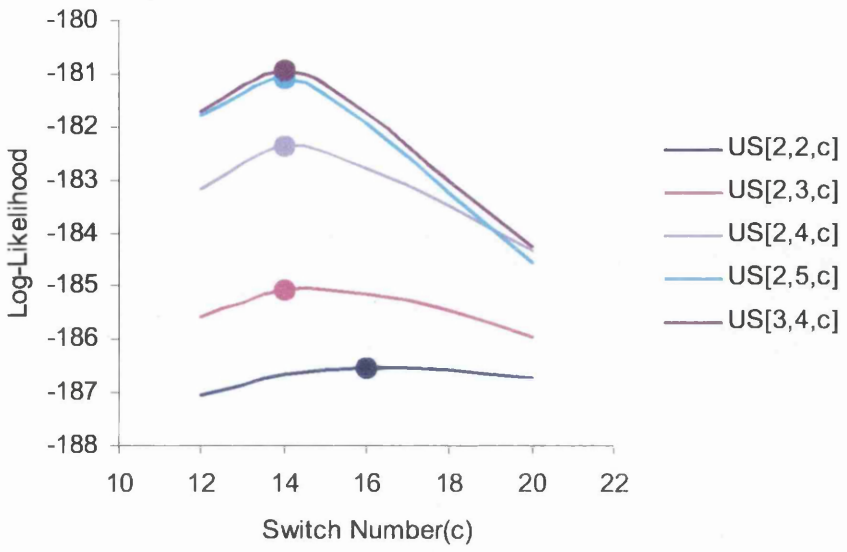

Figure 9.25: The optimum number of switches found by plotting the maximum likelihood possible for the different versions of the Unravelled Slide model.

across a range of values for the number of switches. The maximum point of each graph is indicated by a marker. It is by observing this graph that the second interesting observation becomes apparent. All of the gradual switching models appear to favour a 14 switch structure, while the $US[2,2,c]$ prefers 16 switches. Some light can be shed on this by focussing on the same models but measuring them by the level of noise they infer. If we take the standard deviation of the noise rather than the likelihood as measure of fit then we obtain Table 9.24. Again, when we present this in graphical form, in Figure 9.26, it is clearer. Where a minimum exists we have indicated it with a marker.

The simple $US[2,2,c]$ case has no minimum, nor does the $US[2,3,c]$. These models are able to switch states very quickly, allowing them to soak up some of the noise. For lower noise levels this is not a problem but when it is not always possible to determine the regime by knowledge of the value erroneous sojourns may be created. The models displaying longer switching times have no such problem. The cost of a mistaken switch for these models is much higher. We find that the greater the number of levels, the more clearly defined the optimum point is, in terms of noise levels.

Unravelled Ladder Model

The next natural step after considering the simplest case, the Slide model, is to turn our attention to the Ladder model. Surely this can only improve matters by adding greater flexibility to the switching behaviour. The truth is somewhat less inspiring. We shall first examine a summary of the results we obtained.

Number of Switches(c)	Model US[2, 2, c]	US[2, 3, c]	US[2, 4, c]	US[2, 5, c]	US[3, 4, c]
12	0.8513	0.8501	0.8319	0.8315	0.8271
14	0.8406	0.8339	0.8079	0.8084	0.8038
16	0.8322	0.8291	0.8067	0.8084	0.8053
18	0.8260	0.8269	0.8064	0.8151	0.8136
20	0.8207	0.8257	0.8071	0.8228	0.8215

Table 9.24: A comparison of the standard deviation of residual noise for different versions of the Unravelled Slide model using different numbers of switches.

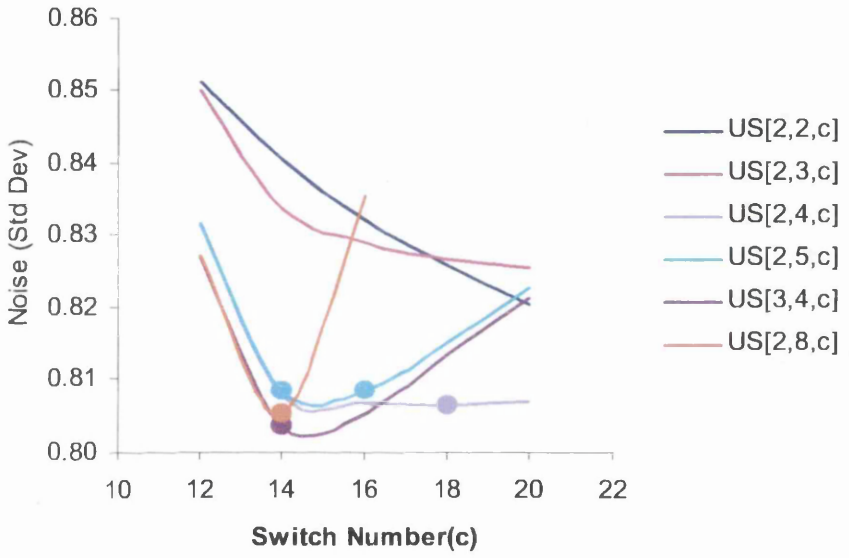

Figure 9.26: The optimum number of switches found by plotting the standard deviation of the residual noise for the different versions of the Unravelled Slide model.

Number of Switches	Model UL[2, 2, c]	UL[2, 3, c]	UL[2, 4, c]	UL[2, 5, c]	UL[3, 4, c]
12	-187.0398	-187.0455	-186.5043	-186.7402	-185.5300
14	-186.6543	-186.4133	-185.6937	-185.8406	-184.5687
16	-186.5183	-186.0593	-185.1705	-185.2403	-184.0279
18	-186.5654	-185.9098	-184.8578	-184.8614	-183.7384
20	-186.7340	-185.8973	-184.6861	-184.6309	-183.5950
22		-185.9768	-184.6095	-184.5008	-183.5384
24		-186.1202	-184.5991	-184.4406	-183.5331
20			-184.6364	-184.4308	-183.5566

Table 9.25: The optimum switch number found by a comparison of the maximised Log-Likelihood for different versions of the Unravelled Ladder model

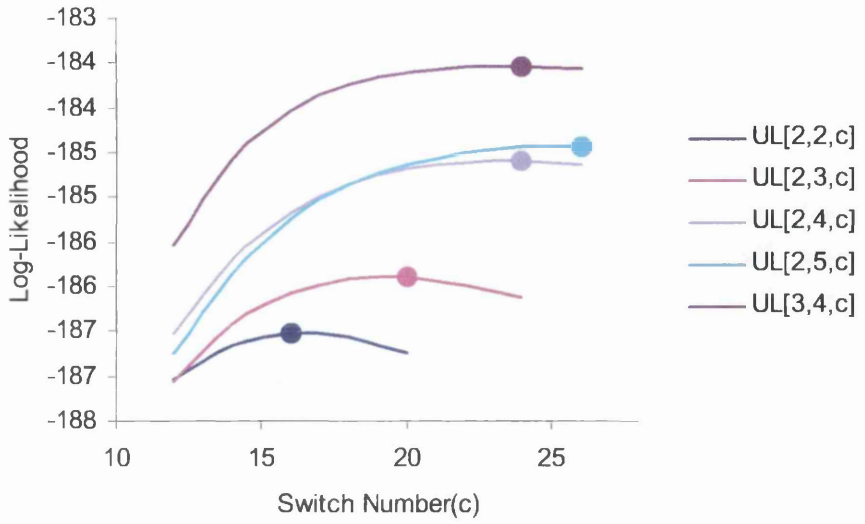

Figure 9.27: The optimum switch number found by plotting the maximised Log-Likelihood for different versions of the Unravelled Ladder model

It is worth reiterating at this point that, while the Asymmetric Slide model is simple to construct the same cannot be said of the Asymmetric Ladder Model. Unlike the symmetric case, where there is no problem in determining the start and end points of a regime transition, when a regime change takes place, multiple solutions exist The particular approach we have chosen was dictated by programming ease rather than any other criterion. As such the results we have obtained when the upper regime consists of more than one level, while valid, cannot be said to be authoritative. We have chosen some of the better performing versions of the Unravelled Ladder model and maximised the log-likelihood for different numbers of switches. The results of this are given in Table 9.25, and then plotted in Figure 9.27.

A quite different picture emerges from the Unravelled Slide model. For the simple model, discouraged from switching by the large gap between levels, there is a clear maximum for 16 switches. As the ladder lengths increase we find the optimum number of switches rising and the consequence of each partial switch becoming less extreme. Despite its flexibility it is this tendency of the model that makes it poor at distinguishing the appropriate number of switches to employ.

This tendency is clear from Table 9.26 as well. The greater the number of levels the smaller the features the model will try to apply itself to. It quickly becomes lost in modelling the noise and loses sight of the regime behaviour, so apparent to the simpler models.

Figure 9.28 shows evidence of this inexorable drive to overfit. As we add more levels the noise levels fall and the likelihood rises. Eventually the switching intensities will become so large as to betray any concept of regime-determined behaviour.

Number of Switches	Model UL[2, 2, c]	UL[2, 3, c]	UL[2, 4, c]	UL[2, 5, c]	UL[3, 4, c]
12	0.8513	0.8591	0.8471	0.8488	0.8399
14	0.8406	0.8462	0.8325	0.8334	0.8218
16	0.8322	0.8378	0.8231	0.8240	0.8117
18	0.8260	0.8323	0.8168	0.8181	0.8043
20	0.8207	0.8284	0.8120	0.8139	0.7979
22		0.8252	0.8079	0.8105	0.7917
24		0.8225	0.8043	0.8077	0.7854
20			0.8010	0.8051	0.7789

Table 9.26: A comparison of the standard deviation of residual noise for different versions of the Uravelled Ladder model using different numbers of switches.

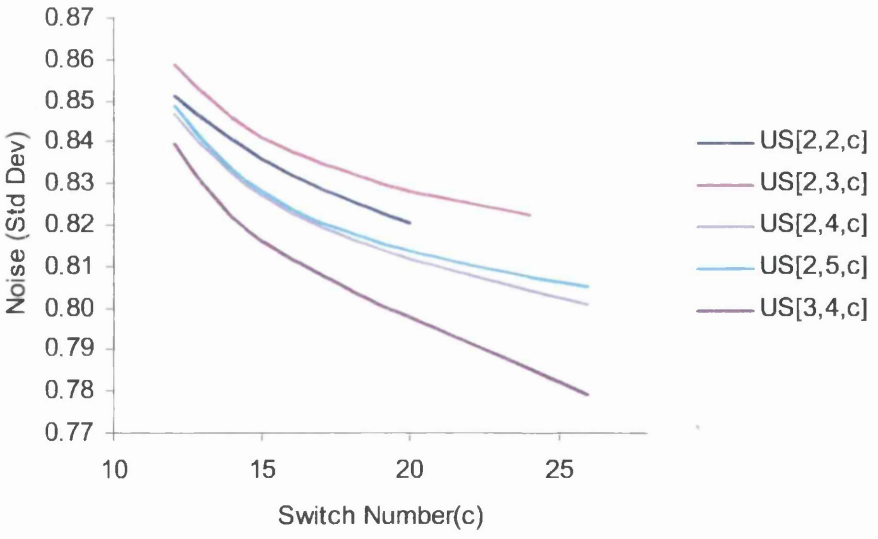

Figure 9.28: Comparing the residual noise for different versions of the Unravelled Ladder model utilising different numbers of switches.

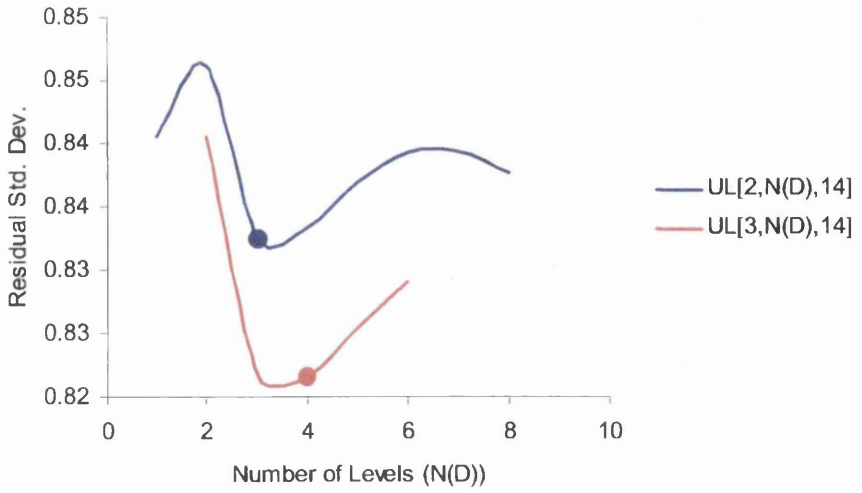

Figure 9.29: A plot of the standard deviation of the residual noise for Unravelled Ladders, with varying numbers of levels in the lower regime.

In order to progress further we choose to accept the weakness of the Ladder model to determine regimes and constrain ourselves to the inference of the Slide model regarding switch numbers. Then we can at least explore whether the switching structure, of which we are already convinced by the Slide model is better modelled with a Ladder. If we plot the noise level for many of the Ladder models we have observed when restricted to 14 switches, then we obtain Figure 9.29. We find a local minimum in the region favoured by the Slide model. There is no such confidence of a unique solution as there was in that case.

9.5.7 Imputing the Best Sequence

So far we have been able to demonstrate evidence of a gradual transition during switches. Without really changing the growth regime structure we can account for around 8% of the variance of the noise. For comparison the autoregressive element of the Hamilton model reduces the noise level by around 13%, although it is worth remembering that this is a fourth-order autoregression. It would be interesting to know exactly how this autoregression was operating and in what way it reduces the noise level.

To do this we must look more closely at the best sequence for $S(n)$. We do this by applying the Line Fit model that we introduced on page 153. It is really a form of parameter augmentation in which we introduce the regimes occupied during the time series as parameters to be estimated. Due to the problems in optimising a likelihood function with so many parameters we only apply this model when we have a very good idea of what the 'best sequence' of regimes should look like. We also calculate the suitability of each regime sequence using a slightly different likelihood function we term the Sequence Likelihood.

Label	Start Date	End Date
A	1952:II	1953:II
B	1956:III	1957:I
C	1959:III	1959:IV
D	1968:III	1969:I
E	1973:II	1974:I
F	1979:I	1979:II
G	1980:III	1981:III

Table 9.27: The start and end positions for the sojourns in the down (or recession) state as predicted by Hamilton's two-state Markov switching model.

This is to avoid some of the problems that have arisen using the filters, problems to be discussed in Chapter 10.

In Table 9.27 we separate the peaks and troughs of the inference for the two-level model (identified in Figure 9.24) into apparent sojourns in the falling state. We labelled these using the letters A-F with the approximate positions, determined by output of the two-state Ladder model

We are careful to assume that any best sequence for $\{s(n)\}$ would incorporate all of these switches given that there may be costs incurred in terms of likelihood. We shall consider each sequence as containing of some, or all, of these sojourns and attempt to find a best sequence within this restriction. To do this we turn to working with the Line Fit model and find the sequence MLE. This model is defined in Section 8.8. Optimising this model is considerably more problematic with the Ladder models. The likelihood space is richly littered with local maxima and some of the parameters (namely the switch points) are discrete variables. As a result the optimisation is exceedingly tricky and only a very large number of attempts and careful exploration of the possibilities ensures that a maximum has been found. Fortunately we have good reason to have trust in the existence of a regime structure with only a few sojourns and can work only with solutions consisting of these.

We were quickly able to determine that those solutions utilising fewer than 4 switches fall way behind in terms of both likelihood and noise levels. Of those solutions that were credible we obtained maximised sequence log-likelihood values ($\ln(MLE_\Sigma)$). These are shown in Table 9.28.

The two cases that stand out are ABEFG, with log-likelihood -191.2676, and ABCDEFG, with variance 0.7996. Both are relevant, for different reasons and they are shown together in Graph 9.30.

It is worth commenting that these sojourns are shorter than those indicated by the Ladder and Slide models and also shorter than those indicated by the same models when autoregressive noise is added. We then take some of the better performing models (and regime sequences) and re-evaluate them using a 4th order autoregression (with coefficients ϕ_i) to model the residual noise. This should give us something like the best of the set of regime sequences that Hamilton's filter was indicating. The parameter estimates and

Sojourns	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_1$	$\hat{\sigma}$
ABEG	-193.8579	0.3127	0.0331	-1.0987	0.9202	0.8742
ABCEG	-196.0615	0.3619	0.0419	-1.0817	0.9354	0.8603
ABDEG	-196.1789	0.3382	0.0422	-1.0199	0.9448	0.8585
ABEFG	-191.2676	0.4210	0.0413	-1.3476	0.9283	0.8351
ABCDEG	-197.9204	0.3789	0.0512	-1.0103	0.9604	0.8439
ABCEFG	-193.1984	0.4044	0.0508	-1.1690	0.9641	0.8168
ABDEFG	-193.4175	0.3796	0.0512	-1.1057	0.9738	0.8154
ABCDEFG	-194.5511	0.4161	0.0604	-1.0912	0.9899	0.7996

Table 9.28: A comparison of the maximised sequence likelihood for the Line Fit model LF[2,2] using different sojourn patterns.

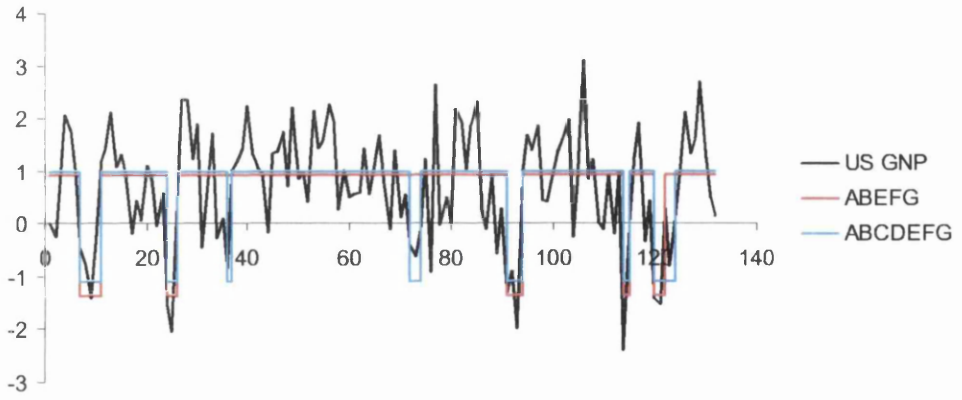

Figure 9.30: The growth rate of the US GNP series combined with two of the best sequences of $S(n)$ proposed by the $LF[2,2]$ model for two different sojourn patterns.

Sojourns	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0$	$\hat{l}_1$	$\hat{\sigma}$
ABEG	-191.0834	0.3688	0.0327	-1.184	0.8929	0.8609
ABEFG	-188.5315	0.4209	0.0413	-1.2958	0.922	0.8179
ABCDEG	-190.1618	0.1877	0.0589	-0.4397	1.1155	0.7701
ABCEFG	-190.5852	0.4655	0.0501	-1.2699	0.9372	0.807
ABDEFG	-188.8483	0.2909	0.0532	-0.8937	1.0306	0.777
ABCDEFG	-187.4716	0.2178	0.0691	-0.5416	1.1511	0.7315

Sojourns	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
ABEG	0.1723	0.0828	-0.0336	-0.0081
ABEFG	0.1718	0.0698	0.0108	-0.065
ABCDEG	0.0781	0.0027	-0.2223	-0.2424
ABCEFG	0.1627	0.0457	0.0646	-0.045
ABDEFG	0.1907	0.0445	-0.0841	-0.2237
ABCDEFG	-0.0361	-0.1074	-0.2859	-0.2928

Table 9.29: A comparison of the maximum sequence likelihood estimates for the Line Fit model LF[2,2] using different sojourn patterns (when the residuals are modelled using an AR(4))

maximised sequence likelihood are given in Table 9.29.

We see an interesting pattern emerging from this table. For the full set of 7 switches the autoregressive pattern Hamilton has led us to expect is clearly visible. We find the first two coefficients small compared with the third and fourth. When we remove certain sojourns this pattern disappears quite quickly. In fact only those lines of fit including the sojourn D seem to display high values for the last coefficient. But this is not the only noticeable feature. The line (ABCDEFG) achieving the highest likelihood value when AR noise is included is not one of the best performing without that noise. With careful examination some local maxima appear in the positions of the AR-free solutions. This is hardly surprising, given the number of local maxima in this type of likelihood space, but it is likely that the autoregression is doing more than just explaining the noise. By enabling the signal to be transformed without any cost (in terms of likelihood) it is allowing some of the more improbable models (without AR) to become competitive. A good example of this is the full 7-switch regime structure, ABCDEFG. This has a log-likelihood value of -194.5. This is approximately 4% as probable as the best fitting model which has a log-likelihood of around -191.3.

The question now is how competitive the gradual switching models are when applied using a Line Fit model. There are many different possibilities when choosing a model and each model, and regime structure, must be systematically explored to ensure we have not missed the overall maximum sequence likelihood value. Though this is time consuming, the routines are extremely fast and can process huge numbers of cases very quickly. As such we are fairly sure we have representative solutions. We first consider the models with no autoregressive noise, and the benchmarks against which we are testing at a maximum

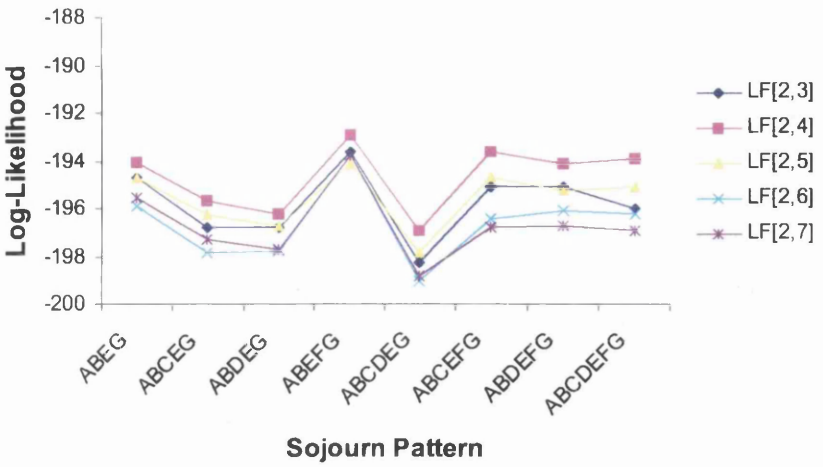

Figure 9.31: A comparison of maximised sequence log-likelihood of $LF[2, N_{(D)}]$ models for some of the most likely sojourn patterns.

likelihood value of -191.2627 and a minimum variance of 0.7996. We shall consider first those models with only gradual switching during the switch from the upper to lower regime. Figure 9.31 shows the performance of the $LF[2, N_{(D)}]$ models for different regime switching patterns.

The best performing appears to be the $L[2, 4]$ model. This is borne out if we measure the performance in terms of the estimated standard deviation of the noise, rather than the likelihood. This is displayed in Figure 9.32. It is also of note that the different possibilities of regime switching behaviour are arranged approximately in order of likelihood. The 7 switch case does seem to be the most convincing of the possibilities, as it results in higher likelihood values for all models than any other combination of pattern and model.

We go on to present the same information for the $LF[3, N_{(D)}]$ and $LF[4, N_{(D)}]$ models in Figures 9.33, 9.34, 9.35 and 9.36. The decision of which of these models performs best is not as clear. The best performing in terms of likelihood is $LF[3, 4]$ but the lowest noise levels are usually recorded by $LF[3, 5]$. Of the $LF[4, N_{(D)}]$ models we find $LF[4, 4]$ clearly the best. We can also see irregularities in the scores of these models. The graphs are not as predictable as in previous cases. This appears to be due to the model being capable of modelling more levels than are required by the data. There is a tendency for the level associated with the lower regime to become more volatile as the model switches between utilising all the available levels or only some. We are now in a position to compare the best model in each category with each other to determine the best fitting model overall. The results of this comparison are given in Tables 9.30 and 9.31 and then presented visually in Figures 9.37 and 9.38.

We could come to different conclusions depending on how we choose to measure the models. Overall the highest likelihood and lowest noise levels are given by $LF[3, 4]$, closely

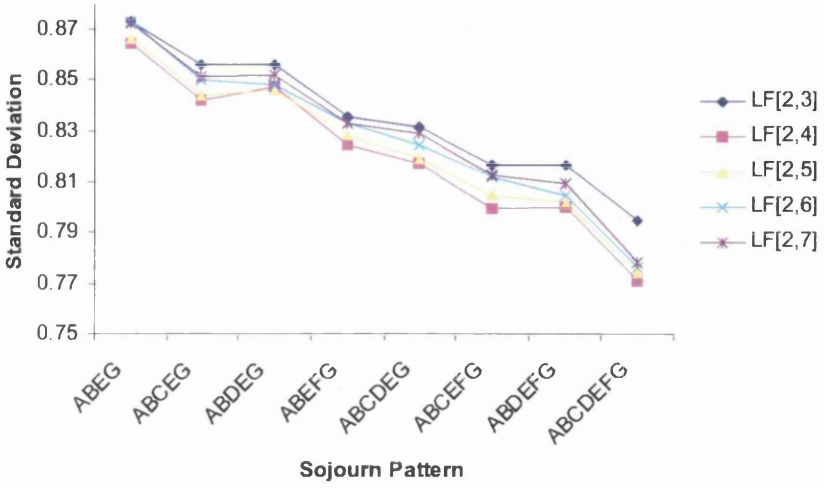

Figure 9.32: A comparison of standard deviation of residual noise of $LF[2, N_{(D)}]$ models for some of the most likely sojourn patterns.

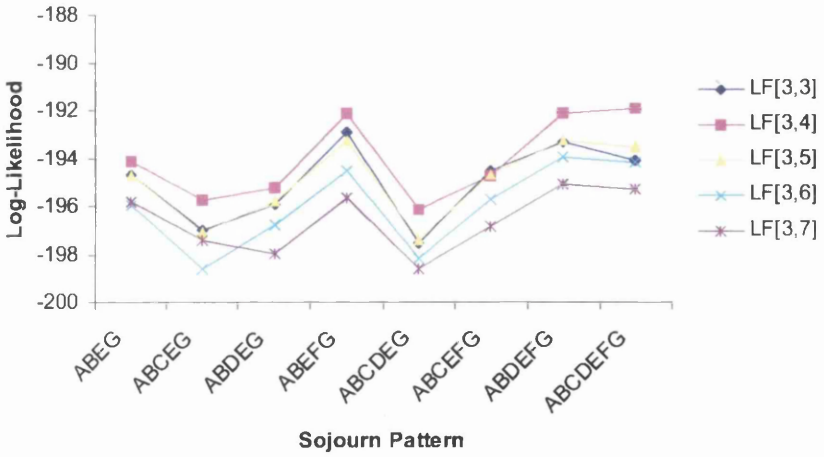

Figure 9.33: A comparison of maximised sequence log-likelihood of $LF[3, N_{(D)}]$ models for some of the most likely sojourn patterns.

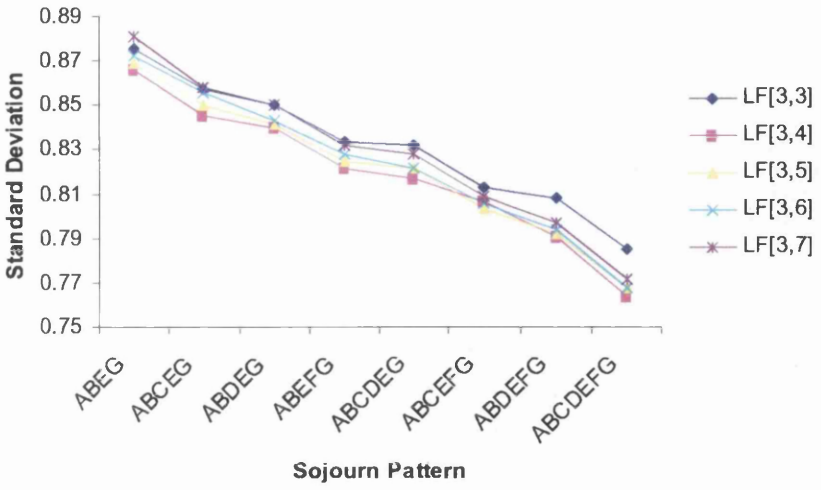

Figure 9.34: A comparison of standard deviation of residual noise of $LF[3, N_{(D)}]$ models for some of the most likely sojourn patterns.

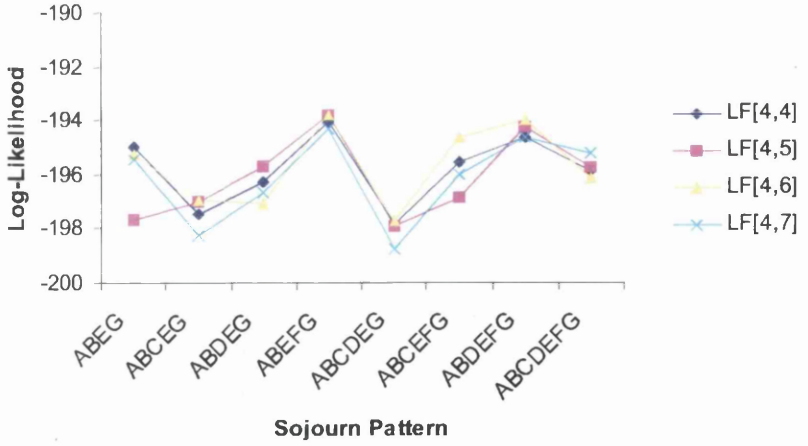

Figure 9.35: A comparison of maximised sequence log-likelihood of $LF[4, N_{(D)}]$ models for some of the most likely sojourn patterns.

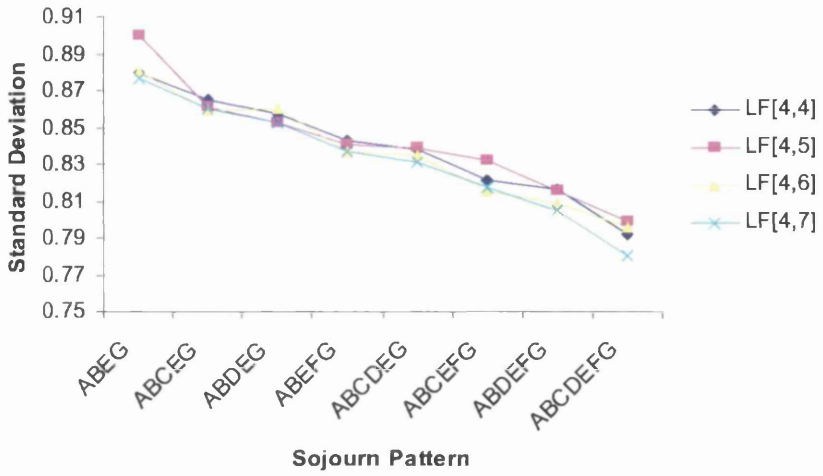

Figure 9.36: A comparison of standard deviation of residual noise of $LF[4, N_{(D)}]$ models for some of the most likely sojourn patterns.

Sojourns	LF[2, 4]	LF[3, 4]	LF[3, 5]	LF[4, 4]
ABEG	-194.0555	-194.0711	-194.6710	-194.9722
ABCEG	-195.6397	-195.7239	-197.0488	-197.4665
ABDEG	-196.2372	-195.2358	-195.7690	-196.2810
ABEFG	-192.8949	-192.1330	-193.2963	-194.0762
ABCDEG	-196.8999	-196.1733	-197.3726	-197.8110
ABCEFG	-193.5888	-194.7348	-194.6486	-195.5243
ABDEFG	-194.1041	-192.1136	-193.2335	-194.6230
ABCDEFG	-193.8598	-191.9329	-193.5451	-195.8591

Table 9.30: A comparison of the maximised sequence log-likelihood of the leading Line Fit models for different sojourn patterns.

Sojourns	LF[2, 4]	LF[3, 4]	LF[3, 5]	LF[4, 4]
ABEG	0.8649	0.8664	0.8690	0.8794
ABCEG	0.8422	0.8457	0.8499	0.8647
ABDEG	0.8475	0.8396	0.8416	0.8569
ABEFG	0.8247	0.8213	0.8246	0.8426
ABCDEG	0.8171	0.8167	0.8214	0.8382
ABCEFG	0.7994	0.8064	0.8033	0.8215
ABDEFG	0.7999	0.7904	0.7923	0.8159
ABCDEFG	0.7714	0.7639	0.7672	0.7919

Table 9.31: A comparison of the standard deviation of the residual noise for each of the leading Line Fit models for different sojourn patterns.

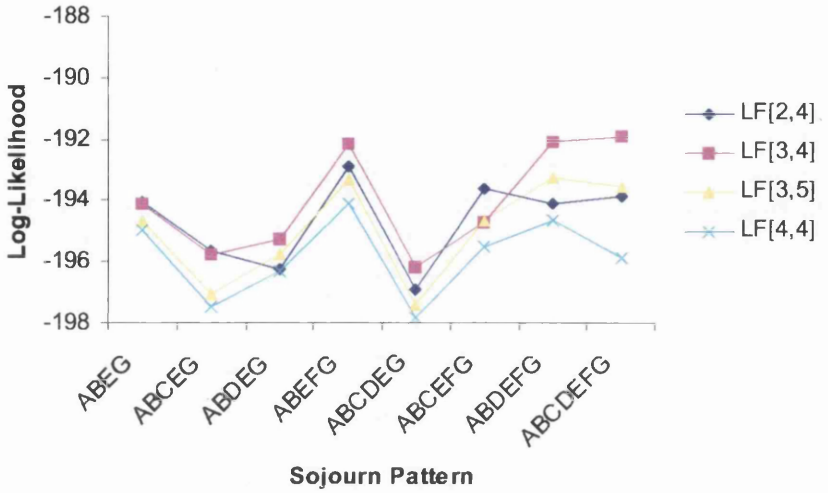

Figure 9.37: A comparison of maximised sequence log-likelihood of the best performing Line Fit models for the most likely sojourn patterns.

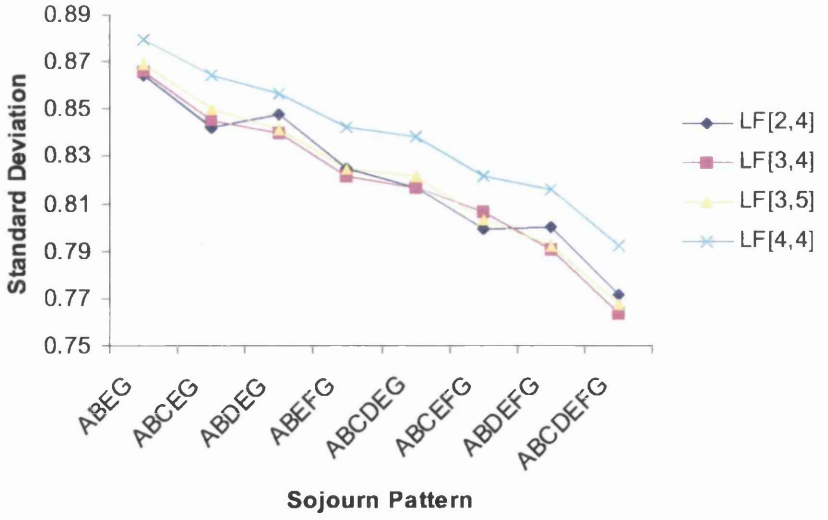

Figure 9.38: A comparison of the standard deviation of the residual noise of the best performing Line Fit models for the most likely sojourn patterns.

	Log-Likelihood	Std. Dev.	Variance
LF[2, 2]	-191.2676	0.8351	0.69739
LF[3, 5]	-193.2963	0.8246	0.67997
Gain	-2.0287	-0.0105	-0.0174
Change(%)		-1.3%	-2.5%

Table 9.32: A comparison between the noise levels and maximised sequence log-likelihood for LF[2,2] and LF[3,5] where they conform to an ABEFG sojourn pattern.

	Log-Likelihood	Std. Dev.	Variance
LF[2, 2]	-191.2676	0.8351	0.69739
LF[3, 4]	-192.1330	0.8213	0.67453
Gain	-0.8654	-0.0138	-0.0229
Change(%)		-1.7%	-3.3%

Table 9.33: A comparison between the noise levels and maximised sequence log-likelihood for LF[2,2] and LF[3,4] where they conform to an ABEFG sojourn pattern.

followed by $LF[3, 5]$. Both of them outperform the $LF[2, 2]$ case on both counts. The two key sojourn patterns in the $LF[2, 2]$ case are the cases ABEFG and ABCDEFG. We now require a slightly more precise comparison between them. In Table 9.32 we compare $LF[2, 2]$ and $LF[3, 5]$ in terms of their maximised sequence likelihood and the standard deviation of the residual noise at the SMLE, given that they broadly conform to the ABEFG sojourn pattern. Table 9.33 gives the same comparison between $LF[2, 2]$ and $LF[3, 4]$. Tables 9.34 and 9.35 repeat these measurements for another pattern of sojourns.

The gain is marginal in the first case amounting to only a slight improvement in likelihood. In the second case, however, the story is quite different. The difference in likelihood is sizeable and there is a distinct reduction in noise level. As the comparison is between different models it is difficult to say whether this consitutes a 'significant' improvement or not. All we can say is that there is plenty of evidence that this kind of gradual switching mechanics may occur in real world data. The parameter estimates for these two models are shown in Tables 9.36 and 9.37.

	Log-Likelihood	Std. Dev.	Variance
LF[2, 2]	-194.5511	0.7996	0.63936
LF[3, 5]	-193.5451	0.7672	0.588596
Gain	1.0060	-0.0324	-0.0508
Change(%)		-4.1%	-7.9%

Table 9.34: A comparison between the noise levels and maximised sequence log-likelihood for LF[2,2] and LF[3,5] where they conform to an ABCDEFG sojourn pattern.

	Log-Likelihood	Std. Dev.	Variance
LF[2, 2]	-194.5511	0.7996	0.63936
LF[3, 4]	-191.9329	0.7639	0.583543
Gain	2.6182	-0.0357	-0.0558
Change(%)		-4.5%	-8.7%

Table 9.35: A comparison between the noise levels and maximised sequence log-likelihood for LF[2,2] and LF[3,4] where they conform to an ABCDEFG sojourn pattern.

Sojourn Pattern	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
ABEG	-194.0711	0.2278	0.0343	-1.1199	0.9758	0.8664
ABCEG	-195.7239	0.2554	0.0438	-1.1447	1.0073	0.8457
ABDEG	-195.2358	0.2329	0.0445	-1.0303	1.0259	0.8396
ABEFG	-192.1330	0.2436	0.0441	-1.1596	1.0289	0.8213
ABCDEG	-196.1733	0.2553	0.0545	-1.0591	1.0595	0.8167
ABCEFG	-194.7348	0.2455	0.0550	-1.0316	1.0733	0.8064
ABDEFG	-192.1136	0.2455	0.0550	-1.0777	1.0826	0.7904
ABCDEFG	-191.9329	0.2647	0.0655	-1.1057	1.1189	0.7639

Table 9.36: The maximum sequence likelihood estimates for the parameters of the LF[3,4] model using different sojourn patterns.

Sojourn Pattern	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
ABEG	-194.6710	0.2162	0.0345	-1.2602	0.9759	0.8690
ABCEG	-197.0488	0.2230	0.0448	-1.1644	1.0200	0.8499
ABDEG	-195.7690	0.2230	0.0448	-1.1966	1.0251	0.8416
ABEFG	-193.2963	0.2140	0.0452	-1.1823	1.0425	0.8246
ABCDEG	-197.3726	0.2363	0.0554	-1.1903	1.0640	0.8214
ABCEFG	-194.6486	0.2279	0.0559	-1.1787	1.0823	0.8033
ABDEFG	-193.2335	0.2126	0.0569	-1.0833	1.1038	0.7923
ABCDEFG	-193.5451	0.2246	0.0685	-1.0919	1.1471	0.7672

Table 9.37: The maximum sequence likelihood estimates for the parameters of the LF[3,5] model using different sojourn patterns.

Sojourn Pattern	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
ABEFG	-188.5315	0.4209	0.0413	-1.2958	0.9220	0.8179
ABCDEFG	-192.9228	0.3935	0.0608	-1.0657	1.0045	0.787

Sojourn Pattern	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
ABEFG	0.1718	0.0698	0.0108	-0.0650
ABCDEFG	0.1201	-0.0505	0.0629	0.1184

Table 9.38: The results of maximising the sequence likelihood of the LF[2,2] with residual noise modelled using an AR(4). In this case the hill climber is started from the best sequence of LF[2,2]

Sojourn Pattern	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
ABEFG	-188.3305	0.3620	0.0419	-1.1614	0.9454	0.8110
ABCDEFG	-187.4716	0.2178	0.0691	-0.5416	1.1511	0.7315

Sojourn Pattern	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
ABEFG	0.1896	0.0997	-0.0205	-0.1606
ABCDEFG	-0.0361	-0.1074	-0.2859	-0.2928

Table 9.39: The results of maximising the sequence likelihood of the LF[2,2] with residual noise modelled using an AR(4). In this case the hill climber is started from the the parameter estimates (and sequence estimate) obtained by Hamilton

9.5.8 Auto-Regressive Noise

The final question for us to ponder in relation to these type of models concerns the kind of auto-regressive noise (AR(4)) found in Hamilton's original Markov switching model. How does it change the way the model fits and what effect will it have on our gradual switching models. In order to study this we could add it directly to the Ladder models but this would prove rather time consuming, given the number of levels we would be using. Instead we add it to the Line Fit model, a process which is very straightforward We denote the Line Fit model with added 4th order autoregressive noise $LFAR[N_{(U)}, N_{(D)}]$. If we consider the two most significant sojourn patterns and apply a 4th order autoregressive process to model the noise we find that the first results do not match those obtained by Hamiltons filter. It differs both in the predicted sojourns and in that there is little evidence of autoregressive noise. However, it seems that this is a local maximum only. If we restart the hill climber from other positions (including Hamiltons original parameter estimates) we obtain another, better maximum. After exhaustive repetition of this process it seems likely that this represents the global maximum. Not only are they a distinct improvement, they are very close to Hamiltons original estimates.

They show similar evidence of higher order autoregressive behaviour in the residuals. What has changed from the first case to the second is not simply due to a case of finding another, better, solution to the same problem but a marked change in the sojourns in the

Model	Direction of Regime Switch (Dn refers to a transition down to regime 0)													
	Dn	Up	Dn	Up	Dn	Up	Dn	Up	Dn	Up	Dn	Up	Dn	Up
LFAR$^{(1)}$[2,2]	5	9	22	24	34	35	70	72	89	92	112	113	118	122
LF[2,2]	5	9	22	24	34	35	70	73	89	92	112	113	118	122
LFAR$^{(2)}$[2,2]	5	9	19	24	32	35	70	75	87	92	109	113	116	123
LF[3,4]	4	8	20	24	32	34	69	72	87	92	110	113	116	122
LF[3,5]	4	8	19	24	32	35	68	72	86	92	109	113	116	122

Table 9.40: A comparison between the best sequences suggested by the different solutions we obtained.

	LFAR$^{(1)}$[2,2]	**LF[2,2]**	**LFAR$^{(2)}$[2,2]**	**L[3,4]**	**L[3,5]**
LFAR$^{(1)}$[2,2]	0	1	16	12	17
LF[2,2]	1	0	15	15	18
LFAR$^{(2)}$[2,2]	16	15	0	10	9
LF[3,4]	12	15	10	0	5
LF[3,5]	17	18	9	5	0

Table 9.41: A comparison of the similarity of the most likely regime sequence for different models. Similarity is measured by the number of times the sequences disagree.

regime pattern. We shall concentrate now on the ABCDEFG sojourn pattern as this seems the most likely. While the first solution for the Line Fit model with autoregressive noise, $LFAR^{(1)}[2,2]$ (see Table 9.38) showed only 1 change of regime switch time from those for $LF[2,2]$, the second solution $LFAR^{(2)}[2,2]$ (see Table 9.39) differs in 7 of the 14 switch points. The beginning and end points of each of these 7 periods of recession for the best sequence for each model are shown in Table 9.40.

One way we can measure the similarity (or otherwise) of the different solutions is to count the number of times the regime differs between a pair of models over the 131 points in the time series. We can tabulate this measure of similarity (see Table 9.41) and then perform an analysis to see how closely related the different groups are.

The results of this analysis are displayed in Figure 9.39. The three main groups are apparent and $LFAR^{(2)}[2,2]$ is closer to the Ladder models than to $LF[2,2]$.

The way in which $LFAR^{(2)}[2,2]$, $LFAR[3,4]$ and $LFAR[3,5]$ distort the regime pattern from that of $LF[2,2]$ is quite similar in many ways. If we look more closely at the actual effect on the timings we find that all three move to earlier downward switching (see 9.42).

The question we begin to ask ourselves is whether the AR noise and the Ladder models are both capitalising on the same behavioural mechanics. This is borne out further if we look at the Mean Squared Difference (MSD) between the level of the signal for each of the the models. This information is given in 9.43. Each of the locations refers to different subsets of the series. The row marked 'During [3,4] Recession' includes those observation times that fall within recessions predicted by the $LF[2,2]$ model. Those 'Close to Switches' are those within 4 time units of a predicted switch point (for any model).

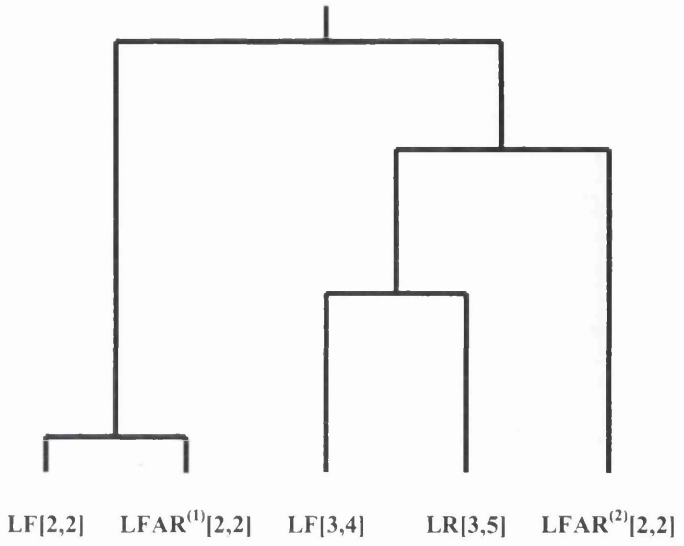

Figure 9.39: A tree structure showing the similarity between the switching point sequences suggested by the different models

	DownSwitch		*UpSwitch*	
	Early	*Late*	*Early*	*Late*
$\mathbf{LFAR^{(2)}[2,2]}$	5	0	0	2
$\mathbf{LF[3,4]}$	7	0	2	0
$\mathbf{LF[3,5]}$	7	0	1	0

Table 9.42: The deviation of the best sequence of several models, from that of the two-level signal found using LF[2,2].

Location	**Mean Square Difference in Signal**			**Diff**
	$LFAR^{(2)}[2,2]$ *vs* $LF[3,4]$	$LFAR^{(2)}[2,2]$ *vs* $USGNP$	$LF[3,4]$ *vs* $USGNP$	
Full Time Series	0.2283	0.5351	0.5836	0.0485
During [2,2] Recession	0.4462	0.5358	0.5432	0.0074
During [3,4] Recession	0.3391	0.4396	0.4505	0.0109
Close to Switches	0.4002	0.4879	0.5337	0.0458

Table 9.43: Mean Sum Difference of the levels of the signal for several models during subsets of the series

Model	Log-Likelihood	$\hat{a}$	$\hat{b}$	$\hat{l}_0^{(U)}$, $\hat{l}_0^{(D)}$	$\hat{l}_{N_{(U)}-1}^{(U)}$, $\hat{l}_{N_{(D)}-1}^{(D)}$	$\hat{\sigma}$
LF[2, 2]	-194.5511	0.4161	0.0604	-1.0912	0.9899	0.7996
LFAR[2, 2]	-187.4716	0.2178	0.0691	-0.5416	1.1511	0.7315
LF[3, 4]	-191.9329	0.2647	0.0655	-1.1057	1.1189	0.7639
LFAR[3, 4]	-189.7035	0.2555	0.0661	-1.0123	1.1219	0.7497

Model	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
LF[2, 2]	0.0000	0.0000	0.0000	0.0000
LFAR[2, 2]	-0.0361	-0.1074	-0.2859	-0.2928
LF[3, 4]	0.0000	0.0000	0.0000	0.0000
LFAR[3, 4]	-0.0371	0.0070	-0.1421	-0.1463

Table 9.44: The maximum sequence likelihood estimates of the parameters of different models using the ABCDEFG sojourn pattern.

	Without AR Noise	With AR noise	Change	%
$LF[2, 2]$	0.7996	0.7315	-0.0681	-8.5%
$LF[3, 4]$	0.7639	0.7497	-0.0142	-1.9%

Table 9.45: The reduction in noise level due to the addition of autoregressive noise.

Although the $LF[3, 4]$ records higher MSD with the data series the difference in the MSD between the models is actually smaller during the switching episodes that at other times.

The final step for this data set is to take the Line Fit model and add autoregressive noise to it. When we do this with the $LF[2, 2]$ model we find a reduction in noise levels of over 8%. If some of this can be explained by gradual switching mechanics we would not expect another similarly large reduction. Given the time it takes to explore the complex parameter space when working with lines of best fit, no attempt is made at this stage to be comprehensive. We only consider the best performing of the AR-free models, the $LF[3, 4]$, as a candidate. It is likely, as with the $LF[2, 2]$ case, that the precise specification of the model that performs best without AR, continues to dominate with it. But as the likelihood spaces for these models are complex and the results intended only as illustrative we shall restrict ourselves to the $LF[3, 4]$ model. The autoregressive noise is added in precisely the same way as with the $LF[2, 2]$ model we have been working with already. Despite the drawbacks of working with the line of best fit it is likely, in such a simple example, that our maximum sequence likelihood estimate will be either the global maximum or closely related to it. The parameters of the fitted model are shown in Table 9.44.

The improvement is hardly stunning. We find only a slight reduction in noise levels after adding AR to the $LF[3, 4]$ model. In fact the noise levels do not even fall to the same levels as the $LFAR[2, 2]$ model (see 9.45).

Whereas the $LF[2, 2]$ model lost over 8% of its noise through introducing an autoregression the $L[3, 4]$ fared much less well, only dropping around 2%. We can also see another

suggestion of the reduced importance of the autoregression in the much smaller coefficient values. The fact that they are still a long way from zero tells us that there is still some systematic behaviour pattern left in the data. This may be, as Hamilton proposed, due to the way the data was collated and de-seasonalised. But if this is not the case, or only partly so, then it suggests that, even if there is some form of gradual switching mechanics at play here, we have not necessarily specified the model in such a way to fully exploit it.

9.5.9 Summary

The earlier data sets were chosen for their compliance with gradual switching models. They were relatively short and contained behaviour that strongly suggested the kind of mechanics we sought. They served a useful purpose in introducing the methods we have been proposing and allowing an overview of the strengths (and weaknesses) of working with gradual switching models. This final data set was not selected as such. Given that Hamilton's paper served as a starting point for this piece of research it seemed fitting to apply these new tools to it. The results have been mixed. We present here a summary of the steps taken in this case study and our conclusions.

The first step was to construct the model used by Hamilton and obtain his results. This was quickly done. We then explored the possibility of slightly modifying the smoothing algorithm to work with a distribution based on previous switching times rather than a full history (Pointer Filter). The results we obtained were surprisingly close to the results obtained by the original smoother. In a case such as this where switching maybe relatively infrequent this may allow higher order autoregressions or larger numbers of levels to be used without introducing unmanageable matrix sizes. The attempts made to use this approach to introduce gradual switching were of limited success. The complex structure of the likelihood space led to difficulty obtaining maxima. As this was not considered central this line of research was abandoned.

The Ladder model was introduced but the levels of noise were simply too high for any clear distinction to be made between the regime led structure highlighted in previous work and a random walk between the levels of the process. Given the fact that we had found similar problems with previous data sets, more ideal in terms of structure, this was not surprising.

The Slide model produced more positive results, as has been the case earlier. The rigid structure of the model ensured that switching was not undertaken lightly ensuring the retention of the expected regime structure. The results of this must be interpreted carefully as the structure of the model drives down the likelihood value. But when a consideration is made of the level of noise we find the $S[3,5]$ is favoured, with lesser support for neighbouring models including $S[2,7]$ and $S[2,8]$. It is likely that this does represent evidence of gradual switching although a more formal test would need to be devised.

In an attempt to provide such a formal test a more general model was proposed capable of approximating both the two-regime, two-level Markov chain and a Slide model. In order

to ensure the results of this test can be treated as reliable we checked to see how closely we can approximate the tested models. The closeness of the approximation is probably good enough to treat the conclusions as relevant although obvious concerns remain. We find a sufficiently high test statistic for the likelihood ratio test to conclude there is evidence for the existence of gradual switching.

Following on from this we introduce the Unravelled model and use it to estimate the number of regime switches that are predicted to occur within the data set. We find that when working with the two-state model it predicts 16 switches although two are so close as to be combined when autoregressive noise is added. We opt for the, only slightly less probable, 14 switch solution and define these 7 possible sojourns in the lower regime. We then work across a range of models, testing each for different numbers of regimes. We conclude that there are an even number of switches and that all Slide models, apart from $L[2,2]$, agree on 7 regimes of recession. Several models perform well under these conditions, namely $S[2,4,14]$, $S[2,7,14]$, $S[2,8,14]$ and $S[3,4,14]$. The reduction is noise levels is sufficient to conclude they are relevant

We also show that that the Ladder models predict a number of switches increasing with the number of levels of the model, highlighting the problem with working with Ladders. Even under such highly constrained circumstances it seems difficult to obtain any useful inference from the Ladder model when working with anything other than ideal or simulated data.

We then take another approach by taking the understanding of the data set we have obtained through the use of our other models and impute the best sequence of regimes. The consideration here is not so much with finding the best solution as with constraining the ladder models in a way that make it possible to test for the existence of gradual switching rather than a test of the capabilities of the Ladder model. To optimise efficiently using only a sequence likelihood algorithm is tricky to say the least. We only attempted to use it to modify a proposed regime pattern and optimise within its neighbourhood. Extensive repetition ensured that the results were reasonably likely to be very close to the global maxima. We found that two models stood out from the rest, the $LF[3,4]$ and $LF[3,5]$ with sojourn pattern ABCDEFG (see Figure 9.24). They were capable of reducing the sequence log-likelihood and noise level significantly, when a consideration is made of the higher order of the added autoregression in Hamilton's model.

Finally we fitted a handful of the best performing models but with autoregressive noise added. We found large reductions in the noise levels for the $LFAR[2,2]$ model, as did Hamilton. We found much smaller reductions in the noise levels for the gradual switching models, only falling to similar levels as those shown by $LFAR[2,2]$. It seems likely that some of the autoregressive pattern could be explained by gradual switching. To go further in confirming this we found much reduced coefficient estimates for the autoregressive noise.

In conclusion, we find that in all the different approaches we have taken to try to uncover and model gradual switching mechanics in this particular data set we have found very

similar results. Almost exclusively the most favoured model incorporates a short transition during an upward regime shift and a longer transition for a downward shift. The models would suggest a transition into recession taking somewhere between three quarters and two years, while the transition during recovery will take only 1 or 2 quarters. We even found experimental evidence to support this type of model. It seems likely that some form of gradual switching does take place in this data set although the continued (although reduced) existence of autoregressive behaviour may suggest that we have not hit upon the precise specification.

Chapter 10

Working with Gradual Switching Models

In the light of the developments made in this work we consider the problems of working with gradual switching models, and discuss their possible solutions and propose areas for future study.

10.1 A Summary of Developments

From the beginning our intention was not to introduce new models for the sake of it. It is likely the models we worked with, as with the Beta generation, may have many interesting properties or have greater possibilities than we have exploited. Nevertheless our interest was always primarily in the capacity to detect and model the transition between levels of a two-regime Markov process.

The first model we focussed on was the Filtered Markov process. Prompted by the observed signal of a two-level Markov process when viewed through a low-pass filter, this stochastic process proved interesting in ways we had not anticipated. It quickly became clear that the model would be far too flexible and adaptable to be effectively applied to many real case studies, apart from those which we know are controlled by appropriate underlying mechanics. Neither are the models well served in terms of the algorithms we use to fit them, which reward structure on one level while ignoring it on another. It was this difficulty in levering this model into the role we had chosen for it that drove us to experiment with discrete approximations. Using these we were able to at least measure the bounds within which we were capable of operating.

The Filtered Markov model was a little more productive in an entirely unforseen way. When we discovered it had a familiar stationary distribution the advantage only seemed significant in that it was easy to work with. Quickly, though, it became obvious that we could use this model to generate Beta variates. Surprisingly enough a comparison with the currently available algorithms showed that it was very competitive in terms of simplicity

and speed.

From the Filtered Markov model the next natural step was to the Ladders, a general class of models that could capture a wide range of switching behaviour. Great improvements were indeed possible with these simple models, but they still had a tendency to move away from the purpose for which they were intended and towards entirely different solutions. This should perhaps not be seen as a problem in general, but in so far as our intention was to restrict them to one area it proved problematic. The development into the Slide models finally gave us this capacity we had been seeking to tie the models to regimes rather than rapid switching. They tended to match the clear patterns indicated by the simple Markov models, even going further in providing stronger inference about current regime.

Further progress was possible through the introduction of the Unravelled Ladder models. By forcing the process to make only a fixed number of changes we were able to draw inference only from a realistic switching pattern, unlike the original algorithms that may draw likelihood from a host of improbable scenarios. They also allow us to introduce the number of switches a process may have made as a parameter that can be estimated. These Unravelled models can also be combined with full sample smoothing to help provide even stronger inference on regime distribution.

Finally, when faced with the problem of demonstrating the existence of gradual switching mechanics in econometric data, namely records of the US GNP figures, we turned to working with Line Fit models to impute the most likely regime history. As far as possible we had sought to avoid working with these due to their complex structure and the difficulties in finding global maxima. The extensive knowledge we had built up with the array of models we had used gave us an edge and allowed us, in all probability, to find the overall solutions.

The evidence we uncovered did indeed suggest that this data set appears to exhibit gradual switching dynamics. We found lower levels of noise for the gradual switching Line Fit models, and a good agreement between their predictions and official datings of business cycles. These regime dates were as good as those produced by Hamilton's original model, but only those obtained with the inclusion of autoregressive noise. We would even suggest that at least some of the autoregressive noise evident in this, earlier, model is probably due to some kind of gradual switching behaviour during regime transitions.

10.2 Possible Developments

In the course of this research certain possibilities have arisen that seem to me worthy of further examination. In some of these cases a certain amount of work was done that has not been included here, either because no useful progress was made or it did not fit well into any of the different sections of this thesis while other directions were simply left open. Of the first category we should draw attention to the most interesting cases.

10.2.1 Random Variate Generation

Initially it was quite a pleasant surprise when the Filtered Markov Process turned out to have a Beta stationary distribution. We experimented with variations on this theme such as extending the stochastic process to higher numbers of dimensions. This had only limited success, throwing up distributions like Dirichlet's. Despite this it seems likely that there is more potential in this area for development. There are close links between the Beta distribution and the Generalised Hypergeometric distribution. The distributions of both the Beta and Filtered Markov distribution are constructed using Generalised Hypergeometric Functions. It is possible that some general relationship exists between this kind of process and a subclass of the Hypergeometric. This work was well beyond the scope of this thesis and was not explored.

10.2.2 Smoothing Algorithms

Over the course of this research we have used the smoothing algorithm, or versions of it, extensively. In particular we have found it to be extremely effective in identifying the distribution of the process between the two regimes. We have also found, in the simplest case, that it relies heavily on the fact that the choice between the two regimes is rather stark. The potential for mis-classification is relatively slight. When extra levels are added to the regimes the possibility of mis-classification becomes quite large and switching too easy. This is especially true when you consider that most of the time we are working with differenced processes in which the consequence of a poorly determined switch point may be only minor. This, coupled with the the capacity of the process to draw inference from switching that is not well supported by the data, can lead to difficulties with more complex models. There are two elements we feel could be addressed in order to improve the effectiveness of these smoothers; working directly with integrated process (rather than differenced series) and incorporating future data in the inference procedure.

Integrated Processes

It has sometimes been frustrating, when working with gradual switching models, to see them enthusiastically model every little twist and turn of the data set while ignoring larger scale structure. This has been true even with simulated data, although the problem is much worse with genuine time series. While appreciating that algorithms are not compelled to fit the pattern we want and may have good reason for preferring a higher frequency model, this can be somewhat frustrating. One alternative is to apply prior distributions to the parameters of the model so as to constrain it to the range we prefer. As far as possible we tried to avoid doing this so as to avoid finding ourselves in the situation of obtaining no more or less than we requested. The source of the problem becomes apparent when you consider the differenced process. The long term shifts in mean growth, obvious in an integrated process, hardly register when the same data is differenced. As with any Markov

switching model the most valuable inference you can make concerns the current state of the process. For anything except low levels of noise the distribution can appear unimodal and any separation is impossible if the order of the data is not considered. It is precisely this that rendered the Methods of Moments so limited for use with time series data. It was only with huge data sets that the subtle features that differentiate the mixed model from a single distribution become apparent.

In working with the smoothing algorithm it became clear that even with relatively high levels of noise it was often possible to obtain quite good estimates of the distribution between regimes. When future data, or indeed the full sample, was used to enhance the inference this was especially true. We tried to find ways to estimate the current value of the process from the history of the growth rate. Normally this is not possible as the history is not recorded but estimates of the expected growth rate at each point can be recalled and used to produce an expected current value. When the inference about the current state is good, tending close to 0 or 1, this estimate is quite good. In this way a short sighted placement of a switchpoint could have long term negative consequences. The method worked thus:

We start the process at $x(0)$ at time $n = 0$.

The process has levels l_0 to $l_{N_{(L)}}$.

After the first Markov step of the process we have residuals r_i, associated with the levels l_i, where

$$r_i(1) = y(1) - (x(0) + l_i)$$

rather than

$$r_i(1) = y(1) - l_i$$

Treating the residual as any error term we can obtain $f(y(1)|l_i)$. By taking these likelihood values for i we can obtain $p_i(n) = Pr(S(n) = i)$ for $i = \{0, 1\}$.
If we define $E(1)$ as the expected growth at time $n = 1$ then

$$\begin{aligned} E(1) &= l_0.p_0(1) + l_1.p_1(1) + x(0) \\ E(n+1) &= l_0.p_0(n) + l_1.p_1(n) + x(n) \end{aligned}$$

We then use this to obtain the residuals at time n by

$$r_i(n) = y(n) - (E(n-1) + l_i)$$

In practice with simple examples the results were encouraging. The problems arose in two ways.

First the algorithm works by seriously penalising switching. As a result there is a tendency for the process to switch only after strong evidence has arrived to support it. As the noise levels rise higher and higher this delay in recognising the switch points can become

Figure 10.1: A time series obtained by taking a Moving Average of the record of monthly car registrations in the US.

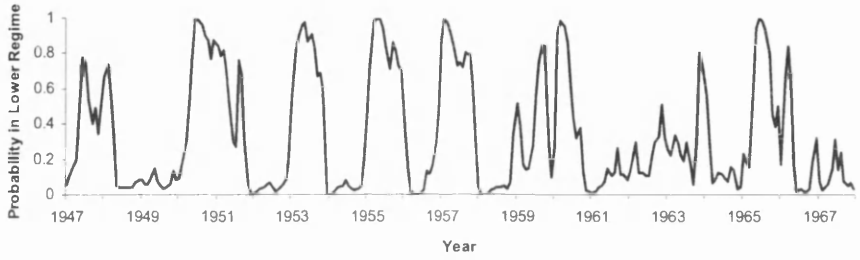

Figure 10.2: An example of the inference about the regime history (of the car registration data) using only a Basic filter.

problematic. The effect on this method is to ensure the current value of the process, as modelled through the expected growth rates, tends to lag behind the time series.

Secondly, the separation between states need to be clear, even if not consistent, so the process can construct good estimates of the long term growth rate. In order to do this it is helpful to bring in future data or even the full sample to make inference. We can give an example of this, below. We take the time series, illustrated in Figure 9.2 on page 159, representing the Registration of Cars in the United States. A moving average is then taken of the data, giving the more amenable series shown in Figure 10.1. The order of the moving average was set to 12 as there was significant annual periodicity in the data. The symmetric Slide model was fitted for a range of appropriate switch lengths and Likelihood favoured S[8,8]. When we examine the series of distributions of inferred regime the algorithm is usually quite clear (see Figure 10.2).

But to see how much this inference is improved by adding inference from the full sample a comparison should be make with Figure 10.3. Here we find almost no uncertainty on the current regime. Several quite noticeable episodes, namely around 1948 and 1963, are completely absent. When the algorithm can separate the different regimes as clearly as this it would be quite possible to estimate the current position by summing the expected

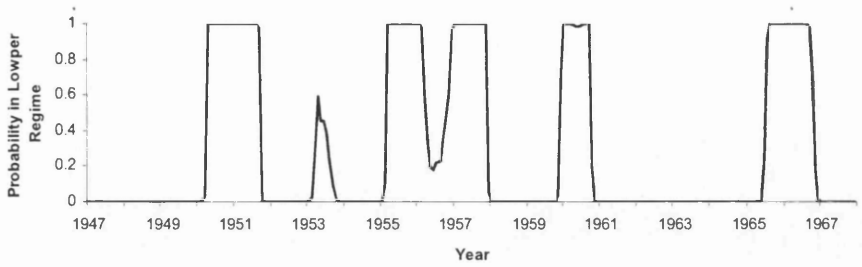

Figure 10.3: An example of improved estimation of the regime switching history possible when using Full Sample Inference.

growth rate at each point.

Overall this approach showed some promise but was not perfected. Quite soon it was abandoned in favour of working with the Unravelled models. We remain convinced that a version of this method could work effectively.

Incorporating Future Data

Extending the smoothing algorithm to include future data adds to the complexity of the process but not excessively so. Hamilton uses it to improve the inference about the current regime of the signal although his method is tedious to code. When it is applied through matrix algebra it becomes much easier and faster to apply. No attempt is made to include this future data in the evaluation of likelihood but the consequences of not doing so for the two-level model are not catastrophic. As we add more levels, we incorporate more flexibility. And this flexibility leads to instability. This is not a failure of the model, rather a failure of the algorithm, as the information necessary to detect a change of regime is spread over several time intervals rather than just one. What is needed is some way of penalising a switch in regime not by assigning a punitive likelihood cost to it, as is currently done, but by exploring the future consequences of such a move. In order to do this an entirely new way of obtaining the likelihood will be necessary, as this likelihood cost for switching must be paid in order to allow the switching intensity parameters to be accurately estimated. What we propose is that it is possible to obtain the inferred distribution of regimes first, then use this to find the likelihood. This approach could be seen as a development of the application of the EM algorithm to this problem. In the EM algorithm the full sample inference about the regime history is used to weight the data set from which sample estimates are obtained giving rise to the next set of parameter estimates. In this new variation the full sample estimates are obtained and then used to weight the transitions before evaluating the likelihood.

For instance, if we have a process, $y(n)$, comprising a signal with two levels (l_0 and l_1) and Gaussian noise.

We then run the smoothing algorithm using parameter estimates $\{a, b, l_0, l_1, \sigma\}$ to produce the series of distributions $\Pr[(S(n) = s(n) \mid \mathbf{y_N}]$ for all n.

Then for each n we have

$$\begin{bmatrix} \Pr[(S(n) = 0 \mid \mathbf{y_N}] \\ \Pr[(S(n) = 1 \mid \mathbf{y_N}] \end{bmatrix}$$

And we take

$$\begin{bmatrix} \Pr[(S(0) = 0 \mid \mathbf{y_N}] \\ \Pr[(S(0) = 1 \mid \mathbf{y_N}] \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix}$$

Where π is a suitable prior distribution e.g. the stationary disribution.

We then estimate

$$\Pr[S(n) = i \;\&\; S(n+1) = j \mid \mathbf{y_N}] = \Pr[(S(n) = i \mid \mathbf{y_N}].\Pr[(S(n+1) = j \mid \mathbf{y_N}]$$

And obtain the log-likelihood from a summation of these terms.

$$\begin{aligned} l(\mathbf{\Theta}, \mathbf{\Pi}) &= \sum_{n=1}^{n}\sum_{i=0}^{1}\sum_{j=0}^{1}\log_n(\Pr(S(t-1) = i \;\&\; S(t) = j)) \\ &= \sum_{n=1}^{n}\sum_{i=0}^{1}\sum_{j=0}^{1}\log_n(\Pr[(S(n) = i \mid \mathbf{y_N}].\Pr[(S(n+1) = j \mid \mathbf{y_N}]) \end{aligned}$$

Again, we had some success with this for simple cases but the method had a specific weakness. This was that there is no direct link between the value taken at consecutive time points. The process can happily draw inference from a transition from $S(t-1) = 0$ to $S(t) = 0$ and then from a transition from $S(t) = 1$ to $S(t+1) = 0$ a moment later. Of course in doing this it does not require a switching intensity at all, leading to a serious underestimation of these parameters. A lack of time precluded a thorough exploration of this problem although the method shows promise and with suitable modification would be likely to provide an alternative algorithm for fitting this kind of model.

Bibliography

Acemoglu, D. & Scott, A. (1994), Asymmetries in the Cyclical Behaviour of U.K. Labour Markets. *Economic Journal*, **104**: 1303–1323.

Albert, J. & Chib, S. (1993), Bayes Inference Via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts. *Journal of Business and Economic Statistics*, **11**: 1–15.

Atkinson, A. (1979), A Family of Switching Algorithms for the Computer Generation of Beta Random Variables. *Biometrika*, **66**: 141–145.

Bacon, D. & Watts, D. (1971), Estimating the Transition Between Two Intersecting Lines. *Biometrika*, **58**: 525–534.

Baum, L., Petrie, T., Soules, G. & Weiss, N. (1970), A Maximisation Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Mathematical Statistics*, **41**: 164–171.

Boldin, M. (1996), A Check on the Robustness of Hamilton's Markov Switching Model Approach to the Econometric Analysis of the Business Cycle. *Studies in Non-Linear Dynamics and Econometrics*, **1**(1): 35–46.

Box, G. & Jenkins, G. (1970), *Time Series Analysis, Forecasting and Control.* Holden Day.

Brillinger, D. (1975), *Time Series: Data Analysis and Theory.* Holden Day.

Brockwell, P. & Davies, R. (2002), *Introduction to Time Series and Forecasting.* Springer-Verlag, 2nd edition.

Brockwell, P. & Davis, R. (1991), *Time Series: Theory and Methods.* Springer.

Buckle, R., Haugh, D. & Thomson, P. (2002), Growth and Volatility Regime Switching Models for New Zealand GDP Data, treasury Working Paper Series 02/08 New Zealand Treasury.

Campbell, J., Lo, A. & MacKinlay, A. (1997), *The Econometrics of Financial Markets.* Princeton University Press.

Cecchetti, S., Lam, P. & Mark, N. (1990), Mean Reversion in Equilibrium Asset Prices. *American Economic Review*, **80**(3): 398–418.

Cheng, R. (1978), Generating Beta Variates with Non-Integral Shape Parameters. *Communications of the Association of Computing Machinery*, **21**: 317–322.

Clements, M. & Krolzig, H.-M. (1998), A Comparison of the Forecast Performance of Markov-Switching and Threshold. Autoregressive Models of U.S. GNP. *Econometrics Journal*, **1**(1): 47–75.

Cosslett, S. & Lee, L.-F. (1985), Serial Correlation in Discrete Variable Models. *Journal of Econometrics*, **27**: 79–97.

Davis, M. (1984), Piecewise-Deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models. *Journal of the Royal Statistical Society B*, **46**: 353–388.

Dempster, L. & Rubin (1977), Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1): 1–38.

Devroye, L. (1986), *Non-Uniform Random Variate Generation.* Springer-Verlag.

Diebold & Rudebusch (1996), Measuring Business Cycles: A modern perspective. *Review of Economic Studies*, **78**: 67–77.

Diebold, F. (1998), The Past, Present and Future of Macroeconomic Forecasting. *Journal of Economic Perspectives*, **12**(2): 175–192.

Diebold, F., Lee, J. & Weinback, G. (1994), Regime Switching with Time Varying Transition Probabilities. In: C. Hargreaves (editor), *Non-Stationary Time Series Analysis and Cointegration*, Oxford University Press, pp. 283–302.

Elderton, W. & Johnson, N. (1969), *System of Frequency Curves.* Cambridge University Press.

Engle, C. & Hamilton, J. (1990), Long Swings in the Dollar: Are they in the data and do the markets know it? *The American Economic Review*, **80**(4): 689–713.

Engle, R. (1982), Autoregressive Conditional Heteroskedasticity with Estimates of Variance of United Kingdom Inflation. *Econometrica*, **50**(4): 987–1008.

Filardo & Gordon (1994), International co-movements of business cycles., research Working Paper 94-11, Federal Reserve Bank of Kansas.

Fitzhugh, R. (1983), Statistical Properties of the Asymmetric Random Telegraph Signal, with Applications to Single Channel Analysis. *Mathematical Biosciences*, **64**: 75–89.

Garcia, R. (1998), Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models. *International Economic Review*, **39**(3): 763–788.

Ghysels, E. (1994), On the periodic structure of the business cycle. *Journal of Business and Economic Statistics*, **12**(3): 289–298.

Goodwin, T. (1993), Business cycle analysis with a Markov switching model. *Journal of Business and Economic Statistics*, **11**(3): 331–339.

Granger, C. & Anderson, A. (1978), *An Introduction to Bilinear Time Series Models.* Vandenhoeck und Ruprecht.

Hamilton, J. (1989), A New Approach to the Economic Analysis of Non-Stationary Time Series and the Business Cycle. *Econometrica*, **57**(2): 357–384.

Hamilton, J. (1990), Analysis of Time Series Subject to Changes in Regime. *Journal of Econometrics*, **45**: 39–70.

Hamilton, J. & Susmel, R. (1994), Autoregressive Heteroskedasticity and Changes in Regime. *Journal of Econometrics*, **64**: 307–333.

Hansen, B. (1992), The Likelihood Ratio Test Under Non-Standard Conditions: Testing the Markov Switching Model of GNP. *Journal of Applied Econometrics*, **7**: S61–S82, special Issue on Non-Linear Dynamics and Econometrics.

Hartley, H. (1958), Maximum Likelihood from Incomplete Data. *Biometrics*, **14**(2): 174–194.

Jalali, A. (2003a), Filtering Markovian Signals, unpublished.

Jalali, A. (2003b), On Central Moments of Beta Distributions, unpublished.

Jalali, A. (2003c), On Markovian Signals, unpublished.

Jalali, A. (2003d), Parameter Estimation for 2X2 Markovian Filtered Signals, unpublished.

Jalali, A. (2003e), The Structure of Growth, unpublished.

Kähler, J. & Marnet, V. (1994), International Business Cycles and Long-Run Growth: An Analysis with Markov Switching and Co-Integration Models. In: K. Zimmerman (editor), *Output and Employment Fluctuations*, Physica Verlag.

Keynes, J. (1936), *The General Theory of Employment, Interest and Money.* MacMillan.

Kim, C. (1994), Dynamic Linear Models with Markov Switching. *Journal of Econometrics*, **60**: 1–22.

Kim, C., Morley, J. & Piger, J. (2005), Non-Linearity and the Permanent Effects of Recessions. *Journal of Applied Econometrics*, **20**(2): 291–309, working Paper 2002-014E.

Kim, I. (1993), A Dynamic Programming Approach to the Estimation of Markov Switching Regression Models. *Journal of Statistical Computation and Simulation*, **45**: 61–76.

Konno, T. & Fukushige, M. (2002), The Canada-United States Bilateral Import Demand Functions: Gradual Swtiching in Long-Run Relationships. *Applied Economics Letters*, **9**: 567–570.

Koopmans, L. (1995), *The Spectral Analysis of Time Series.* Academic Press.

Krolzig, H. (1997), International Business Cycles: Regime Shifts in the Stochastic Process of Economic Growth., applied Economics Discussion Paper 194, University of Oxford.

Krolzig, H. & Lütkepohl, H. (1995), Konjunkturanalyse mit Markov-Regimewechselmodellen. In: K. Oppenländer (editor), *Konjunkturindikatoren. Fakten, Analysen*, Oldenbourg, pp. 177–196.

Lam, P.-S. (1990), The Hamilton Model with a General Autoregressive Component. *Journal of Monetary Economics*, **26**: 409–432.

Leamer, E. (1978), *Specification Searches: Ad Hoc Inference with Experimental Data.* Wiley.

Makridakis, S., Wheelwright, S. & R.J., H. (1998), *Forecasting: Methods and Applications.* Wiley, 3rd edition.

McCulloch, R. & Tsay, R. (1994), Bayesian Analysis of Autoregressive Time Series Via the Gibbs Sampler. *Journal of Time Series Analysis*, **15**: 235–250.

Metropolis, N. & Ulam, S. (1949), The Monte Carlo Method. *Journal of the American Statistical Association*, **44**(247): 335–341.

Neftci, S. (1984), Are Economic Time Series Asymmetric over the Business Cycle? *Journal of Political Economy*, **92**(2): 307–328.

Nelson, C. (1973), *Applied Time Series Analysis for Managerial Forecasting.* Holden Day.

Ohtani, K., Kakimoto, S. & Abe, K. (1990), A Gradual Switching Regression Model with a Flexible Transition Path. *Economics Letters*, **32**: 43–48.

Orchard, T. & Woodbury, M. (1972), The Missing Information Principle: Theory and Applications. *Proc. 6th Berkeley symposium on Mathematics Statistics and Probability*, **1**: 697–715.

Pawula, R. (1970), The Transition Probability Density Function of the Low-Pass Filtered Random Telegraph Signal. *International Journal of Control*, **12**(1): 25–32.

Pawula, R. (1986), On Filtered Binary Processes. *IEEE Transactions on Information Theory*, **IT-32**(1): 63–72.

Phillips, K. (1991), A Two-Country Model of Stochastic Output with Changes in Regime. *Journal of International Economics*, **31**: 121–142.

Priestley, M. (1981), *Spectral Analysis and Time Series*, volume 2. Academic Press.

Roberts, H. (1992), *Data Analysis for Managers.* Scientific Press, (see Marija Nobusis's 1981 SPSS primer).

Sensier, M. (1996), *Investigating Business Cycle Asymmetries in the U.K.* Ph.D. thesis, University of Sheffield, ph.D Thesis.

Slutsky, E. (1937), The Summation of Random Causes as the Source of Cyclic Processes. *Econometrica*, **5**(2): 105–146.

Stephens, M. (1974), EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of American Statistical Association*, **69**: 730–737.

Sundberg, R. (1974), Maximum Likelihood Theory for Incomplete Data from an Exponential Family. *Scandinavian Journal of Statistics*, **1**: 49–58.

Tong, H. (1990), *Non-Linear Time Series: A Dynamical Systems Approach.* Clarendon Press.

Varoufakis, Y. & Sapsford, D. (1991), Discrete and Smooth Switching Regressions for Australian Labour Productivity Growth. *Applied Economics*, **23**: 1299–1304.

Wonham, W. (1959), Transition Probability Densities of the Smoothed Random Telegraph Signal. *Journal of Electronic Control*, **6**: 376–384.

Yule, G. (1927), On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfers Sunspot Numbers. *Philosophical Transactions of the Royal Society, Series A*, **226A**: 267–298.