Inference in Bayesian Time-Series Models

Christopher Ian Bracegirdle

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy of **University College London**.

22$^{\text{nd}}$ January, 2013

Statement of Originality

I hereby declare that:

- I understand what is meant by plagiarism;
- I understand the implications of plagiarism;
- I have composed this thesis entirely by myself; and
- this thesis describes my own research.

Chris Bracegirdle
University College London
22nd January, 2013

Abstract

Time series—data accompanied with a sequential ordering—occur and evolve all around us. Analysing time series is the problem of trying to discern and describe a pattern in the sequential data that develops in a logical way as the series continues, and the study of sequential data has occurred for a long period across a vast array of fields, including signal processing, bioinformatics, and finance—to name but a few. Classical approaches are based on estimating the parameters of temporal evolution of the process according to an assumed model. In econometrics literature, the field is focussed on parameter estimation of linear (regression) models with a number of extensions. In this thesis, I take a Bayesian probabilistic modelling approach in discrete time, and focus on novel inference schemes. Fundamentally, Bayesian analysis replaces parameter estimates by quantifying uncertainty in the value, and probabilistic inference is used to update the uncertainty based on what is observed in practice. I make three central contributions. First, I discuss a class of latent Markov model which allows a Bayesian approach to internal process resets, and show how inference in such a model can be performed efficiently, before extending the model to a tractable class of switching time series models. Second, I show how inference in linear-Gaussian latent models can be extended to allow a Bayesian approach to variance, and develop a corresponding variance-resetting model, the heteroskedastic linear-dynamical system. Third, I turn my attention to cointegration—a headline topic in finance—and describe a novel estimation scheme implied by Bayesian analysis, which I show to be empirically superior to the classical approach. I offer example applications throughout and conclude with a discussion.

CONTENTS

Part I ∼ Overview

Part II ∼ Contribution

Part III $\sim$ Conclusions and Extensions

Appendices

List of Figures

Statistics, Econometrics, and Machine Learning

Switch-Reset Models: Exact and Approximate Inference

Heteroskedastic Linear Dynamical System

Bayesian Conditional Cointegration

List of Algorithms

Statistics, Econometrics, and Machine Learning

Switch-Reset Models: Exact and Approximate Inference

Heteroskedastic Linear Dynamical System

Bayesian Conditional Cointegration

PART I

Overview

CHAPTER 1

Introduction

"Certainty? In this world nothing is certain but death and taxes."

—Benjamin Franklin, 1706-1790

1.1 Time Series

The term 'time series' refers, in general, to sequential data of any form. The data may take discrete or continuous values, and form a time series because they are provided with discrete indices indicating an ordering. The index may correspond to the onset of time, with for example price or position data, or have no intuitive meaning with for example the Fibonacci sequence, or genetic sequence data.

The problem of searching for patterns in observed sequential data is both well known and widely studied, transcending fields of research and types of data. There is a very wide variety of problems and correspondingly wide variety of solutions. In general, one is interested in modelling time series in order to understand more about something that occurs over time, in order to make decisions, predictions, or simply to understand more about the world.

A great example of the impact of development in time series modelling is the concept of 'cointegration', which is concerned with the links between two or more things that happen over time simultaneously and a topic I turn to in detail in chapter 5. The concept was a key part in the award of the Nobel prize for economics to Engle and Granger in 2003.

1.2 The Bayesian Way

In my approach to time-series analysis, I aim to consider the uncertainty in assumptions I make about how a series evolves and the parameters I use in modelling—making the problem one of determining patterns in data under uncertainty. I then seek an understanding of the temporal relationships between what is observed. Whilst very little is ever certain, dealing with uncertainty is by its nature a difficult problem for mathematics, which is founded on sound logical inference. Broadly speaking, one may have some prior belief about what a value is likely to be, and seek to update that belief based on what is observed in practice. From this inferred understanding of a 'random' process, one may refine an understanding of the process to make better predictions, or alternatively derive other conclusions that may be helpful. This approach to quantifying uncertainty is broadly understood as the modern approach to 'Bayesian statistics', which fundamentally relies on the definition of conditional independence to update one's belief based on data.

1.3 Research Scope and Methodology

My research is primarily focussed on exact inference in Bayesian time-series models in closed form. Observations are assumed to be made in discrete time, which is to say that the evolution of a process is observed at a finite number of time-points, usually at common intervals. When no closed form algorithm is available or efficient, I have taken the approach of analytical approximation, rather than stochastic approaches. In particular, I have not undertaken any work in the field of estimation of posteriors through sampling, since my main focus is on understanding the mathematical properties relating to exact probabilistic inference in time-series models. When exact inference is not possible, I aim to understand what are the mildest analytical approximation approaches that render the inference problem tractable.

My contributions aim to be methodological with broad applicability; I have however shown some example applications towards the end of each contribution chapter. Such experiments with data are not intended as exhaustive data analysis but rather to indicate how the models perform on real problems.

1.4 Contribution

The contributions of this thesis can be ascribed to three main planks, each forming one of three main chapters. First, I discuss a class of latent Markov model with internal process resets, and demonstrate how exact inference algorithms can be derived; I extend the model to allow state switches also, and show an approximate inference routine that is computationally efficient. The resulting switching model is very inexpensive to deploy when compared with switching models without internal process resets such as the exponentially-complex switching linear-dynamical system. Second, I discuss the problem of unknown variance regimes in Bayesian time series models. By appealing to my earlier work on internal process reset models, I show an inference scheme for a 'regime-switching' model in which the regime variance parameters are not known *a priori*. This simplifies the requirement to specify parameters for each state in a switching system and reduces the burden of specifying highly-sensitive values in advance, detecting regimes with 'unknown-but-different' characteristics. Third, I take a slightly different tack and present a novel approach to the estimation of a cointegration relationship. I show that my Bayesian-inspired approach to estimation in this model is empirically superior to the classical method, while remaining both flexible and simple. Finally, I apply the reset model to cointegration estimation to show how a cointegration relationship can be detected even if it doesn't apply for the whole time series sample. Previous work on this 'segmented cointegration' problem has placed *a priori* restrictions on the number of regimes.

1.5 Application

The algorithms and methods presented in this thesis are expository in nature, but designed with a view to real-world applications. One clear application domain is in the world of quantitative finance; for example, having reliable generative models for the evolution of an asset price enables one to benefit from any predicted price increase or decrease by taking the appropriate position through buying or selling the asset (respectively). The work of chapter 5 is focussed on the concept of cointegration, and it is useful to motivate the concept with a simple explanation of the benefit of the model. If it can be shown reliably that the prices of two assets move together in some well-defined way (such as a linear relationship provided by cointegration), one may seek to profit from this knowledge if the two assets temporarily move out of sync. For example, suppose that it is known that the price of asset a tends to be approximately twice that of asset b, but that right now the prices are about the same. We may profit on the expected return to equilibrium by buying asset a and selling asset b; using this approach, we are only interested in the relationship between the two prices not the absolute prices of a or b since we have taken an offsetting position. This motivates a desire for a reliable method to detect these relationships—something discussed in detail in chapter 5.

1.6 Thesis Structure

The main body of this thesis is split into three parts. Part I includes this introductory chapter and goes on to set out significant background material relevant to the results of the proceeding work, from the fields of machine learning, statistics, and econometrics. Part II comprises the main novel analytical contribution of the thesis and is split into three chapters, each detailing an approximately homogeneous contribution. Taken together, the three planks of work are related under the broader umbrella of novel approaches to Bayesian inference in time series. Finally, Part III ties the thesis together by providing summary conclusions of Part II and discusses some opportunities for further study. Several appendices are also included, setting out some key definitions and derivations.

1.7 Notation

Representing distributions over many variables can be confusing and it is necessary to set out how I deal with the problem. Throughout, a variable is written as a lower-case letter such as x, and for time series, the sequential index is given by a subscript such as x_t for the value of variable x at point t. When considering collections of the variable x over a set of indices $\{a, a+1, \ldots, b-1, b\}$, I abbreviate the series $x_a, \ldots, x_b \equiv x_{a:b}$. In the case $b < a$, I assume $x_{a:b} = \emptyset$, and also $x_n = \emptyset$ for all $n < 1$. When a variable has multiple dimensions, I write it as a vector $\mathbf{x}$. Matrices are written in capitals, $\mathbf{M}$.

Throughout the thesis, the operator $p(\cdot)$ represents a probability density function in an intuitive way. The distribution x conditioned on a realised value y is written as $p(x|y)$.

A major consideration in Bayesian inference is marginalisation, or 'summing out' variables from a distribution. For discrete variables, I write this as

$$\sum_x p(x,y) \equiv p(y)$$

and for continuous x, I use integral notation

$$\int_x p(x,y) \equiv p(y)$$

which incorporates a compact notation for integrals; the notation for integration throughout generalises as

$$\int_{z_{a:b}} f(z_{a:b}) \equiv \int \cdots \int f(z_{a:b}) \; \mathrm{d}z_b \, \mathrm{d}z_{b-1} \, \cdots \, \mathrm{d}z_a.$$

When refering to the probability density function $p(\cdot)$, the function will generally factorise into components with a fixed analytical form, each corresponding to one of a set of well-known probability distributions. Density functions, along with some interesting properties, for the relevant distributions are given in appendix A.

Finally, angled brackets represent expectation, $\langle f(x) \rangle \equiv \mathbb{E}\left[f(x)\right] \equiv \int_x f(x)\, p(x)$.

CHAPTER 2

Statistics, Econometrics, and Machine Learning

"As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality."

—Albert Einstein, 1879-1955

2.1 Modelling Under Uncertainty

The field of statistics is fundamentally based on the study of data. Often, data contain randomness or uncertainty for any number of possible reasons, which makes statistics a natural home for the mathematics of uncertainty. Machine learning, generally speaking, is the study of methods to extract information from data in an automated fashion, and the association with statistics and an interest in dealing under uncertainty follows naturally. For a machine learning practitioner, working in the presence of uncertainty and evolving reason based on experience are fundamental concepts. The axioms of probability provide a mathematical framework and calculus for the expression of 'chance', and whilst some artificial intelligence practitioners argue that probability is incomplete or inappropriate for modelling real-life uncertainty (Cheeseman (1988) provides an interesting discussion), others disagree—notably Pearl (1988) supports probability as a mechanism for uncertain reasoning. The use of such probabilistic reasoning underpins this thesis.

By following the accepted rules of probability distributions, one may make inferences about unknown quantities of interest relating to the data we observe. The approach of Bayesian analysis can be thought of as a framework to quantify and update uncertainty, expressed numerically, based on experience in the 'real-world' in the form of data. The mechanism is to place a prior distribution over the variable to express the *a priori* belief, and to update the distribution to form the posterior based on what is observed by appealing to the axioms of probability to provide a calculus of uncertainty. In a data model, one may describe mathematically a process of generating data, and seek to show that the model matches that which we observe. A frequentist approach would be to seek estimates of quantities of interest in such a model—the model parameters—and through construction of a test statistic, seek evidence that the model so described 'matches' what we experience. By contrast, one may incorporate uncertainty about the assumed parameters of the model into the generating process, rather than find a point estimate for the "best" parameter value. This is the approach of Bayesian statistics, which seeks to quantify the uncertainty in parameter values rather than find 'estimators'. Normally, a Bayesian statistician would introduce a prior distribution over the uncertain parameters, itself characterised by a set of *hyper*-parameters.

Underpinning the dichotomy of statistical approaches is a fundamental discussion about the interpretation of probability. The frequentist church of probability can broadly be thought to consider an experiment as a single example in an infinite sequence of possible (independent) repetitions of the same experiment. In this scheme, a probability of a particular outcome simply represents the long-run rate (*frequency*) of such outcome and can only be considered relevant when the underlying experiment is well-defined, random, and repeatable. By contrast, a Bayesian naturally interprets probability as an abstract concept that describes a "degree of belief" in an outcome. This approach can be thought of as a "fuzzy" extension of propositional logic; Bayesian probability having an evidential interpretation, the Bayesian is happy to update their belief as their experience develops.

This thesis is far from a treatise on the merits of Bayesian inference; rather, I take a pragmatic approach to reasoning under uncertainty—naturally preferring Bayesian statistics—and aim to develop methods with broad applicability and demonstrable benefits to real-world problems.

Bayesian probabilistic inference. In particular, inference is based on repeated application of the definition of conditional probability,

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

which characterises the impact of knowledge of y on the variable x. Bayes' rule[1] trivially follows from this definition, as

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

in which the *posterior* $p(x|y)$ is found based on the *prior* $p(x)$, *likelihood* $p(y|x)$ and *marginal likelihood* or *model evidence* $p(y)$.

Conditional independence

A key concept in Bayesian statistics is conditional independence. The variable x is conditionally independent of y given z, written $x \perp\!\!\!\perp y \,|\, z$ if $p(x,y|z) \equiv p(x|z)\,p(y|z)$. From this we may see (by an application of the definition of conditional probability) that equivalently, $p(x|y,z) \equiv p(x|z)$.

[1] The definition of conditional probability is often confused with Bayes' rule: since the latter trivially follows from the former, I make no further distinction throughout the thesis.

Intuition about what is meant by conditional independence can be developed by characterising the idea that if x is conditionally independent of y given z, that once z is known, y tells us nothing more about x—and x similarly tells us nothing more about y.

2.1.1 Bayesian Generative Modelling

When talking about 'Bayesian' models, I refer to inference in a generative probabilistic model. Generative models are probabilistic frameworks for generating observable data, so named because they may be used to simulate examples of data similar to that one may observe. In machine learning, specifying a generative model is an intermediate step to forming a conditional density function, from which parameter updates are inferred from observed data.

The process of Bayesian modelling can be understood as follows:

1. Identify variables that are observed and variables that impact the process but are not observed (latent variables).
2. Specify a joint density function over all of the variables (observed and latent) that describes how the variables interact. The density function is likely to make assumptions about the conditional independence of the variables, and consists of the assumed analytical form of the distribution along with values for (hyper-)parameters.
3. Perform inference based on real observations. For a model with latent variables, this involves forming a conditional density function for the *a posteriori* belief—the distribution of those variables conditioned on the observations.
4. If required, update the value of the (hyper-)parameters of the model to improve the 'fit'. A common approach is to aim to maximise the marginal likelihood of the observed values by choice of the (hyper-)parameters.

For time series, the approach is to describe a process in terms of temporal relationships between observations that are made in a sequential manner.

The nature of the process of Bayesian generative modelling is such that, from the outset, the assumptions being made about the way the data are generated are set out in a clear and coherent manner. Whilst one may see such assumptions as limiting and a restriction of the method, another viewpoint is that any mathematical model must make assumptions, and that a fortunate property of the Bayesian modelling process is that those assumptions are made explicit at the outset.

The rest of this background chapter is set out as follows. The remainder of this section deals with further introduction to the terminology and concepts in Bayesian modelling. Section 2.2 introduces some widely-known models that are commonly used for time-series, and explains some of the corresponding properties. Section 2.3 introduces a slightly different topic which forms the basis of a major contribution of the thesis, known as 'cointegration' and taken from the field of econometrics. Cointegration specifically deals with the relationship between two or more time-series and therefore fails to fit into the earlier discussion since the method does not make assumptions about the generating process of the observed data. Finally, section 2.4 considers the problem of inference in Bayesian models and the methods that can be used to find the posterior and parameter estimates.

Graphical Models

In machine learning, graphical models are used as visual frameworks for data dependencies. Such depictions of the variables are created to help us develop intuition about how a model generates data, how variables interact, or how we may perform inference efficiently.

There are many different types of graphical model, with different properties and features, and different aims. For my purposes, graphical models are an interesting but non-fundamental way to gain understanding of the models featured in the thesis.

I represent conditional independence in a generative model with a *belief network*[2], a directed graph in which the nodes represent variables and directed edges represent relationships between variables that are not assumed to be independent. A belief network represents a factorisation of the model joint density function as $p(\mathbf{x}) \equiv \prod_i p(x_i \,|\, \mathrm{parents}(x_i))$ in which the incoming edges to node x_i denote the parents. This approach helps to visualise the dependencies of the model to gain an understanding of the relationships between the variables. For an overview of belief networks and other graphical models, see for example Lauritzen (1996) or Barber (2012).

Non-Parametric Modelling

This thesis is focussed exclusively on a paradigm known as *parametric modelling*, which contrasts with the alternative *non-parametric modelling*[3], a discipline featuring broad model classes such as the Gaussian process. Non-parametric models, generally speaking, can be understood as working to refine a distribution over an infinite space of functions, whereas parametric models require specific domain knowledge to inform the structure of a model one may choose to work with. Whilst one may therefore consider non-parametric models to be more generic and less restrictive than parametric models, there are specific reasons that one may prefer a parametric approach. For example, we may have clear domain knowledge that we wish to integrate into a model. More importantly, it is often useful to the practitioner to have knowledge of and work with the specific parameters of the functional form of a model, which non-parametric models are designed to obfuscate.

A good overview of Gaussian process modelling for time series is available from Roberts et al. (2012), which includes discussion of some key concepts discussed in this thesis.

[2] Also known as *Bayesian networks*, *causal networks*, *influence diagrams*, and *relevance diagrams*.

[3] This may be considered as something of a misnomer since any model is characterised by parameters—known as hyperparameters in the case of non-parametric models, which generally avoid parameters that explicitly inform the functional form of curves.

2.1.2 Bayesian Model Selection

It is easy to devise generative models of almost any form. Without a method for assessing the ‘quality’ of a model, one may select any number of spurious data representations. For this reason it is important to consider the problem of *model selection*—which is especially important when one seeks to fit Bayesian models without concrete insight into the underlying generating process.

For a Bayesian model $\mathcal{M}$, the data likelihood $p(\mathcal{D}|\mathcal{M})$ is commonly used as a measure of ‘goodness-of-fit’. The higher the data likelihood, the more likely that the model $\mathcal{M}$ generated the data $\mathcal{D}$.

Occam’s razor is the principle that one should select, between competing hypotheses, the one which is no more complicated than necessary. Bayesian inference tends to penalise overly-complex models, building in an Occam’s razor effect (Jefferys and Berger, 1991). When it is difficult or impossible to calculate the data likelihood for a particular model exactly, it is common to use approximations such as Laplace’s method, or the Bayes Information Criterion, which explicitly penalises model complexity.

There is a broad literature on model selection for Bayesian analysis. Fundamentally, one may rely on the Bayes factor to compare models, based on comparing the data likelihood for each model,

$$\frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}.$$

From this definition we can see that when the likelihood of model 1 is greater than that of model 2, the ratio has a value > 1, and conversely when model 2 shows greater likelihood the factor has value < 1. The magnitude of the value, and in particular how different it is from 1, gives a measure of how much better a model is than the other; Jeffreys (1961, Appendix B) tabulates critical values.

2.2 Time-Series Models

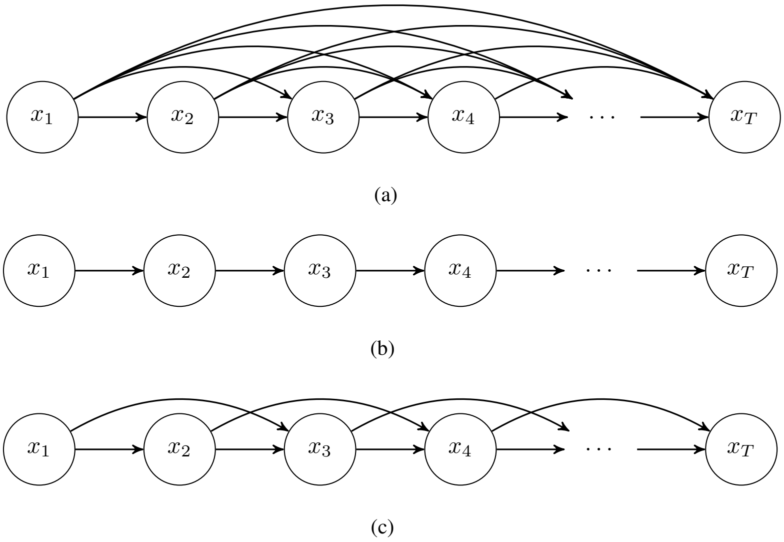
There are a great number of models that have been used for modelling sequential data. In this section, I set out some of the common concepts and generative models used for time-series and attempt to unify some disparate technologies and terminologies from econometrics, statistics, and machine learning.

2.2.1 Markov Chains and State-Space Models

A naïve approach to modelling time series would be to assume that each observation x_t is described by all available historical knowledge of the variable: that is, x_t a function in all of the preceding observations $x_{1:t-1}$. Unfortunately this can be troublesome since that function is difficult to define: a function which uses t data sources may have a number of parameters that depends on t. This provides the motivation for a Markov chain, which seeks to reduce the complexity.

Markov Chains

The idea of a Markov chain is to restrict the ‘memory’ of a process by assuming that all of the knowledge required from the previous observations is contained in a finite number of the most recent values.



(a)

(b)

(c)

Figure 2.1: Belief networks illustrating the conditional independence assumptions of the Markov chain.

An example of a sequence generated by a Markov chain is the Fibonnaci sequence: the well-known series in which each number is obtained by adding together the two preceding numbers. The process is a second-order Markov chain since no further knowledge of the previous values is needed once the two most recent values are known, in order to obtain the next. Whilst the Fibonnaci sequence is deterministic, Markov chains are generally used to describe the conditional independence of an ongoing probabilistic process.

Consider the sequence $x_{1:T}$. We can repeatedly apply the definition of conditional independence to see

$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{1:t-1})$$

in which the distribution of each x_t may depend on all the previous values. If the process were assumed to be an order-n Markov chain, we would write

$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{t-n-1:t-1})$$

which encodes the assumption that only n of the most-recent observations are relevant to the value of each x_t. In this formulation, the 'update' distribution $p(x_t | x_{t-n-1:t-1})$ is known as the *transition* distribution.

In figure 2.1(a), I show the belief network for the fully-connected conditional decomposition of the sequence $x_{1:T}$. Corresponding belief networks for a first and second order Markov chain are given in figures 2.1(b) and 2.1(c) respectively.

Latent Markov Models

The Markov chain is useful in that it allows us to describe a potentially complex process by dealing only with a fixed number of parameters that determine the transition distribution of the Markov chain. We can enrich the model by assuming that the Markov chain is *latent*—that is, unobserved. Normally, we would

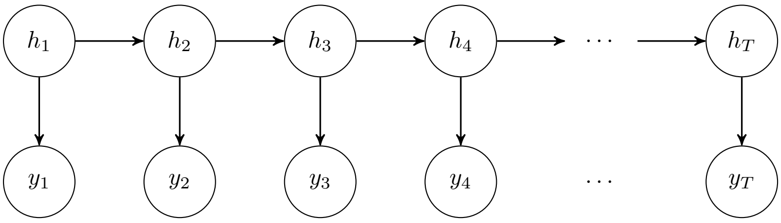


Figure 2.2: Belief network for a latent Markov model.

therefore assume that the observations we make were derived from the Markov chain but not the same as the variables in the Markov chain itself: perhaps instead they are observed with some noise due to the sensor used in making the observation, for example.

A latent Markov model (also known as a state-space model) assumes that the data are generated by an underlying Markov chain, but that the chain is not itself observed. The range of values that can be taken by the latent variable is known as the state space, and the latent value is commonly referred to as the prevailing state of the process. Since the state is never observed, one must make assumptions about the state space and infer the state at each point based on the observations that are made, through the lens of a noise model or transformation.

Normally, we assume the latent variable follows a first-order Markov chain, though it is possible to consider higher-order chains. In the event that the latent variables forming the latent Markov chain take discrete values, the system is known as a *hidden Markov model*. For continuous-valued variables, the system is known as a *dynamical model*. I will normally label the variables forming the latent chain as h, short for “hidden”, representing the fact that the values are not observed.

In general, the joint distribution describing a latent Markov model is given by

$$p(y_{1:T}, h_{1:T}) = \prod_{t=1}^{T} \underbrace{p(h_t|h_{t-1})}_{\text{transition}} \underbrace{p(y_t|h_t)}_{\text{emission}}$$

in which we see that the joint density function factorises into a product of transition terms from the latent Markov chain and *emission* terms that describe how the observations are based on the latent variables.

A belief network which visualises the conditional independence assumptions of a latent Markov model is shown in figure 2.2. For these models, I have so far made no assumptions about the dimension, distribution, or domain of each variable.

Inference in latent Markov models refers to application of Bayesian reasoning in order to update the belief about the distribution of the latent variables based on the observations that are made. Inference is well studied for this class of model, and “message passing” algorithms are commonly used. These algorithms are so called because they generally rely on passing “messages” along the edges of the belief network, which encode information collected from the nodes at the source of the message.

There are several common inference problems for latent Markov models. *Filtering* corresponds to calculating the posterior for the latent variable h_t based on preceding observations $y_{1:t}$. *Smoothing* builds in hindsight to the posterior and seeks the distribution of each h_t based on observations $y_{1:T}$, where in general $T > t$. Calculating the marginal *likelihood* $p(y_{1:T})$, making a *prediction* for y_{T+1}, and *viterbi* (finding the most-likely joint state for $h_{1:T}$) are other common inference problems. In the remainder of the thesis, I focus on the key problems of filtering and smoothing (I will also refer to likelihood—the value is usually available from the filtering problem).

Filtering corresponds to obtaining each $p(h_t|y_{1:T})$. The common approach is to use a message-passing characterisation for the inference problem for "forward inference". I write $p(h_t, y_{1:t}) \equiv \alpha(h_t)$ and the recursion is written[4]

$$\alpha(h_t) = \int_{h_{t-1}} p(h_t, h_{t-1}, y_{1:t}) = p(y_t|h_t) \int_{h_{t-1}} p(h_t|h_{t-1})\, \alpha(h_{t-1}) \tag{2.1}$$

where the posterior for h_t can be easily obtained by normalisation—which corresponds to dividing by the data likelihood $p(y_{1:t})$.

For the problem of smoothing, there are two common approaches using a message passing scheme. Both rely on a "backwards" recursion algorithm. The first involves the calculation of a reverse message for each t, and finally combining these with the filtered posteriors $\alpha(h_t)$ to form the smoothed posterior $p(h_t|y_{1:T})$. For this "forward-backward" algorithm, I write $p(y_{t+1:T}|h_t) \equiv \beta(h_t)$ which is calculated as

$$\beta(h_t) = \int_{h_{t+1}} p(h_{t+1}, y_{t+1:T}|h_t) = \int_{h_{t+1}} p(y_{t+1}|h_{t+1})\, p(h_{t+1}|h_t)\, \beta(h_{t+1}) \tag{2.2}$$

from which the smoothed posterior $\gamma(h_t) \equiv p(h_t|y_{1:T})$ is calculated since

$$\gamma(h_t) \propto p(h_t, y_{1:T}) = p(y_{t+1:T}|h_t)\, p(h_t, y_{1:t}) = \alpha(h_t)\, \beta(h_t)\,. \tag{2.3}$$

Note that each $\beta(h_t)$ is not a distribution in h_t, unlike the $\alpha(h_t)$ messages which are given as distributions in h_t (up to a scaling factor of the likelihood). Whilst this does not prevent us from implementing the β recursion, it can cause computational issues for some important types of model.

The second message-passing approach to finding the smoothed posterior—known as 'correction' smoothing since it 'corrects' each $p(h_t|y_{1:t})$ into the required $p(h_t|y_{1:T})$—calculates the posterior directly using the filtering messages, since

$$\gamma(h_t) = \int_{h_{t+1}} p(h_t, h_{t+1}|y_{1:T}) = \int_{h_{t+1}} p(h_t|h_{t+1}, y_{1:t})\, \gamma(h_{t+1}) \tag{2.4}$$

where the *dynamics reversal* term is calculated as

$$p(h_t|h_{t+1}, y_{1:t}) = \frac{p(h_{t+1}|h_t)\, \alpha(h_t)}{p(h_{t+1}|y_{1:t})\, p(y_{1:t})} \propto p(h_{t+1}|h_t)\, \alpha(h_t)\,. \tag{2.5}$$

The initial step is easily calculated since $\gamma(h_T) \propto \alpha(x_T)$.

Likelihood is used for model selection, and gives the marginal likelihood that the observations were generated by the model $p(y_{1:T})$. Normally this can be calculated easily from normalising the α messages, or iteratively using

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

which is easily available from the filtering recursion equation (2.1).

[4] These recursions hold more generally on replacing integration over any discrete variables by summation.

Linear-Dynamical Systems

For discrete variables the latent Markov model is known as the hidden Markov model, and can be easily specified by providing the fixed values for the transition and emission distributions: the probability of changing between the states (from a fixed discrete set) is specified in the conditional probability table forming the transition probabilities, and the probability of emitting each observation is again specified as a conditional probability table for the emission distribution. These tables are normally considered as matrices (*stochastic* matrices have columns which sum to 1), and if the transition and emission distributions are invariant through time the chain is said to be *homogeneous*.

When the variables take continuous values, the transition and emission distributions can be specified as continuous conditional density functions. However, it can be difficult to find a structure for the transition and emission distributions that allows inference to be performed exactly. I return to the problem of tractable inference, and methods taken to solving the inference problem, in section 2.4. First, we discuss a class of model which has been very widely studied because it is widely applicable and allows the inference to be performed in closed analytical form. This is the linear-dynamical system (LDS)—also known under a variety of different names including a *linear-Gaussian* or *conditionally Gaussian state space model*, and *Kalman filter/smoother*.

The linear-dynamical system has been independently derived a number of times and has been applied to diverse problems including finance, bioinformatics, signal processing, and many more.

The model—as the various names suggest—relies on the Gaussian distribution for both the transition and emission distributions in a latent Markov model, where the mean of each Gaussian is based linearly on the conditionally-dependent 'parent' variable. For the transition, the mean of h_t is a linear function of the previous latent value h_{t-1}; correspondingly, the emitted value y_t comes from a Gaussian distribution where the mean is a linear form of the latent state h_t.

The transition distribution (assuming a homogeneous model with time-invariant parameters) is therefore given as

$$p(\mathbf{h}_t|\mathbf{h}_{t-1}) = \mathcal{N}(\mathbf{h}_t|\mathbf{A}\mathbf{h}_{t-1}, \mathbf{Q}) \tag{2.6}$$

and the corresponding emission distribution is

$$p(\mathbf{y}_t|\mathbf{h}_t) = \mathcal{N}(\mathbf{y}_t|\mathbf{B}\mathbf{h}_t, \mathbf{R}) \,. \tag{2.7}$$

In the algorithms I give below for the filtering and smoothing recursions, I allow the mean of the transition and emission distributions to be augmented with additive constants, given as $\bar{\mathbf{h}}$ and $\bar{\mathbf{y}}$ respectively. For brevity, in the following derivation these means are not included.

The inference can be performed exactly because the family of Gaussian distributions is closed under Bayesian manipulation of linear combinations, and the standard approach to filtering in this model bears the name of Kalman (1960). The results relating to Bayesian manipulation of linearly-Gaussian variables that allow exact inference to be performed are given in summary in appendix B, and referred to in the derivations below.

Algorithm 2.1 LDS standard Kalman filter

```
function LDSFORWARD($\mathbf{f}$, $\mathbf{F}$, $\mathbf{y}$)
    $\boldsymbol{\mu}_\mathbf{h} \leftarrow \mathbf{A}\mathbf{f} + \bar{\mathbf{h}}$, $\boldsymbol{\mu}_\mathbf{y} \leftarrow \mathbf{B}\boldsymbol{\mu}_\mathbf{h} + \bar{\mathbf{y}}$          ▷ Mean of $p(\mathbf{h}_t, \mathbf{y}_t | \mathbf{y}_{1:t-1})$
    $\boldsymbol{\Sigma}_\mathbf{h} \leftarrow \mathbf{A}\mathbf{F}\mathbf{A}^\top + \mathbf{Q}$, $\boldsymbol{\Sigma}_\mathbf{y} \leftarrow \mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top + \mathbf{R}$          ▷ Marginal covariances
    $\mathbf{f}' \leftarrow \boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}\right) + \boldsymbol{\mu}_\mathbf{h}$, $\mathbf{F}' \leftarrow \boldsymbol{\Sigma}_\mathbf{h} - \boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}^\top$          ▷ Conditioning
    return $\mathbf{f}'$, $\mathbf{F}'$
end function
```

Filtering for the linear-dynamical system is based on the Kalman filtering pass (Kalman, 1960), and inference can be computed in closed form. We write equation (2.1) as

$$\alpha(\mathbf{h}_t) = \mathcal{N}(\mathbf{y}_t | \mathbf{B}\mathbf{h}_t, \mathbf{R}) \int_{\mathbf{h}_{t-1}} \mathcal{N}(\mathbf{h}_t | \mathbf{A}\mathbf{h}_{t-1}, \mathbf{Q})\, \alpha(\mathbf{h}_{t-1})$$

and it follows from this recursion that the Gaussian distribution is a conjugate form for α. To see this, assume $\alpha(\mathbf{h}_{t-1}) = \mathcal{N}(\mathbf{h}_{t-1} | \mathbf{f}_{t-1}, \mathbf{F}_{t-1})\, p(\mathbf{y}_{1:t-1})$, and by using Corollary B.3

$$\begin{aligned}
\alpha(\mathbf{h}_t) &= p(\mathbf{y}_{1:t-1})\, \mathcal{N}(\mathbf{y}_t | \mathbf{B}\mathbf{h}_t, \mathbf{R}) \int_{\mathbf{h}_{t-1}} \mathcal{N}(\mathbf{h}_t | \mathbf{A}\mathbf{h}_{t-1}, \mathbf{Q})\, \mathcal{N}(\mathbf{h}_{t-1} | \mathbf{f}_{t-1}, \mathbf{F}_{t-1}) \\
&= p(\mathbf{y}_{1:t-1})\, \mathcal{N}(\mathbf{y}_t | \mathbf{B}\mathbf{h}_t, \mathbf{R})\, \mathcal{N}\big(\mathbf{h}_t \big| \mathbf{A}\mathbf{f}_{t-1}, \mathbf{A}\mathbf{F}_{t-1}\mathbf{A}^\top + \mathbf{Q}\big)
\end{aligned}$$

To make notation easier, set $\boldsymbol{\Sigma}_\mathbf{h} = \mathbf{A}\mathbf{F}_{t-1}\mathbf{A}^\top + \mathbf{Q}$. Then we use Gaussian conditioning Corollary B.6,

$$\begin{aligned}
&\mathcal{N}(\mathbf{y}_t | \mathbf{B}\mathbf{h}_t, \mathbf{R})\, \mathcal{N}(\mathbf{h}_t | \mathbf{A}\mathbf{f}_{t-1}, \boldsymbol{\Sigma}_\mathbf{h}) \\
&\quad = \underbrace{\mathcal{N}\Big(\mathbf{h}_t \Big| \mathbf{M}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right) + \mathbf{A}\mathbf{f}_{t-1}, \boldsymbol{\Sigma}_\mathbf{h} - \mathbf{M}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}^\top\Big)}_{p(\mathbf{h}_t | \mathbf{y}_{1:t})} \underbrace{\mathcal{N}\big(\mathbf{y}_t \big| \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top + \mathbf{R}\big)}_{p(\mathbf{y}_t | \mathbf{y}_{1:t-1})}
\end{aligned} \tag{2.8}$$

where $\mathbf{M} = \boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top\left(\mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top + \mathbf{R}\right)^{-1}$. By also setting $\boldsymbol{\Sigma}_\mathbf{y} = \mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top + \mathbf{R}$, we have $\alpha(\mathbf{h}_t) = \mathcal{N}(\mathbf{h}_t | \mathbf{f}_t, \mathbf{F}_t)\, p(\mathbf{y}_{1:t})$ where we have the filtering recursion equations

$$\mathbf{f}_t = \boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right) + \mathbf{A}\mathbf{f}_{t-1} \tag{2.9}$$

$$\mathbf{F}_t = \boldsymbol{\Sigma}_\mathbf{h} - \boldsymbol{\Sigma}_\mathbf{h}\mathbf{B}^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{B}\boldsymbol{\Sigma}_\mathbf{h}^\top. \tag{2.10}$$

The update is shown as algorithm 2.1.

Likelihood. The data likelihood can be easily calculated since

$$p(\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{y}_t | \mathbf{y}_{1:t-1})$$

where each factor is given from equation (2.8) as

$$p(\mathbf{y}_t | \mathbf{y}_{1:t-1}) = \mathcal{N}(\mathbf{y}_t | \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \boldsymbol{\Sigma}_\mathbf{y})\,.$$

Smoothing. The two approaches to smoothing set out in equations (2.2) and (2.4) are both applicable to the linear-dynamical system. For the β smoother of equation (2.2), since as noted above the messages are not distributions in the latent variable $\mathbf{h}_t$, the components cannot be characterised as distributions with

Algorithm 2.2 LDS canonical backward update

```
function $\textsc{LDSCanonicalBackward}(k, \mathbf{p}, \mathbf{P}, \mathbf{y})$
    $\mathbf{M} \leftarrow \mathbf{B}^\top \mathbf{R}^{-1}\mathbf{B} + \mathbf{Q}^{-1} + \mathbf{P}, \mathbf{b} \leftarrow \mathbf{B}^\top \mathbf{R}^{-1}\left(\mathbf{y} - \bar{\mathbf{y}}\right) + \mathbf{Q}^{-1}\bar{\mathbf{h}} + \mathbf{p}$    ▷ Complete square
    $\mathbf{P}' \leftarrow \mathbf{A}^\top \left[\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{M}^{-1}\mathbf{Q}^{-1}\right]\mathbf{A}$    ▷ Quadratic component
    $\mathbf{p}' \leftarrow \mathbf{A}^\top \mathbf{Q}^{-1}\left[\mathbf{M}^{-1}\mathbf{b} - \bar{\mathbf{h}}\right]$    ▷ Linear component
    $k' \leftarrow \frac{k}{\sqrt{|2\pi\mathbf{R}||\mathbf{QM}|}} \exp -\frac{1}{2}\left[\left(\mathbf{y} - \bar{\mathbf{y}}\right)^\top \mathbf{R}^{-1}\left(\mathbf{y} - \bar{\mathbf{y}}\right) + \bar{\mathbf{h}}^\top \mathbf{Q}^{-1}\bar{\mathbf{h}} - \mathbf{b}^\top \mathbf{M}^{-1}\mathbf{b}\right]$
    return $k', \mathbf{p}', \mathbf{P}'$
end function
```

Algorithm 2.3 LDS moment-canonical combination

```
function $\textsc{LDSMomentCanonicalCombine}(w, \mathbf{f}, \mathbf{F}, k, \mathbf{p}, \mathbf{P})$
    $\mathbf{G} \leftarrow \left[\mathbf{F}^{-1} + \mathbf{P}\right]^{-1}$
    $\mathbf{g} \leftarrow \mathbf{G}\left[\mathbf{F}^{-1}\mathbf{f} + \mathbf{p}\right]$
    $w' \leftarrow \frac{wk}{\sqrt{|\mathbf{FM}'|}} \exp -\frac{1}{2}\left\{\mathbf{f}^\top \mathbf{F}^{-1}\mathbf{f} - \left[\mathbf{F}^{-1}\mathbf{f} + \mathbf{p}\right]^\top \mathbf{M}'^{-1}\left[\mathbf{F}^{-1}\mathbf{f} + \mathbf{p}\right]\right\}$
    return $w', \mathbf{g}, \mathbf{G}$
end function
```

recursions for sufficient statistics (moments) as for the Kalman filtering routine given above. However, the messages can be considered in the form of quadratic-exponential components, and characterised with sufficient coefficients in canonical form. This approach is similar to the *information parameterisation* set out by Cappé et al. (2005)—the "Backward Information Recursion" corresponds to the derivations for the backwards recursion given in this section.

In the case of a linear dynamical system, we assume the β messages form squared-exponential messages in canonical form given by $\beta(\mathbf{h}_{t+1}) = k_{t+1} \exp -\frac{1}{2}\left(\mathbf{h}_{t+1}^\top \mathbf{P}_{t+1}\mathbf{h}_{t+1} - 2\mathbf{h}_{t+1}^\top \mathbf{p}_{t+1}\right)$. Each message is then parametrised by the coefficients $\mathbf{P}_{t+1}$ and $\mathbf{p}_{t+1}$ and the 'weight' k_{t+1}. Whilst the weight k_{t+1} may not in general be of interest since the combination of the α and β messages shown in equation (2.3) provides the smoothed posterior of the latent variable up to a normalisation constant, I retain the term here since it will be required for models of interest in chapter 3.

To derive a recursion for these messages I write equation (2.2) as

$$\begin{aligned}\beta(\mathbf{h}_t) = \int_{\mathbf{h}_{t+1}} & \frac{1}{\sqrt{|2\pi\mathbf{R}|}} \exp -\tfrac{1}{2}\left(\mathbf{y}_{t+1} - \mathbf{B}\mathbf{h}_{t+1} - \bar{\mathbf{y}}\right)^\top \mathbf{R}^{-1}\left(\mathbf{y}_{t+1} - \mathbf{B}\mathbf{h}_{t+1} - \bar{\mathbf{y}}\right) \\ & \times \frac{1}{\sqrt{|2\pi\mathbf{Q}|}} \exp -\tfrac{1}{2}\left(\mathbf{h}_{t+1} - \mathbf{A}\mathbf{h}_t - \bar{\mathbf{h}}\right)^\top \mathbf{Q}^{-1}\left(\mathbf{h}_{t+1} - \mathbf{A}\mathbf{h}_t - \bar{\mathbf{h}}\right) \\ & \times k_{t+1} \exp -\tfrac{1}{2}\left(\mathbf{h}_{t+1}^\top \mathbf{P}_{t+1}\mathbf{h}_{t+1} - 2\mathbf{h}_{t+1}^\top \mathbf{p}_{t+1}\right).\end{aligned}$$

By setting $\mathbf{M} = \mathbf{B}^\top \mathbf{R}^{-1}\mathbf{B} + \mathbf{Q}^{-1} + \mathbf{P}_{t+1}$ and $\mathbf{b} = \mathbf{B}^\top \mathbf{R}^{-1}\left(\mathbf{y}_{t+1} - \bar{\mathbf{y}}\right) + \mathbf{Q}^{-1}\bar{\mathbf{h}} + \mathbf{p}_{t+1}$ we can complete the square[5] and write

$$\beta(\mathbf{h}_t) = \frac{k_{t+1}\sqrt{|2\pi\mathbf{M}^{-1}|}}{\sqrt{|2\pi\mathbf{R}|\,|2\pi\mathbf{Q}|}} \exp -\tfrac{1}{2}\left(\mathbf{y}_{t+1}-\bar{\mathbf{y}}\right)^\top \mathbf{R}^{-1}\left(\mathbf{y}_{t+1}-\bar{\mathbf{y}}\right)$$
$$\times \exp -\tfrac{1}{2}\left(\mathbf{A}\mathbf{h}_t+\bar{\mathbf{h}}\right)^\top \mathbf{Q}^{-1}\left(\mathbf{A}\mathbf{h}_t+\bar{\mathbf{h}}\right)$$
$$\times \exp \tfrac{1}{2}\left[\mathbf{b}+\mathbf{Q}^{-1}\mathbf{A}\mathbf{h}_t\right]^\top \mathbf{M}^{-1}\left[\mathbf{b}+\mathbf{Q}^{-1}\mathbf{A}\mathbf{h}_t\right]$$

from which I aim to write $\beta\left(\mathbf{h}_t\right) = k_t \exp -\frac{1}{2}\left(\mathbf{h}_t^\top \mathbf{P}_t \mathbf{h}_t - 2\mathbf{h}_t^\top \mathbf{p}_t\right)$. Then noting $\dim \mathbf{M} = \dim \mathbf{Q}$, we see

$$\begin{aligned}
\mathbf{P}_t =& \mathbf{A}^\top\left[\mathbf{Q}^{-1}-\mathbf{Q}^{-1}\mathbf{M}^{-1}\mathbf{Q}^{-1}\right]\mathbf{A}\\
\mathbf{p}_t =& \mathbf{A}^\top\mathbf{Q}^{-1}\left[\mathbf{M}^{-1}\mathbf{b}-\bar{\mathbf{h}}\right]\\
k_t =& \frac{k_{t+1}}{\sqrt{|2\pi\mathbf{R}|\,|\mathbf{Q}\mathbf{M}|}} \exp -\tfrac{1}{2}\left[\left(\mathbf{y}_{t+1}-\bar{\mathbf{y}}\right)^\top\mathbf{R}^{-1}\left(\mathbf{y}_{t+1}-\bar{\mathbf{y}}\right)+\bar{\mathbf{h}}^\top\mathbf{Q}^{-1}\bar{\mathbf{h}}\right] \times \exp \tfrac{1}{2}\mathbf{b}^\top\mathbf{M}^{-1}\mathbf{b}.
\end{aligned}$$

The complete update is shown in algorithm 2.2.

Once the backwards recursion has been completed in this canonical form, the α and β messages can be combined in accordance with equation (2.3) to yield the required smoothed posterior γ. To derive this, consider the filtered component $\mathcal{N}(\mathbf{h}_t|\mathbf{f},\mathbf{F})$ with an associated weight[6] w and a canonical component $k\exp -\frac{1}{2}\left(\mathbf{h}_t^\top\mathbf{P}\mathbf{h}_t - 2\mathbf{h}_t^\top\mathbf{p}\right)$. Then the posterior component found by multiplying the filtered Gaussian component by the β message in canonical form is given by

$$\frac{w}{\sqrt{|2\pi\mathbf{F}|}}\exp -\tfrac{1}{2}\left(\mathbf{h}_t-\mathbf{f}\right)^\top\mathbf{F}^{-1}\left(\mathbf{h}_t-\mathbf{f}\right)\times k\exp -\tfrac{1}{2}\left(\mathbf{h}_t^\top\mathbf{P}\mathbf{h}_t-2\mathbf{h}_t^\top\mathbf{p}\right)$$
$$= \frac{wk}{\sqrt{|2\pi\mathbf{F}|}}\exp -\tfrac{1}{2}\left\{\mathbf{h}_t^\top\left[\mathbf{F}^{-1}+\mathbf{P}\right]\mathbf{h}_t-2\mathbf{h}_t^\top\left[\mathbf{F}^{-1}\mathbf{f}+\mathbf{p}\right]+\mathbf{f}^\top\mathbf{F}^{-1}\mathbf{f}\right\}.$$

Set $\mathbf{M}' = \mathbf{F}^{-1}+\mathbf{P}$. Then the posterior component is given by

$$\frac{wk}{\sqrt{|\mathbf{F}\mathbf{M}'|}}\exp -\tfrac{1}{2}\left\{\mathbf{f}^\top\mathbf{F}^{-1}\mathbf{f}-\left[\mathbf{F}^{-1}\mathbf{f}+\mathbf{p}\right]^\top\mathbf{M}'^{-1}\left[\mathbf{F}^{-1}\mathbf{f}+\mathbf{p}\right]\right\}$$
$$\times\mathcal{N}\left(\mathbf{h}_t|\mathbf{M}'^{-1}\left[\mathbf{F}^{-1}\mathbf{f}+\mathbf{p}\right],\mathbf{M}'^{-1}\right)$$

and the moments of the resulting Gaussian component are given as

$$\begin{aligned}
\mathbf{G} &= \left[\mathbf{F}^{-1}+\mathbf{P}\right]^{-1}\\
\mathbf{g} &= \mathbf{G}\left[\mathbf{F}^{-1}\mathbf{f}+\mathbf{p}\right]
\end{aligned}$$

which completes the calculation of the smoothed posterior using the backward information recursion.

[5] Completing the square corresponds to replacing

$$\exp -\tfrac{1}{2}\left(\mathbf{x}^\top\mathbf{M}\mathbf{x}-2\mathbf{c}^\top\mathbf{x}\right)\equiv\exp -\tfrac{1}{2}\left[\left(\mathbf{x}-\mathbf{M}^{-1}\mathbf{c}\right)^\top\mathbf{M}\left(\mathbf{x}-\mathbf{M}^{-1}\mathbf{c}\right)-\mathbf{c}^\top\mathbf{M}^{-1}\mathbf{c}\right]$$

where for the purposes of this section I replace $\mathbf{x}\to\mathbf{h}_{t+1}$ and $\mathbf{c}\to\mathbf{b}+\mathbf{Q}^{-1}\mathbf{A}\mathbf{h}_t$.

[6] The filtering recursions for the linear-dynamical system yield messages with a single Gaussian component so the weight coefficient is not normally necessary; however, the weight component is included in this derivation for the case of a mixture of Gaussian distributions which is relevant to chapter 3.

Algorithm 2.4 LDS standard RTS correction update

```
function LDSBackward($\mathbf{g}$, $\mathbf{G}$, $\mathbf{f}$, $\mathbf{F}$)
    $\boldsymbol{\mu}_\mathbf{h} \leftarrow \mathbf{A}\mathbf{f} + \bar{\mathbf{h}}$, $\boldsymbol{\Sigma}_\mathbf{h} \leftarrow \mathbf{A}\mathbf{F}\mathbf{A}^\top + \mathbf{Q}$        ▷ Statistics of $p(\mathbf{h}_{t+1}|\mathbf{y}_{1:t})$
    $\overleftarrow{\mathbf{A}} \leftarrow \mathbf{F}\mathbf{A}^\top\boldsymbol{\Sigma}_\mathbf{h}^{-1}$, $\overleftarrow{\boldsymbol{\Sigma}} \leftarrow \mathbf{F} - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}$, $\overleftarrow{\boldsymbol{\mu}} \leftarrow \mathbf{f} - \overleftarrow{\mathbf{A}}\boldsymbol{\mu}_\mathbf{h}$        ▷ Reversal $p(\mathbf{h}_t|\mathbf{h}_{t+1}, \mathbf{y}_{1:t})$
    $\mathbf{g}' \leftarrow \overleftarrow{\mathbf{A}}\mathbf{g} + \overleftarrow{\boldsymbol{\mu}}$, $\mathbf{G}' \leftarrow \overleftarrow{\mathbf{A}}\mathbf{G}\overleftarrow{\mathbf{A}}^\top + \overleftarrow{\boldsymbol{\Sigma}}$        ▷ Backward propagation
    return $\mathbf{g}'$, $\mathbf{G}'$
end function
```

The direct γ approach is known for the linear-dynamical system as the RTS correction smoother (Rauch, Tung, and Striebel, 1965), and is known to be more numerically stable than the information filter (Verhaegen and Van Dooren, 2002). For this reason, I focus on the correction smoother in the rest of the thesis.

In the linear-Gaussian case, the 'dynamics reversal' term from equation (2.5) can be calculated exactly by properties of Gaussian variables under linear transformation, as given in appendix B. Again we use Gaussian conditioning Corollary B.6,

$$\begin{aligned} p(\mathbf{h}_t|\mathbf{h}_{t+1}, \mathbf{y}_{1:t})\, p(\mathbf{h}_{t+1}|\mathbf{y}_{1:t}) &= \mathcal{N}(\mathbf{h}_{t+1}|\mathbf{A}\mathbf{h}_t, \mathbf{Q})\,\mathcal{N}(\mathbf{h}_t|\mathbf{f}_t, \mathbf{F}_t) \\ &= \underbrace{\mathcal{N}\Big(\mathbf{h}_t\Big|\overleftarrow{\mathbf{A}}\,(\mathbf{h}_{t+1} - \mathbf{A}\mathbf{f}_t) + \mathbf{f}_t, \mathbf{F}_t - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}_t^\top\Big)}_{p(\mathbf{h}_t|\mathbf{h}_{t+1}, \mathbf{y}_{1:t})}\underbrace{\mathcal{N}\big(\mathbf{h}_{t+1}\big|\mathbf{A}\mathbf{f}_t, \mathbf{A}\mathbf{F}_t\mathbf{A}^\top + \mathbf{Q}\big)}_{p(\mathbf{h}_{t+1}|\mathbf{y}_{1:t})} \end{aligned}$$

where $\overleftarrow{\mathbf{A}} = \mathbf{F}_t\mathbf{A}^\top\left(\mathbf{A}\mathbf{F}_t\mathbf{A}^\top + \mathbf{Q}\right)^{-1}$. The first of the two resulting terms is the dynamics reversal, and this allows us to see that the Gaussian distribution is a conjugate form for the message $\gamma(\mathbf{h}_t) = \mathcal{N}(\mathbf{h}_t|\mathbf{g}_t, \mathbf{G}_t)$; appealing to equation (2.4) the backward propagation is given according to

$$\begin{aligned} \gamma(\mathbf{h}_t) &= \int_{\mathbf{h}_{t+1}} \mathcal{N}\Big(\mathbf{h}_t\Big|\overleftarrow{\mathbf{A}}\,(\mathbf{h}_{t+1} - \mathbf{A}\mathbf{f}_t) + \mathbf{f}_t, \mathbf{F}_t - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}_t^\top\Big)\,\mathcal{N}(\mathbf{h}_{t+1}|\mathbf{g}_{t+1}, \mathbf{G}_{t+1}) \\ &= \mathcal{N}\Big(\mathbf{h}_t\Big|\overleftarrow{\mathbf{A}}\,(\mathbf{g}_{t+1} - \mathbf{A}\mathbf{f}_t) + \mathbf{f}_t, \overleftarrow{\mathbf{A}}\mathbf{G}_{t+1}\overleftarrow{\mathbf{A}}^\top + \mathbf{F}_t - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}_t^\top\Big) \end{aligned}$$

and the updates for the backwards recursion are given as

$$\mathbf{g}_t = \overleftarrow{\mathbf{A}}\,(\mathbf{g}_{t+1} - \mathbf{A}\mathbf{f}_t) + \mathbf{f}_t \tag{2.11}$$

$$\mathbf{G}_t = \overleftarrow{\mathbf{A}}\mathbf{G}_{t+1}\overleftarrow{\mathbf{A}}^\top + \mathbf{F}_t - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}_t^\top. \tag{2.12}$$

The final recursive update is shown in algorithm 2.4.

Switching Linear Dynamical Systems

Whilst the linear-dynamical system is both easy to calculate and applicable to many real-world problems, the assumption of a fixed Gaussian distribution for each of the transition and emission distributions can be restrictive. A number of extensions to the model have been proposed, with varying complexity, and one of the best-known is the switching linear-dynamical system. For this model, instead of having a single distribution for each of the transition and emission distributions, there are instead a fixed number of each (all linear-Gaussian as with the simple linear-dynamical system) to choose from. The 'chosen' distribution is selected according to a second internal state. A switching linear dynamical system is then a heterogenous dynamical model in which the parameters defining the transition and emission distributions may change from time to time according to the prevailing internal state, as opposed to the time-invariant homogeneous model of the simple linear-dynamical system.

The switching model represents a variant latent Markov model, augmented with an additional discrete state variable s_t. The state is assumed to be selected at each time point from a finite set of S such states, $s_t \in S$, and a Markov chain with transition $p(s_t|s_{t-1})$ is assumed to govern the discrete state—a belief network is shown in figure 2.3. The prevailing state determines which parameters are applicable to the transition and emission distributions. These 'switching' latent Markov models are also known as *conditional Markov models*, *jump Markov models*, *switching models*, and *changepoint models* (I come to a thorough discussion of changepoint models in section 2.2.2). In the case of a switching linear-dynamical system, for each state s_t the transition and emission distributions are assumed to be linear-Gaussian.

Unfortunately, using the switching linear-dynamical system in real problems is usually prohibitively expensive. This is because choosing from any of the S states at each time t means there are a possible S^T different combinations. Because the mean of the transition distribution is always a linear form of the previous latent value, the result is that at the end of the chain, there are S^T possibilities for the Gaussian representing the latent variable.

This widely-known fact means that the complexity of exact inference scales exponentially in the length of the observed sequence, since at every observation time, each of the S latent states introduces an additional component. The computational complexity is a major problem and means inference is prohibitively expensive for all but trivially-short sequences. Much research has been done into how one may approximate the posterior in order to balance the accuracy and computation time of the algorithm for the particular case of the switching linear-dynamical system, and Frühwirth-Schnatter (2006) provides an overview of some common methods. As described therein, approaches include deterministic approximation by restricting the number of components in the posterior by ignoring 'unlikely' components or otherwise merging the individual distributions, or stochastic sampling estimation methods. I discuss approximation schemes for inference problems in section 2.4 and provide further references to approximation methods discussed in the literature.

2.2.2 Changepoint Models

The term 'changepoint' can be confusing because it is a general term that refers to any model which permits a change in the model parameters or structure of the generating process at one or more unknown point(s). This could take the form of a change of internal state, or a 'break' in the model dynamics, for example—a change-point may occur when there is a regime switch or other abrupt change in the structure of the generating process. By this broad definition a change of the state of a switching linear dynamical system, which corresponds to a change of moments in the governing distribution, could be thought of as a change-point.

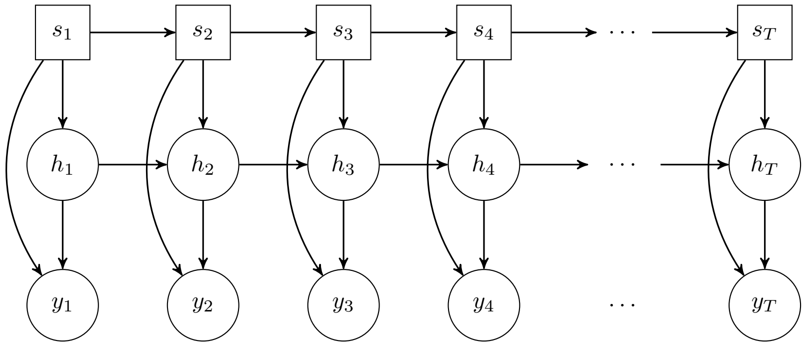


Figure 2.3: Belief network for a switching linear dynamical system. Rectangles are used for the discrete variables s_t in order to differentiate them from the continuous-valued variables. The discrete switch $s_t \in \{1, \ldots, S\}$ selects from a set of S distinct transition and emission distributions.

For my purposes, change-points are defined as the boundaries of a partition of the time-series data into disjoint, contiguous segments (described as product partition models by Barry and Hartigan (1992)). The process generating each segment is assumed to be independent of the (observations in) other segments. Another way to think of this is that conditioned on a partition $\mathcal{P}$, the data in a segment $\{a : b\} \subset \{1 : T\}$ are independent of other data,

$$p(y_{a:b}|\mathcal{P}) \equiv p(y_{a:b}|y_{1:a-1}, y_{b+1:T}, \mathcal{P})$$

Such a broad concept can apply to many different classes of model. The seminal works by Page (1954, 1955) introduced the notion of change-points and many works since have focused on detecting change-points of a sequence for broad classes of models.

Normally, the number and position of the changepoints is not known, and the modeller has the task of trying to ascertain when changepoints occur. In general, one may try to ‘sample’ different combinations by conditioning on a specific occurrence of changepoints, and comparing how likely it is that the data were generated with different such combinations. For some studies, there are limiting assumptions about the number of changepoints, such as only one or two. By contrast, a Bayesian approach to change-point analysis would more naturally aim to avoid these limiting assumptions, and attempt to place a distribution over the number and location of the changepoints and this is the approach taken in this thesis.

There are classic datasets for change-point problems. These include the coal-mine disaster data of Jarrett (1979) and well-log data first used by Ó Ruanaidh and Fitzgerald (1996) and made available by Fearnhead and Clifford (2003); both of these problems are well-studied in the literature. For the coal-mine disasters, the data are given as the number of days between disasters in coal mines between 1851 and 1962. The problem is to detect the introduction of the Coal Mines Regulations Act in 1887, and normally a Poisson model is assumed for the data. The well-log data are based on geophysical measurements of the nuclear magnetic response of the Earth in order to detect the structure of the rock strata below, normally assumed to have Gaussian noise.

To develop some intuition for the problem of change-point detection, I briefly consider the second of these problems using well-log data. The observed data, comprising sensor readings, are assumed to be constant for a layer of rock of any particular type, plus random noise. For different types of rock, the value is assumed to be different. The data are therefore usually modelled as Gaussian-distributed values with unknown mean, where the problem is translated to one of inferring the mean of the readings in each partition of the data, where the 'correct' partition is unknown. I give an example with the well-log data in section 3.5.1.

More recently, work has focused on on-line algorithms for the detection of change-points, and highly efficient algorithmic approaches are necessary. Both Fearnhead and Liu (2007) and Adams and MacKay (2007) offer on-line change-point detection algorithms, by characterising a change-point system with a variable capturing the run length—the time of the last change-point or the time since the last change-point—which enables exact calculation of the filtered posterior. These problems have broad applicability since such algorithms may permit prediction for broad classes of model. Fearnhead and Liu (2007) provide an approximation based on particle filtering to reduce the complexity of the algorithm, while the modular, broadly-structured work of Adams and MacKay (2007) is aims to deliver an algorithm with linear time-complexity.

Methods for detecting change-points in the literature are normally based on minimising a cost function over possible change-point locations. The cost function usually incorporates a cost component for a segment of observations that is evaluated for different combinations of change-points. As set out by Killick et al. (2011), the two most prevalent methods for multiple change-point detection are the approximate algorithm "binary segmentation" (Scott and Knott, 1974) and exact "segment neighbourhoods" (Auger and Lawrence, 1989).

Binary Segmentation is a divide-and-conquer, greedy strategy for identifying change-points. The cost function is first calculated for (all) possible locations of a single change-point and compared with the cost of no changepoint. If a change-point is chosen, the segment is then sub-divided, and the algorithm is repeated in each sub-segment until no more changepoints are detected. The greedy nature of the algorithm means that the method may fail to find an optimal setting, however.

Segment Neighbourhoods is an exact algorithm that relies on dynamic programming to evaluate the cost function for all possible segments.

Killick et al. (2011) propose a linear-time algorithm by extending an alternative method with a simple rule to prune irrelevant options from a recursive approach. Such methods based on evaluating a cost function generally make no statement about the expense of evaluating the cost function, however, which may be linear in the length of the sequence. What is more, such algorithms cannot be considered as Bayesian approaches since the result is a set of "detected" change-points, rather than a statement of belief in each point as a change-point (a posterior distribution for the occurrence of a change-point at each observation). Meanwhile, both Garnett et al. (2009) and Roberts et al. (2012) discuss changepoint analysis in the context of non-parametric Gaussian processes, with the former focussed on the problem of sequential prediction while marginalising the location of a fixed number of changepoints over a moving window. As discussed in section 2.1.1, this thesis focusses on parametric models.

2.2.3 Econometric Models

Time series analysis is a well-trodden path for applied economists, and the need to study the evolution of economic variables over time has given rise to the field of econometrics. Econometrics is the study of statistical models for estimating, testing, and evaluating economic relationships and theories. Important works include the forecasting of interest rates and inflation rates, for example, and forecasting is a significant concern in the literature. To this end, practitioners have devised time-series models ranging from the very simple to the very complex. Normally, an econometrician would apply empirical analysis to use data (observations) to seek or test an economic theory or relationship, and such econometric models are generally constructed with simple building blocks that rely heavily on linear regression models estimated with ordinary least squares.

Ordinary Least-Squares

A broad branch of econometric literature is focussed on analysis of regression models: a time series $y_{1:T}$ is estimated based on regressors $x_{i,1:T}$, $i = 1, \ldots, p$ according to

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{i,t} + \epsilon_t$$

in which the regression coefficients β_i are generally estimated by ordinary least-squares which I describe here. This estimation method is so-called because it is based on choosing the coefficients in order to minimise the sum of squared error terms $\sum \epsilon_t^2$. The solution is normally found by first rewriting the regression equation in vector notation,

$$y_t = \mathbf{x}_t^\top \boldsymbol{\beta} + \epsilon_t, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{x}_t = \begin{bmatrix} 1 \\ x_{1,t} \\ \vdots \\ x_{p,t} \end{bmatrix}$$

where I have augmented the data vector $\mathbf{x}_t$ to incorporate the constant term β_0 in the inner product. The sum of squared errors is then given as

$$\sum_{t=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} \left(y_t - \mathbf{x}_t^\top \boldsymbol{\beta}\right)^2 = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_T^\top \end{bmatrix}$$

and since this is a quadratic function the minimum can be found by differentiating with respect to the parameters $\boldsymbol{\beta}$ and setting the result to 0 to find the stationary point.

The solution is then given by

$$\boldsymbol{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2.13}$$

Under the assumptions of the "classical linear model" (set out by Wooldridge (2009)), the noise terms ϵ_t (also known as the *residuals*) are assumed to be independently and identically normally distributed,

$$\epsilon_t \sim \mathcal{N}\big(0, \sigma^2\big) .$$

Under the assumption of normally-distributed errors, the ordinary least-squares estimate for the linear regression coefficients corresponds to the maximum likelihood estimator. I do not derive the result here, but it can be shown simply by taking the log of the likelihood function and maximising the resulting analytical form—a similar argument is used in chapter 5.

OLS is used as an estimation scheme in a wide variety of models for time-series. In the following sections, I describe some of the most popular models set out in econometrics literature and place these models in a descriptive framework with simple belief networks, before moving on to discuss some of properties of estimation using OLS.

Moving Average Processes

One of the most simple models described in econometrics texts (see for example Hamilton (1994)) is the moving average process, where each term in the sequence is specified in terms of random noise variables ϵ_t. The most simple, first-order model is written as

$$x_t = \alpha + \epsilon_t + \varphi\epsilon_{t-1} \tag{2.14}$$

with ϵ_t generated from an independent and identically distributed process with mean 0 and constant variance σ^2.

The name arises since each value in the process is assumed to be deterministically calculated as a weighted average of the two noise terms ϵ_t and ϵ_{t-1}. The concept is easily extended to a q-order model with coefficients φ_j by writing

$$x_t = \alpha + \epsilon_t + \sum_{j=1}^{q} \varphi_j \epsilon_{t-j}$$

and belief networks for the first order and general order moving average processes are given in figure 2.4, based on stacking the observations as $\hat{\mathbf{x}}_t$ for the general-order process.

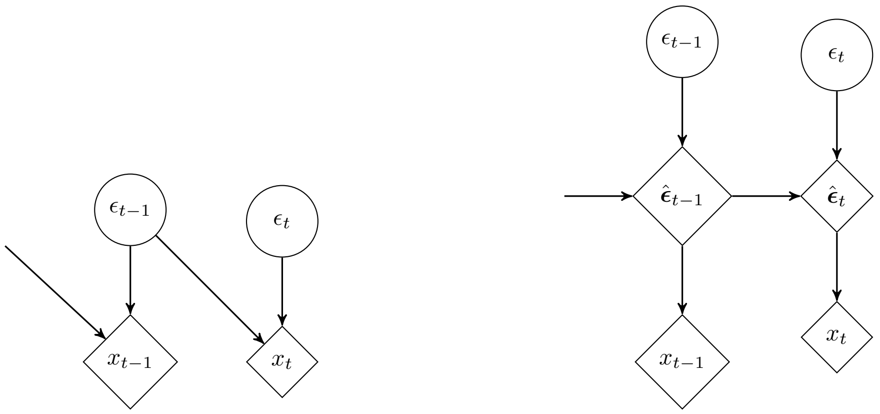
In econometrics literature, it is common to denote an order-q moving average process as MA(q).

Autoregressive Processes

One of the most widely-used and potentially more easily-applicable time-series models is the autoregressive (AR) process, in which the sequence is assumed to be generated using linear regression in which the regressors are simply lagged values of the dependent variable itself. In the most simple first-order case, the model therefore assumes that each observed value x_t is deterministically dependent on the previous value x_{t-1}, subject to the addition of the random noise term ϵ_t.

Such a first-order autoregressive model is therefore written as

$$x_t = \alpha + \varrho x_{t-1} + \epsilon_t$$



(a) First order MA

(b) General MA

Figure 2.4: Belief networks for MA processes. Continuous-valued stochastic variables are shown in circles, deterministic variables with diamonds.

with ϵ_t generated from an independent and identically distributed process with mean 0 and constant variance σ^2.

The general order autoregressive model is given as

$$x_t = \alpha + \sum_{i=1}^{p} \varrho_i x_{t-i} + \epsilon_t$$

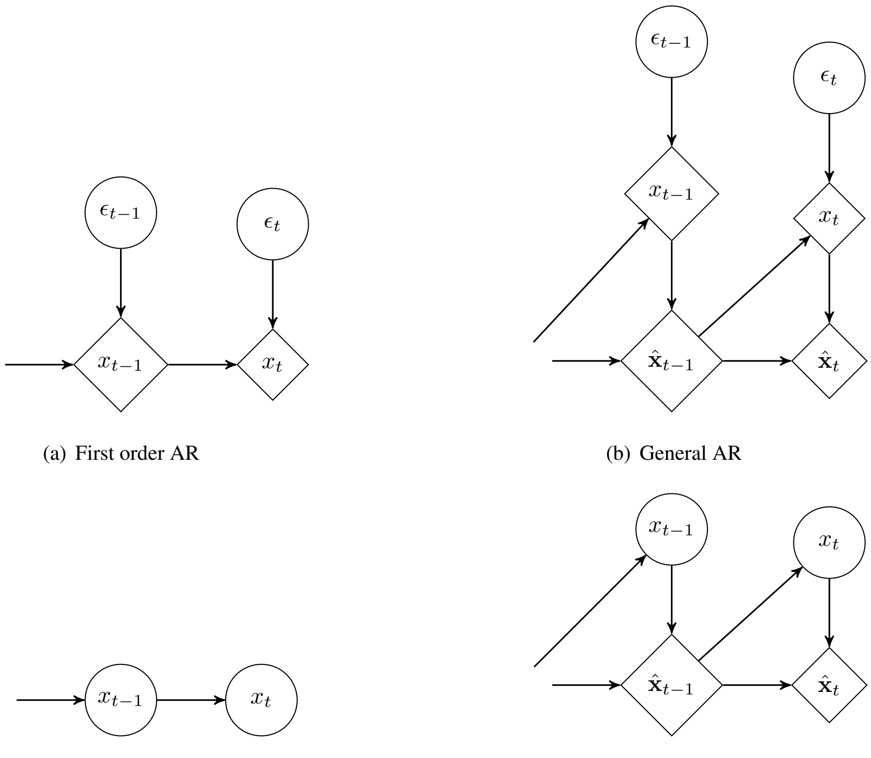
from which it is straightforward to see that the finite-order autoregressive process describes a Markov process. To illustrate, I have plotted belief networks for the processes in figure 2.5. The first-order model is shown as figure 2.5(a) and as a Markov chain in figure 2.5(c) after integrating the noise variables ϵ_t. By stacking the observations as $\hat{\mathbf{x}}_t$, the belief network for a general autoregressive process is shown in figures 2.5(b) and 2.5(d).

In econometrics literature, it is common to denote an order-p autoregressive model as AR(p).

AR and MA The first-order autoregressive model can be written in the form of a moving average process by repeatedly expanding the recursion,

$$\begin{aligned}
x_t &= \alpha + \varrho x_{t-1} + \epsilon_t \\
&= \alpha + \varrho \left(\alpha + \varrho x_{t-2} + \epsilon_{t-1}\right) + \epsilon_t \\
&= \left(\alpha + \epsilon_t\right) + \varrho \left(\alpha + \epsilon_{t-1}\right) + \varrho^2 \left(\alpha + \epsilon_{t-2}\right) + \ldots \\
&= \frac{\alpha}{1-\varrho} + \epsilon_t + \sum_{i=1}^{\infty} \varrho^i \epsilon_{t-i}
\end{aligned}$$

from which we can gain some intuition about the stability of the process when $|\varrho| < 1$.



(a) First order AR

(b) General AR

(c) First order AR, integrated out noise term

(d) General AR, integrated out noise variable

Figure 2.5: Belief networks for AR processes. Continuous-valued stochastic variables are shown in circles, deterministic variables with diamonds.

Autoregressive-Moving Average Processes

It is also common to combine the building blocks of the AR and MA processes to create a so-called ARMA process

$$x_t = \alpha + \sum_{i=1}^{p} \varrho_i x_{t-i} + \epsilon_t + \sum_{j=1}^{q} \varphi_j \epsilon_{t-j}$$

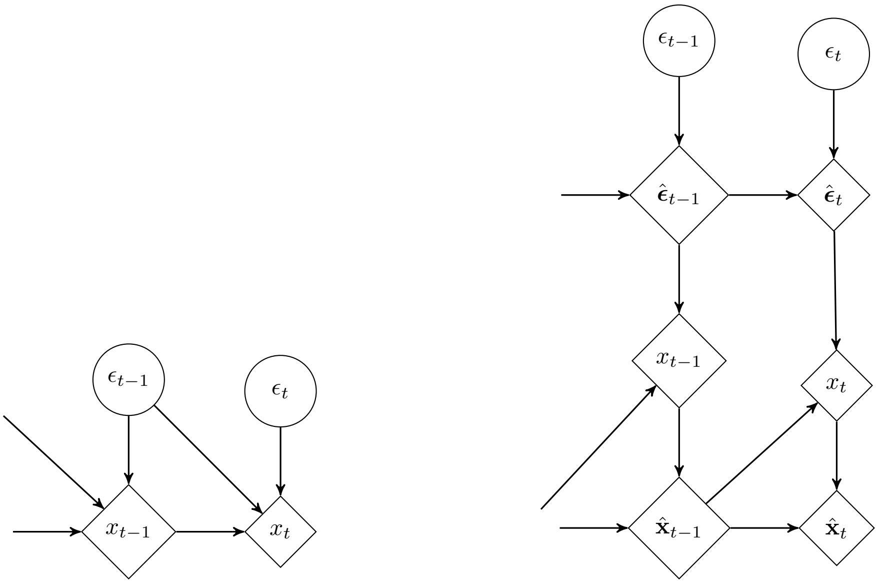
for which belief networks for first-order models and general models based on stacking the variables are given in figure 2.6. In the literature, such a model is referred to as ARMA(p, q); a thorough treatment of these models is presented by Hamilton (1994).

Conditional Heteroskedasticity models

Barber (2012) provides an insightful treatment of conditional heteroskedasticity models, which inspires the overview I give here to complete the overview of standard econometric time-series models.

Heteroksedastic models are designed to deal with sequences that exhibit changes in variance over the life of the process. This is important in many application contexts of interest—in finance, for example, volatility is known to increase in times of economic uncertainty.

The approach of conditional heteroskedasticity models is to consider the conditional variance of the process over time, and model the realised conditional variance as a process itself. The term ‘conditional variance’ refers to a model for the variance of the process conditioned on previous observations.



(a) First order ARMA

(b) General ARMA

Figure 2.6: Belief networks for ARMA processes. Continuous-valued stochastic variables are shown in circles, deterministic variables with diamonds.

In general, the approach is to consider some model for a process

$$x_t = f(x_{1:t-1}) + \epsilon_t$$

where f is some well-defined econometric model, normally a sum of autoregressive terms. The interest is then in the noise process ϵ_t. Under the simple AR/MA processes described above, the noise is assumed to have constant variance σ^2, whereas conditional heteroskedasticity models permit time-varying noise σ_t^2, where Var$(\epsilon_t) = \sigma_t^2$.

The variance process is modelled based on the realised squared-errors ϵ_τ^2 from earlier observations x_τ, $\epsilon_\tau = x_\tau - f(x_{1:\tau-1})$ for $\tau < t$.

ARCH. Developed by Engle (1982), Autoregressive Conditional Heteroskedasticity models use an autoregressive process to model an estimate of the variance process,

$$\begin{aligned}\sigma_t^2 &= \alpha + \sum_{i=1}^{p} \varrho_i \epsilon_{t-i}^2 \\ &= \alpha + \sum_{i=1}^{p} \varrho_i \left(x_{t-i} - f(x_{1:t-i-1})\right)^2 .\end{aligned}$$

Engle developed a lagrange multiplier test for the presence (and lag length) of ARCH errors.

GARCH. Bollerslev (1986) generalised the ARCH model by modelling the conditional variance (squared errors) using a mixed ARMA model in place of the more simple autoregressive model used by ARCH.

This is written as

$$\begin{aligned}\sigma_t^2 &= \alpha + \sum_{i=1}^{p} \varrho_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \varphi_j \sigma_{t-j}^2 \\ &= \alpha + \sum_{i=1}^{p} \varrho_i \left(x_{t-i} - f(x_{1:t-i-1})\right)^2 + \sum_{j=1}^{q} \varphi_j \sigma_{t-j}^2.\end{aligned}$$

Stationary Processes

The literature on econometric models can be confusing on the point of *stationary processes*, or the property of *stationarity*. Strictly speaking, a stochastic process is stationary if the values $x_{\mathcal{T}}$ for any subsequence of indices $\mathcal{T} \subset \mathbb{N}$ has the same joint distribution after shifting the subsequence in time, $p(x_{\mathcal{T}}) \equiv p(x_{\mathcal{T}+\tau})$, $\tau \geq 1$. Such a definition refers to an infinite process, whilst of course in any kind of analysis, one must work with a finite sample of observations that I index $1, \ldots, T$. Stationarity so defined is a property of the underlying stochastic process, and it can be very difficult to detect the property based on the finite sample we are able to observe (though non-stationary processes are often easy to spot).

Normally, a weaker statement of stationarity is sufficient, and for the rest of the thesis I focus on a more useful definition for stationarity.

For our purposes, a sample is stationary if it has finite and constant first and second moments, and the joint distribution of any pair of values depends only on the difference in index. That is, $\langle x_t \rangle = \mathbb{E}[x_t] \in \mathbb{R}$ and Var$(x_t) \in \mathbb{R}$ are constant, and $p(x_t, x_s)$ depends on $|t - s|$ not t or s. Normally, we will assume the mean $\langle x_t \rangle = 0$. A slightly weaker form of this definition, in which the property on pairs of values is replaced by considering the covariance instead of distribution, is often used and referred to as *covariance stationarity* or *weak stationarity*. An associated property of *weak dependence* is described by Wooldridge (2009).

For the first-order moving average model set out in equation (2.14), stationarity holds for any value of the coefficient φ. It is straightforward to show that the mean and variance of the process are constant in terms of α, σ^2 and φ. When looking at variables two or more periods apart, we see that the values are independent, and the distribution of neighbouring values is relatively simple.

Intuitively, stationarity describes the idea of "stability" in the process over time. The concept is a key concern to ensure relations that are found are meaningful.

Order of integration. From any series $x_{1:T}$ we can form a new series $x_{2:T}^1$ by taking the difference $x_t^1 = x_t - x_{t-1}$. Repeating this differencing process d times, a series is order d integrated, written $I(d)$, if the series x_t^d formed from repeatedly taking the difference d times yields a stationary series. A stationary series is said to be integrated of order 0, written $I(0)$.

Spurious Regression

When two variables x_t and y_t are found to be correlated, it may be the case that the variables are related due to a third variable z_t. If one were to control for the effect of z_t on each of x_t and y_t, the partial effect between x_t and y_t is removed; the relationship is then explained. In this sense the initial relation between x_t and y_t was a spurious relation.

However, there is an additional complication when dealing with series x_t and y_t that are not stationary $I(0)$ series. It was shown by Granger and Newbold (1974) that, in a large proportion of simulations, a significant relationship between two series was detected even though those series were generated independently. Here, x_t and y_t are assumed to be independent random walks defined by

$$\begin{aligned} x_t &= x_{t-1} + \eta^x_t, \quad \eta^x_t \sim \mathcal{N}\big(0, \sigma^2_x\big) \\ y_t &= y_{t-1} + \eta^y_t, \quad \eta^y_t \sim \mathcal{N}\big(0, \sigma^2_y\big) \end{aligned}$$

and when estimating the regression

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

with ordinary least-squares a significant relation is wrongly detected in a high proportion of cases. The true value of the coefficient β_1 is 0, since the series are independent. However, in this case the residuals $\epsilon_t = y_t - \beta_0$ themselves form a random walk with mean $\langle \epsilon_t \rangle = \langle y_1 \rangle - \beta_0$; as I describe in the following section, this clearly violates the assumptions needed for unbiased and consistent estimation with OLS.

This is the problem of spurious regression, and to avoid working with a meaningless relation, the applied economist is generally trained to first ensure that the underlying series are stationary. To do this, a practitioner will generally take differences to reduce the order of integration until the processes are stable.

Properties of least-squares estimators

When considering estimators for parameters, one may pick a value for the parameter in any number of ways. When picking an estimator, there are properties of the choice of estimate that one may consider in order to determine how 'good' an estimator is expected to be.

Two primary concerns are *consistency* and *bias*. An estimator $\hat{\beta}$ for a parameter β is consistent if, as the sample size $T \to \infty$, $\hat{\beta} \to \beta$. That is, as more data are observed, the estimate tends towards its true value (in probability). Consistency is generally understood to be the most simple requirement of an estimator—our intuition may agree that if an estimate does not get to the right answer in the limit of infinite data, another choice could be better! The estimator $\hat{\beta}$ for a parameter β is unbiased if $\left\langle \hat{\beta} \right\rangle_{p(x_{1:T})} \equiv \mathbb{E}_{p(x_{1:T})}\left[\hat{\beta}\right] = \beta$, and the *bias* is the difference $\left\langle \hat{\beta} \right\rangle - \beta$.

For widely-used models, notably the autoregressive process, it is common to estimate the coefficients by application of ordinary least-squares (OLS) regression, and the values provided by OLS are known to have useful properties. I therefore return to the general form linear equation applicable to OLS,

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{i,t} + \epsilon_t$$

and consider estimation according to equation (2.13). Consistency and bias of the OLS estimates can be shown under specific assumptions.

Consistency. OLS estimates are known to be consistent when the data $x_{i,t}$ and y_t are stationary and the regressors $x_{i,t}$ are *contemporaneously exogenous*, which means

$$\langle \epsilon_t \rangle_{p(\epsilon_t | x_{1:p,t})} = 0.$$

Another characterisation of the latter condition is

$$\langle \epsilon_t \rangle = 0, \quad \mathrm{Cov}(x_{i,t}, \epsilon_t) = \langle x_{i,t} \epsilon_t \rangle = 0, \quad i = 1, \dots, p.$$

Bias. Unfortunately, the requirements on the underlying process for OLS to provide unbiased estimates are somewhat stronger. In particular, the contemporaneous exogeneity condition stated for consistency is insufficient. Instead, a stronger condition of exogeneity is required, in which the regressors are *strictly exogenous*, with the definition that

$$\langle \epsilon_t \rangle_{p(\epsilon_t | x_{1:p,1:T})} = 0.$$

The difference is that the strict condition requires the errors ϵ_t to be uncorrelated with the regressors $x_{i,s}$ for all s, compared with the contemporaneous condition that only considers $s = t$. Of course, when the mean $\langle \epsilon_t \rangle = 0$ and the errors are independent of the data $\epsilon_t \perp\!\!\!\perp x_{1:p,1:T}$, strict exogeneity automatically holds and the OLS estimates are unbiased.

Anything that causes the error at time t to be correlated with any of the explanatory values $x_{i,t}$ in any time period causes the strict exogeneity assumption to fail, and we can no longer rely on the estimator to be unbiased.

Discussion: Econometric models

This section has set out some of the most prevalent models used in the econometrics community. Broadly speaking, the modeller takes a modular approach and the applied economist generally appeals to intuition about the quantities of interest to inspire a choice of model, before estimating the parameters of such model and considering goodness-of-fit and/or test statistics to reach some conclusion about the efficacy of the representation. The econometric modelling frameworks set out above make strong assumptions about the underlying generating process of the data, notably time-invariant parameters. This has led to a large number of extension methods, including switching models that allow for changes in the value of the (regression) parameters at one or more unknown point(s), amongst other approaches.

Hamilton (1994) is a standard text for time-series methods and also provides a good starting point for the approach taken to regime-switching systems by econometricians. A readable introduction to GARCH, along with a little on regime-switching models as approached by econometrics literature is available from Alexander (2008). Finally, thorough treatment of the properties of OLS estimators is available from Wooldridge (2009).

2.2.4 Autoregressive Latent Variable Models

In section section 2.2.3 I discussed a broad class of model according to a linear equation based on previous values of a sequence, known as autoregressive models. Such a model can be described by the equation

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \epsilon_t$$

from which we see that the model is parameterised by coefficients β_i. By stacking the parameters and variables, we can write the equation in vector form

$$y_t = \beta_0 + \hat{\mathbf{y}}_{t-1}^{\top}\boldsymbol{\beta} + \epsilon_t, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \hat{\mathbf{y}}_{t-1} = \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix}.$$

In the simplest case as described so far, the coefficients are assumed to remain constant over time. However, there may be reasons that we wish to define a richer model which allows the value of the coefficients to vary over time, replacing $\boldsymbol{\beta}$ with a time-dependent version $\boldsymbol{\beta}_t$.

Based on a time-varying structure, it is natural to consider how we can model the process $\boldsymbol{\beta}_{1:T}$ itself. Fortunately, we can consider $\boldsymbol{\beta}_t$ as a *latent variable* in a model of choice. For example, we may wish to model the coefficients with a first-order autoregression

$$\boldsymbol{\beta}_t = \mathbf{A}\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t$$

based on a transition matrix $\mathbf{A}$, which would model the process with first-order Markov chain. Based on the emission of observed values that follows from the regression equation

$$y_t = \beta_0 + \hat{\mathbf{y}}_{t-1}^{\top}\boldsymbol{\beta}_t + \epsilon_t$$

we have exactly the definition of a latent Markov model, as shown in figure 2.2 upon replacing $h_t \to \boldsymbol{\beta}_t$. Notably, adding the assumption of Gaussian noise for the emission ϵ_t and the transition $\boldsymbol{\eta}_t$, this is exactly the definition of the linear-dynamical system of section 2.2.1 upon replacing $\mathbf{h}_t \to \boldsymbol{\beta}_t$. The transition and emission distributions from equations (2.6) and (2.7) are then

$$\begin{aligned} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}) &= \mathcal{N}(\boldsymbol{\beta}_t|\mathbf{A}\boldsymbol{\beta}_{t-1}, \mathbf{Q}) \\ p(y_t|\boldsymbol{\beta}, \hat{\mathbf{y}}_{t-1}) &= \mathcal{N}(y_t|\beta_0 + \hat{\mathbf{y}}_{t-1}^{\top}\boldsymbol{\beta}_t, \sigma^2) \end{aligned}$$

in which $\hat{\mathbf{y}}_{t-1}^{\top}$ plays the role of emission matrix. The standard inference routines of the linear-dynamical system are hence applicable to this time-varying autoregressive model.

Both Fox et al. (2008) and Cassidy and Penny (2002) also discuss placing autoregressive models in a state-space form.

2.3 Cointegration

The idea of cointegration is based on the concept that whilst it may be very difficult to predict the value of a time-series or find a reliable model for how it evolves, it may be easier to find a way to model the relationship between two or more series. As I go on to describe, in contrast to the time-series models of section 2.2, a cointegration model need not make strong assumptions about the underlying generating process of the data, and this is why the introduction of the topic has been postponed to this separate section. This is in contrast to strongly-defined autoregressive models which specifically model the evolution of one or more processes.

2.3.1 Introduction to Cointegration

Whilst an individual time-series may not be predictable, the relationships between two series may be more predictable. For example, the series $x_{1:T}$, $y_{1:T}$ formed by

$$x_{t+1} = x_t + \epsilon_{t+1}, \quad y_{t+1} = y_t + \epsilon_{t+1}, \quad \epsilon_t \sim \mathcal{N}(0,1) \tag{2.15}$$

both follow an unpredictable random walk. However, the future difference $x_{t+1} - y_{t+1} = x_t - y_t$, is perfectly predictable given knowledge of the current difference. In economics and finance such instances are common; for example underlying mechanisms such as limited money supply may forge dependencies between time-series (Dickey et al., 1991).

Cointegration refers to a property that two time-series may jointly hold. For the purposes of this thesis, x_t and y_t (both $I(d)$ series, $d > 0$) are said to be cointegrated if there is a linear regression relationship between the two variables that forms a stationary process (the definition of cointegration from Granger (1986); Engle and Granger (1987) given above is broader, allowing the linear form to be integrated $I(b)$, $0 \leq b < d$). I write such a relationship[7] as

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{2.16}$$

where the residuals process ϵ_t is stationary.

It is of significant interest in finance to find pairs of asset price series that are cointegrated—such estimation underpins one of the classical statistical arbitrage strategies known as ‘pairs trading’, see for example Vidyamurthy (2004). Industry practitioners may compare thousands of pairs of asset prices to detect a cointegration relationship; the key interest is in accurately detecting the presence of a stable relation.

In the example shown in equation (2.15), both x and y are each integrated of order one, $I(1)$ random walks; as shown above, there is a simple linear combination that is a stationary $I(0)$ series. More generally, two series $x_{1:T}$ and $y_{1:T}$ are cointegrated if they are each individually integrated and a linear combination of the two is integrated with a lower order (Granger, 1986; Engle and Granger, 1987). Hamilton (1994) and Wooldridge (2009) provide thorough introductions to the topic of cointegration.

For my purposes, searching for a cointegration relationship therefore requires an estimation of the linear relationship, and evidence that the resulting sequence is stationary, noting that the two series are themselves non-stationary (integrated of order at least 1).

[7] It is straightforward to add other exogenous regressors such as a trend term in which the relation is written

$$y_t = \alpha + \beta x_t + \gamma t + \epsilon_t.$$

The simple linear regression form of the relationship between x_t and y_t given in equation (2.16) naturally appeals to the standard technique of estimation using ordinary least-squares set out in section 2.2.3. Importantly, however, we are dealing with the specific situation set out in the earlier discussion on the topic of spurious regression in which the series are not from stationary processes. This runs contrary to the standard approach of application in linear regression models, in which the practitioner would normally ensure the underlying series are stationary by taking differences. This is fundamentally why the second stage of ensuring stationarity of the residuals ϵ_t is so important to the problem of detecting cointegration. What is more, and as noted in the discussion of section 2.2.3, ordinary least-squares is not an unbiased estimator in this case (though the consistency property still holds), as I discuss below.

Stationarity of $\epsilon_{1:T}$. As set out in section 2.2.3, a time-series sample is stationary if $\langle x_t \rangle = \mathbb{E}[x_t] \in \mathbb{R}$ and $\text{Var}(x_t) \in \mathbb{R}$ are constant, and $p(x_t, x_s)$ depends on $|t - s|$ not t or s. For cointegration we therefore seek a method to determine whether these conditions hold based on observed data.

Normally, the approach is to write the process $\epsilon_{1:T}$ in an autoregressive form

$$\epsilon_t = \sum_{i=1}^{p} \phi_i \epsilon_{t-i} + \eta_t$$

where η_t is an independent noise process with $\langle \eta_t \rangle = 0$, and impose conditions on the coefficients ϕ_i to ensure stationarity. Hamilton (1994) presents a significant amount of detail on the conditions required on ϕ_i for the process to be stable. For example,

$$\begin{aligned} |\phi_1| < 1, \quad & \text{when } p = 1 \\ |\phi_1| < 1, \quad |\phi_2 - \phi_1| < 1, \quad & \text{when } p = 2. \end{aligned}$$

These conditions can be shown by considering the autoregressive recursive form for ϵ_t and imposing the stationarity conditions. In the case $p = 1$ for example, we write

$$\begin{aligned} \langle \epsilon_t^2 \rangle &= \langle \phi^2 \epsilon_{t-1}^2 + 2\phi\epsilon_{t-1}\eta_t + \eta_t^2 \rangle \\ &= \phi^2 \langle \epsilon_t^2 \rangle + \langle \eta_t^2 \rangle && \text{since } \langle \epsilon_{t-1}^2 \rangle = \langle \epsilon_t^2 \rangle, \langle \epsilon_{t-1}\eta_t \rangle = 0 \\ &= \frac{\langle \eta_t^2 \rangle}{1 - \phi^2} \end{aligned} \tag{2.17}$$

which is finite whenever $|\phi| \neq 1$ and positive whenever $|\phi| < 1$.

2.3.2 Engle-Granger Estimation

Testing for and estimating a cointegration relationship is classically a two-step process (Granger, 1986; Engle and Granger, 1987). Firstly, the regression equation is estimated based on a simple ordinary least-squares fit to minimise $\sum_t (y_t - \alpha - \beta x_t)^2$. Subsequently, a test for a unit root in the residuals $\epsilon_t \equiv y_t - \alpha - \beta x_t$ is performed using the *Dickey-Fuller* test (Dickey and Fuller, 1979), see for example Harris and Sollis (2003), which tests the hypothesis that $\phi = 1$ against the alternative $|\phi| < 1$[8]. In the

[8] The standard Dickey-Fuller test focusses on the simplest, first-order autoregression for ϵ_t; the *augmented* Dickey-Fuller test (Said and Dickey, 1984) extends the test to unknown-order processes. For brevity, in this thesis I focus on the simplest case of first-order autoregression.

case that $\phi = 1$, it is well-known that OLS delivers a spurious regression; in respect of cointegration, the problem of spurious regression can be understood as follows. In the event that there is in fact no long-run equilibrium between two series y_t and x_t, the series are not cointegrated; of course, the solution of OLS regression remains well-defined. Since there is no long-term relationship in the generating process, the OLS result is spurious and tells us nothing meaningful about the data.

Corresponding to the stationarity conditions on ϕ set out above, the null and alternative hypotheses of the Dickey-Fuller test are written

$$\begin{aligned} \mathcal{H}_0 : &\quad \phi = 1 &&\quad \text{makes } \epsilon_t \text{ a random walk} \\ \mathcal{H}_1 : &\quad |\phi| < 1 &&\quad \text{makes } \epsilon_t \text{ stationary} \end{aligned}$$

and notably, under the null hypothesis the OLS estimation was spurious which makes the two-stage cointegration detection process conceptually undesirable. The asymptotic distributions of the estimates are somewhat complicated and critical values for the tests have to be tabulated to incorporate the uncertainty in estimation of the parameters with OLS (Wooldridge, 2009; Watson and Teelucksingh, 2002).

Bias of OLS

In this cointegration model, the assumptions for consistency of OLS as set out earlier still hold; however, the stronger assumptions needed for unbiasedness of OLS are violated. By expanding the autoregression for ϵ_t we can see that the strong exogeneity condition no longer holds,

$$\begin{aligned} \epsilon_t &= \phi\left(y_{t-1} - \alpha - \beta x_{t-1}\right) + \eta_t \\ &= \phi y_{t-1} - \phi\alpha - \phi\beta x_{t-1} + \eta_t \end{aligned}$$

so strong exogeneity no longer holds when $\phi \neq 0$. As was shown by Banerjee et al. (1986), the bias for small samples can be considerable. In this case, it is possible an alternative estimator may provide more accurate values for the coefficients α, β.

2.3.3 Error-Correction Models

An alternative approach to cointegration can be considered by modelling the processes with an autoregressive relationship. Normally, this approach is used for *vector cointegration* between more than two series, and follows from representing the series in the vector $\mathbf{y}_t$ with an autoregressive process

$$\mathbf{y}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t-i} + \boldsymbol{\eta}_t$$

from which after some considerable manipulation we can write

$$\mathbf{y}_t - \mathbf{y}_{t-1} = \sum_{i=1}^{p-1} \left(\mathbf{y}_{t-i} - \mathbf{y}_{t-i-1}\right) \left[-\sum_{j=i+1}^{p} \mathbf{A}_j \right] + \mathbf{y}_{t-1} \left[\left(\sum_{i=1}^{p} \mathbf{A}_i \right) - \mathbf{I} \right] + \boldsymbol{\eta}_t.$$

When the series $\mathbf{y}_t$ are random walks $I(1)$, this first difference is hence an $I(0)$ stationary sequence. Importantly, the term in $\mathbf{y}_{t-1}$ with the matrix coefficient

$$\mathbf{B} \equiv \left(\sum_{i=1}^{p} \mathbf{A}_i \right) - \mathbf{I}$$

is known as the *error correction term* and corresponds to the cointegration relationship.

Johansen method. The technique of Johansen (1988) is focussed on extending the notion of cointegration to the vector case, by appealing to the vector error-correction formulation set out above. When dealing with the vector case $\mathbf{y}_t$ there may be more than one cointegrating vector, depending on which permutations of the individual series forming the vector are cointegrated. The Johansen method seeks statistical tests for the order of cointegration amongst d series by considering the rank r of the matrix $\mathbf{B}$; when $r = 0$, there is no cointegration, when $0 < r < d$ there are r cointegration relationships, and when $r = d$ the series are themselves stationary. Importantly, the Johansen method (which uses maximum likelihood to estimate the relationship conditioned on each order of vector cointegration) makes the strong modelling assumption about the series given in the autoregressive formulation above, that the data are $I(1)$ random walks.

2.3.4 Intermittent Cointegration

In practice, there might be reasons that one expects a cointegration relationship to apply only for short periods, or intermittently over time.

A good example is the Interconnector gas pipeline between Bacton, UK and Zeebrugge, Belgium, which allows gas to flow between the two countries, providing a direct link in the gas price at each end. When gas is cheaper in the UK than Belgium, gas can flow through the pipe to Belgium equalising the price, and vice-versa. However, the pipeline closes annually for a short period in order to perform maintenance, and inevitably during this period the link between the prices is temporarily broken. It was shown by Kim (2003) that in general, classical tests for cointegration may fail to detect a relationship even if for a part of the series there is such a relationship.

Previous works in this area typically limit the number of regimes in which cointegration can occur to either only two or three segments (Gregory and Hansen, 1996; Hatemi-J, 2008). The standard approach is to sample different combinations of segments and compare the results.

2.4 Learning By Inferring

When modelling a time-series, one must first select a model, and normally, a model incorporates one or more parameters. In the case of an autoregressive model, for example, the parameters include the linear coefficients $\boldsymbol{\beta}$ and the noise variance σ^2. Once a chosen model has been specified, we are generally interested in understanding what the observations we make are able to tell us about those parameters of the assumed model.

There are ways that one may find values to use for the model parameters based on the data we observe, so called ‘estimators’ such as the solution for $\boldsymbol{\beta}$ using ordinary least-squares already discussed. Although I do not discuss it in this thesis, much work has been done to understand the distribution over such estimators and other potentially useful properties.

A Bayesian, happy to consider probability as an abstract concept representing degree of belief, may be inclined to model parameters such as $\boldsymbol{\beta}$ as variables themselves, complete with prior distributions (normally, themselves requiring specification of hyper-parameters). Then the original problem of estimation, 'learning' from experience (data), is translated into one of probabilistic inference. The Bayesian would appeal to conditional independence, Bayes rule, and the probabilistic toolkit described by the properties of probability distributions to infer a posterior distribution. Of course, there may still be parameters that need to be specified once the posterior is found, and the key concept of updating those parameters based on the inference task is what I refer to as 'learning by inferring'.

Probabilistic inference is then a key part of the analyst's skillset—albeit, normally a computationally challenging one. This discussion on inference for learning is split into two sections corresponding to the key concerns. First, I discuss approaches to inference, with specific reference to difficult inference tasks and how one may work with intractable inference problems. Second, I move on to discuss estimation methods in probabilistic methods—corresponding to 'learning' based on the results of inference.

Depending on the class of model chosen, the inference problems can be simple or computationally troublesome. When the model has a helpful analytic form, it can be possible to calculate the inference problems by characterising the posterior with a fixed number of parameters in a set functional form as with the linear-dynamical system inference schemes set out in section 2.2.1. For other models, it is not possible to find a convenient analytical form for the posterior.

2.4.1 Closed-form Inference

When inference can be performed exactly by evaluating parameters for a fixed functional form of the posterior, we say the inference can be completed in closed form. For example, as set out in section 2.2.1 the linear-Gaussian properties of the linear-dynamical system allow inference to be calculated in closed analytical form. Whilst the same is true of the switching linear-dynamical system—the linear-Gaussian nature means the posterior is given as a mixture of Gaussians—it is usually computationally infeasible to perform these calculations, and instead approximations are made. Later in this section, I discuss methods of estimation and approximation for situations in which the posterior cannot be computed analytically in a sensible time-frame.

There has been much work done to identify properties and relationships between well-known probability distributions that permit simple closed-form algorithms for Bayesian updating. Normally, we seek distributions that leave the functional form unchanged once the prior is updated to form the posterior, conditioned on some data distribution. Combinations of data and prior distributions that permit such closed-form updates are known as conjugate pairs; for example, the Gamma distribution is a conjugate prior for the precision (inverse variance) of a Gaussian variable. This means that we may update the parameters of the Gamma distribution originally forming the prior for the Gaussian variable x once some observations of x are made (I show corresponding results in chapter 4 which appeals to this fact for inference).

Unfortunately, closed-form algorithms that permit exact inference are only available for relatively few model classes in which it is possible to write the posterior distribution in a convenient analytical form. Inevitably, strong modelling assumptions are usually needed; distributions must be known, and fit a helpful functional form. Alternatively, it may simply be impractical to find the posterior in analytical form because the computational problem is simply too complex as with the switching linear-Gaussian model. In the majority of cases exact inference is not feasible, and the modeller can no longer rely on the comfort of exact closed-form inference to fit the model to observations. Instead, one must appeal to an alternative representation of the posterior to work with, and I briefly present two key approaches here. Firstly, I discuss stochastic estimation, which seeks to express the posterior as a set of samples. Secondly, I discuss some deterministic approximation schemes in which one assumes a fixed functional form for the posterior, and selects optimal parameters to fit as closely as possible to the 'true' posterior.

2.4.2 Stochastic Estimation

Stochastic approaches to posterior estimation, also known as *sampling*, and unified under the title *Monte-Carlo* methods—so called for the connotation to gambling. As the gambler may enter a great many bets, so the Monte-Carlo estimator seeks a large number of samples to represent a distribution.

Monte-Carlo methods are founded on the concept that the expectation of some function f in variables $x_{1:n}$ may be estimated as

$$\begin{aligned}\langle f(x_{1:n})\rangle &\equiv \int_{x_{1:n}} f(x_{1:n})\, p(x_{1:n}) \\ &\approx \frac{1}{S}\sum_{s=1}^{S} f(\tilde{x}^{s}_{1:n}) \equiv \tilde{f}\big(\tilde{x}^{1:S}_{1:n}\big)\end{aligned}$$

where $\tilde{x}^{s}_{1:n}$ represents a single sample for all of the variables $x_{1:n}$ from a set of S such samples (Neal, 1993). The modeller's task is therefore to define an algorithm to find a representative sample of points from the required distribution $p(x_{1:n})$ in the latent variables. In principle, the algorithm should pick the samples in such a way that the means correspond exactly,

$$\left\langle \tilde{f}\big(\tilde{x}^{1:S}_{1:n}\big)\right\rangle_{p\left(\tilde{x}^{1:S}_{1:n}\right)} = \langle f(x_{1:n})\rangle_{p(x_{1:n})}$$

in order that $\tilde{f}$ is an unbiased estimator for f (Barber, 2012). When the samples are generated independently, only a small number of samples may be needed to accurately represent the distribution. Unfortunately, however, generating independent samples is in general a difficult task.

There have been a great number of sampling schemes developed, mostly based on a framework known as *Markov Chain Monte-Carlo* (MCMC) methods. As the name suggests, these approaches do not seek independent samples but, rather, invoke a Markov chain for sampling in which samples are found sequentially—each new sample found based on the previous sample. Key algorithms include the Gibbs sampler and the well-known Metropolis-Hastings algorithm. However, MCMC methods can be somewhat difficult to apply to time-series inference tasks like latent Markov models.

I include in this section a brief explanation of the key features of some key Monte-Carlo methods. The remainder of the section is split into two main parts; firstly I introduce key MCMC methods, and secondly talk about how Monte-Carlo methods can be applied to time-series models like the latent Markov model: so-called *sequential Monte-Carlo* (SMC). The issues involved are myriad and complex, and I include this simple discussion for completeness only; the key results of the thesis rely for the most part on closed-form inference schemes and deterministic approximation.

Markov Chain Monte-Carlo methods

MCMC methods describe techniques to draw a representative sample from a distribution $p(x_{1:n})$, and aim to provide a representative sample with good properties in terms of computational efficiency. I introduce two key methods.

Gibbs sampling. From an initial sample $\tilde{x}_{1:n}$, the Gibbs sampler draws subsequent samples by updating each component $\tilde{x}_i$ individually in turn from the conditional distribution $p(\tilde{x}_i|\tilde{x}_{1:i-1}, \tilde{x}_{i+1:n})$, on the assumption that drawing the sample $\tilde{x}_i$ from such distribution is a tractable operation.

Metropolis-Hastings. When it is computationally difficult to draw from the distribution $p(\tilde{x}_{1:n})$ but the distribution can be evaluated cheaply for candidate samples $\tilde{x}^*_{1:n}$, the Metropolis-Hastings algorithm is applicable. Based on the current sample $\tilde{x}_{1:n}$, the algorithm works by picking a candidate sample $\tilde{x}^*_{1:n}$ from a simpler distribution $q(\tilde{x}^*_{1:n}|\tilde{x}_{1:n})$ known as the *proposal distribution*, and admitting the candidate to the sample with probability dependent on the value of $p(\tilde{x}^*_{1:n})$. The new candidate is taken as a sample with probability given by the *acceptance function*; typical acceptance functions include

$$\min\left(1, \frac{p(\tilde{x}^*_{1:n})}{p(\tilde{x}_{1:n})}\right) \quad \text{or} \quad \min\left(1, \frac{q(\tilde{x}_{1:n}|\tilde{x}^*_{1:n})\, p(\tilde{x}^*_{1:n})}{q(\tilde{x}^*_{1:n}|\tilde{x}_{1:n})\, p(\tilde{x}_{1:n})}\right).$$

Although the MCMC methods given here may be difficult to apply to continuous-valued latent Markov chains, such approaches can be used for the discrete components in for example the switching linear-dynamical system: Carter and Kohn (1996) uses Gibbs sampling to deal with intractability of the trajectory of transition and emission states, and the linear-Gaussian dynamics mean inference over the continuous latent state is less troublesome.

Sequential Monte-Carlo

The term Sequential Monte-Carlo refers to a class of methods applicable to the stochastic estimation of time-series models with a latent variable chain. As the title suggests, samples are found for each sequential slice of the state posterior $p(h_t|y_{1:t})$ (smoothing is also dealt with in the literature).

One significant benefit of stochastic approximation is the inherent flexibility. With deterministic methods, in particular when searching for closed-form inference algorithms, we restrict analysis to those models in which the algebra permits analytical solutions. A key example is the restriction of the linear-dynamical system (and associated extensions thereof) to linear-Gaussian dynamics, and this is one of only a few special cases that permit such closed-form treatment. By contrast, any dynamics can be considered with stochastic approximation methods—for example linear-Gaussian latent dynamics and emission in which the latent variable informs the variance (see for example Carvalho and Lopes (2007)).

The idea is based on the concept introduced well by Gordon et al. (1993), that the state posterior (i.e. filtered posterior over the latent variables h_t) may be approximated by essentially propagating the sample through time. The samples are weighted, and the particles are updated according to the transition and emission dynamics in sequence according to each observation. These sequential Monte-Carlo methods are often known as *particle filters* since they filter the samples as 'particles' through the latent trajectory of the system. Normally, importance sampling is used at each time-step to select particles according to the transition distribution $p(h_t|h_{t-1})$ conditioned on each of the particles $\tilde{h}_{t-1}^{1:S}$ from $t-1$.

Sequential importance sampling. The key method for sequential Monte-Carlo is known as sequential importance sampling (set out well by Doucet et al. (2000)). Initially a sample is selected for $t = 1$ according to the initial state distribution $p(h_1|y_1)$. Based on the samples from t with associated weights, new particles for t are sampled according to a transition distribution $q\left(h_t\middle|\tilde{h}_{1:t-1}^s\right)$. Then weights are associated with each particle according to the relative importance of the particle; in general, the weight w_t^s for a particle $\tilde{h}_t^s$ is found as

$$w_t^s \propto w_{t-1}^s \frac{p\left(y_t\middle|\tilde{h}_t^s\right) p\left(\tilde{h}_t^s\middle|\tilde{h}_{t-1}^s\right)}{q\left(\tilde{h}_t^s\middle|\tilde{h}_{1:t-1}^s\right)}$$

where $q(h_t|h_{1:t-1})$ is an importance function that can be sampled from easily—note that the obvious choice defined by p is in general intractable since the normalisation constants are not normally known. See Barber (2012) for more details.

Significant work has been done to attempt to combine MCMC and SMC methods, however the benefits of MCMC are not trivially transferred to continuous dynamics in the time-series context and normally the advantage comes in sampling the joint distribution of the latent Markov chain with model parameters, see for example Andrieu et al. (2010). Gilks and Berzuini (2001) use MCMC methods to introduce diversity in the particles to avoid the sample degenerating through the latent Markov chain, which is a problem for long observation sequences $T \gg 1$.

2.4.3 Deterministic Approximation: Variational Inference

When posterior distributions are difficult to determine and express in analytical closed form, one may estimate the distribution with stochastic estimation using sampling methods, or alternatively form an analytical approximation using variational inference. In essence, variational approximation is concerned with assuming some (usually helpful or simple) functional form for the distribution, and then setting the parameters of the functional form such that the resulting approximate distribution is 'close' to the true distribution.

Often the 'closeness' of the approximating distribution q (known as the *variational distribution*) to the true distribution (posterior) p is considered according to a divergence measure (Minka, 2005). The most common such measure is the Kullback-Leibler divergence,

$$KL(q\|p) \equiv \left\langle \log \frac{q(x)}{p(x)} \right\rangle_{q(x)} \geq 0$$

with the property that

$$KL(q\|p) = 0 \quad \Leftrightarrow \quad p(x) \equiv q(x)\,.$$

From this we can draw intuition that larger values for the Kullback-Leibler divergence correspond to 'less-similar' distributions—the divergence can attain arbitrarily-large values.

The general approach to variational approximation is therefore to first specify a functional form for the approximating distribution $q(x_{1:n})$, which may be for example the fully-factorised distribution $\prod q(x_i)$ in which the variables are assumed independent[9]. It is usual to choose such functional form according to intuition about the model, coupled with concern for computational efficiency. The distribution is assumed to be characterised by a set of parameters θ, and one must then select a method of estimating the parameters θ to best match the true posterior $p(x_{1:n})$. Methods of selecting θ may be developed by intuition or understanding of the posterior, or alternatively sought by minimising a divergence measure like the Kullback-Leibler divergence described above. In any case, one must determine what information about the true posterior $p(x_{1:n})$ is required in order to select θ. The modeller's task is then to find those required properties of $p(x_{1:n})$ to complete each approximate inference step.

I first describe about two key methods for sequential approximate inference[10] before discussing applications to time-series data.

Sequential Approximate Inference

Inference schemes in this section describe how the approximation is updated as information from different observations is incorporated into a posterior, but the data need not be from a sequential time-series.

Assumed Density Filtering. ADF is a one-pass method for approximate inference based on sequentially updating the approximate posterior q once for each observation. The algorithm usually begins with the prior $p(x)$ and updates according to the first observation y^1, the first approximation $q(x)$ chosen to fit closely to the true posterior $p(x)\,p\big(y^1\big|x\big)$. Then the algorithm recursively refines the approximation $q(x)$ for each subsequent observation $i = 2, \ldots, n$ to fit the update $q(x)\,p\big(y^i\big|x\big)$.

Expectation Propagation. EP is a technique first introduced by Minka (2001) that extends ADF by adding further passes through the data to overcome some of the shortcomings of ADF. In particular, ADF is sensitive to the order of the data (when the data are not sequential) and EP tries to build information into the posterior that may have been discarded from earlier observations. To implement EP, one may first

[9] The fully-factorised approach to variational inference corresponds to a method used in the field of statistical physics known as *mean-field theory*.

[10] For the purposes of this section, I distinguish *sequential approximate inference* from *inference with sequential data* (which implies the data are associated with an ordering).

complete inference with ADF, and then perform multiple backward (and forward) updates through the data until the posterior q converges.

An interesting point for discussion with variational approximate inference comes from the choice of parameters θ by minimising the divergence $KL(q\|p)$. The Kullback-Leibler divergence is asymmetric in its arguments: in general, $KL(q\|p) \neq KL(p\|q)$. The choice of which combination to minimise has interesting properties. Minimising $KL(p\|q)$ is known as *moment matching* in the case that q is in the exponential family of distributions since it can be shown that the moments of q must be chosen to match those of the exact posterior p. This approach heavily penalises small values in the variational distribution $q(x)$ when the corresponding value of $p(x)$ is large, so the approximation may overestimate the variance of the true posterior. When fitting a unimodal variational distribution q to a multimodal posterior p, the resulting approximation generally averages over all of the modes. By contrast, minimising $KL(q\|p)$ instead penalises large values in the variational distribution $q(x)$ when the corresponding value of $p(x)$ is small. This means that the variance is normally underestimated, and in particular fitting a unimodal distribution q to a multimodal distribution p will generally fit q to a single mode of p. The discussion on properties of the different orderings for the divergence are discussed by Bishop (2007) and Barber (2012). It is normally the case that EP minimises $KL(p\|q)$.

Approximate Inference with Sequential Data

In the context of time-series, we are specifically interested in variational inference where the data come with an associated ordering.

There are a great many works on approximate inference for latent Markov models, and in particular the switching linear-dynamical system (since as already noted, the number of possible state trajectories is exponential in the number of observations T). A significant proportion can be considered a form of assumed density filtering: one picks an analytical form for the filtering messages α and then devises a method of choosing the parameters to provide a good fit to the true posterior. In the case of a switching linear-dynamical system with S states, it is common to characterise each posterior inference message with exactly S Gaussian components, and collapse the S^2 components found from the sequential update to the S components permitted by the representation—examples include Kim's method (Kim, 1994; Kim and Nelson, 1999) and Expectation Correction (Barber, 2006). Alspach and Sorenson (1972) describe how non-linear dynamics may be approximated with Gaussian mixtures; alternatively, local linearisation is known as the *Extended Kalman Filter* (Jazwinski, 1970).

2.4.4 Parameter Estimation

A data model is normally formed as a set of rules for generating data, for example a regression function. Probabilistic models of the type dealt with throughout this thesis are generally formed by specifying a joint density function that describes a probabilistic interaction between variables of interest, both latent and observed. All models normally have some of parameters that specify the particular characteristics of how data are calculated in some application. For example, a simple regression model is specified by the linear coefficients. Meanwhile, a Gaussian distribution in a probabilistic model is specified by a mean and covariance.

Normally the modeller has reason for selecting a particular model structure over another. But the selection of the parameter values to plug into the model is a key consideration.

The main scheme for selecting parameters is one of estimation, which in general describes schemes to find values of parameters that seem appropriate given some set of data observations. In the nomenclature of machine learning, such estimation based on data corresponds to the nexus concept of *learning*.

The Bayesian would naturally take the view that each parameter should be treated as a variable, place a prior distribution and aim to integrate the parameter out of the likelihood function. However, all but the most trivial probability distributions are themselves specified by one or more parameters—and the Bayesian is required to select appropriate values for these hyper-parameters. Essentially the problem of parameter choice remains; in essence the Bayesian has replaced one set of parameters with another set of hyper-parameters.

The most widely-used approach to estimate the model parameters θ is to maximise the data likelihood as a function in the parameters, $p(\mathcal{D}|\theta)$. We then take

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

so the problem of estimation becomes one of optimisation. Note that maximising the likelihood is equivalent to maximising the $\log$ of the likelihood, since $\log$ is a monotonic increasing function. In the machine learning community it is common to rather speak of minimising a *cost function*, for which we may of course take the negative of the likelihood.

Optimisation is a very broad topic and I do not go into great detail here. Under certain conditions (see for example Boyd and Vandenberghe (2004)) it is possible to optimise a function with respect to a parameter by direct methods, including for convex functions by differentiating the function with respect to the parameter to seek the stationary point. One may need to add Lagrange multiplier terms to deal with constraints on the parameters—for example, one my constrain entries in a conditional probability table to normalise.

Expectation Maximisation. For other functions, optimisation is a difficult problem, and one commonly-used algorithm is known as Expectation Maximisation. The EM algorithm (Dempster et al., 1977) is an iterative algorithm for maximising the data likelihood. Writing v for the observed variables and h for the latent variables, EM is based on the Kullback-Leibler bound

$$0 \leq KL(q\|p) \equiv \left\langle \log \frac{q(h|v)}{p(h|v,\theta)} \right\rangle_{q(h|v)}$$

from which we can write

$$\log p(v|\theta) \geq \underbrace{\langle \log p(h,v|\theta) \rangle_{q(h|v)}}_{\text{energy}} - \underbrace{\langle \log q(h|v) \rangle_{q(h|v)}}_{\text{entropy}} .$$

By considering p to be the variational distribution in which we are interested to set the parameters θ, we can see that we have so formed a lower bound on the data likelihood $p(v|\theta)$, in terms of the fixed posterior $q(h|v)$.

For our purposes, EM corresponds to iteratively maximising this lower bound in a two-stage process. First, inference is performed based on parameter estimates θ^{old}, and this posterior is used as $q(h|v)$ in the second stage, in which the right-hand side of the bound is maximised by choice of θ in the density function of the variational distribution p (only the energy term[11] is relevant). The two stages are then repeated until the algorithm reaches convergence. Even though the EM algorithm only maximises the lower bound on the (log) likelihood, it can be shown (see for example Barber (2012)) that each iteration of the two-stage algorithm is in fact guaranteed to increase the data likelihood itself.

Variational Expectation Maximisation. EM has the property that each iteration is guaranteed to increase the marginal likelihood. However, when we use an approximation to the exact posterior, either deterministic according to a variational technique or stochastic using some form of sampling, this important characteristic no longer holds, and taking EM steps in likelihood space can have unpredictable results. A technique has been developed known as *variational Expectation Maximisation* (see for example Turner and Sahani (2011)) that seeks to deal with maximising the data likelihood in the absence of an exact posterior.

Variational Bayes. An alternative algorithm known as Variational Bayes (VB) is also available for parameter modelling and optimisation. This method (see Attias (2000)) takes the approach of repeatedly fitting a variational distribution $q(h, \theta|v)$ where it is assumed that the parameters are independent of the latent variables,

$$q(h, \theta|v) \equiv q(h|v)\, q(\theta|v)\,.$$

Following this approach we can see VB as a generalisation of EM, where EM makes the assumption that the parameters θ are approximated by a single value $\widehat{\theta}$; the posterior is therefore given as $q(\theta|v) \equiv \delta\left(\theta - \widehat{\theta}\right)$. Using the VB algorithm would allow one to model the parameters θ with a prior distribution and update the sufficient statistics (assuming that the prior distribution has a conjugate form) with each iteration of the algorithm, contrasting with the point estimate provided by EM.

Cassidy and Penny (2002) use VB in a mulivariate autoregressive model similar to the form discussed in section 2.2.4.

[11] Also called the expected completed data log likelihood.

PART II

Contribution

CHAPTER 3

Switch-Reset Models: Exact and Approximate Inference

Reset models are constrained switching latent Markov models in which the dynamics either continues according to a standard model, or the latent variable is resampled. We consider exact marginal inference in this class of models and their extension, the switch-reset models. A further convenient class of conjugate-exponential reset models is also discussed. For a length T time-series, exact filtering scales with T^2 and smoothing T^3. We discuss approximate filtering and smoothing routines that scale linearly with T.

This contribution of this chapter was published in Bracegirdle and Barber (2011).

3.1 Reset Models

The switching linear dynamical model set out in section 2.2.1 represents an example of a latent variable model with an additional internal state. We can therefore identify a class of switching latent Markov model in which the latent variable is augmented with a second discrete variable chain s_t representing the prevailing state. The joint distribution function for such a model is given as

$$p(y_{1:T}, h_{1:T}, s_{1:T}) = \prod_{t=1}^{T} p(h_t|h_{t-1}, s_t)\, p(y_t|h_t, s_t)\, p(s_t|s_{t-1}) \,, \quad h_0 = \emptyset, \quad s_0 = \emptyset \tag{3.1}$$

in which we can deal with discontinuous jumps in the continuous latent state h_t by using a discrete 'switch' variable s_t. A belief network for this joint density is shown in figure 2.3.

These switching models are attractive because evolving state regimes in the chain $s_{1:T}$ allow for rich modelling frameworks that may match our understanding of a system of interest: dynamics may switch to different modes of evolution in certain circumstances, for example. Whilst the flexibility of these models is potentially very powerful, they suffer from a well known computational difficulty as discussed in section 2.2.1 and section 2.4.3. This relates to the fact that the messages in the propagation algorithms (the analogue of equations (2.1), (2.2) and (2.4) applied to the two latent variables s_t, h_t in place of h_t alone) are given as mixtures with a number of components that grows exponentially with time t. This means that marginal inference of quantities such as the filtered posterior $p(h_t|y_{1:t})$ scales with $O\left(S^t\right)$, and for the smoothed posterior $p(h_t|y_{1:T})$ inference scales with $O\left(S^T\right)$. The overcome the computational intractability, a number of different methods have been proposed including the stochastic estimation schemes and deterministic inference methods set out in sections 2.4.2 and 2.4.3.

An alternative, computationally simpler model is obtained by constraining the switch variable to have the effect of cutting temporal dependence[1] in the latent variable chain h_t. To construct such a model, I first define a reset variable $c_t \in \{0, 1\}$ with Markov transition

$$p(c_t = j|c_{t-1} = i) = \tau_{j|i}, \quad i, j \in \{0, 1\}. \tag{3.2}$$

The continuous latent variable then transitions as

$$p(h_t|h_{t-1}, c_t) = \begin{cases} p^0(h_t|h_{t-1}) & c_t = 0 \\ p^1(h_t) & c_t = 1 \end{cases}. \tag{3.3}$$

In this model, the latent binary variable c_t selects one of only two possible dynamics: either a continuation along the default dynamics $p^0(h_t|h_{t-1})$, or a 'reset' of the latent variable, drawing from the reset distribution $p^1(h_t)$. This reset process cuts the dependence on past states[2], see the belief network given in section 3.1.

Finally, the reset model is completed by specifying an emission distribution[3]

$$p(y_t|h_t, c_t) = \begin{cases} p^0(y_t|h_t) & c_t = 0 \\ p^1(y_t|h_t) & c_t = 1 \end{cases}. \tag{3.4}$$

For the reset model, it is well appreciated that filtered marginal inference $p(h_t, c_t|y_{1:t})$ scales as $O\left(t^2\right)$ (see for example Fearnhead and Liu (2007) and Barber and Cemgil (2010)), and smoothed marginal inference $p(h_t, c_t|y_{1:T})$ can be achieved in $O\left(T^3\right)$ time. Whilst this is a great saving from the exponential complexity of the switching model, cubic complexity is still prohibitive for large T and approximations may be required.

The contribution of this chapter is to introduce an exact, numerically stable correction smoothing method for reset models, in addition to demonstrating a fast and accurate linear-time approximation. I also describe an extension, the switch-reset model, which is able to model switching between a set of S continuous latent Markov models, but for which inference remains tractable.

[1] Some refer to equation (3.3) as a 'changepoint' model whilst others use this terminology for any switching latent Markov model of the form equation (3.1). Others refer to the piecewise reset model, section 3.5 as a 'changepoint' model. For this reason, in an attempt to avoid confusion and to better represent the models of the assumption, I refer to equation (3.3) as a reset model.

[2] Similarities can be drawn between the reset model in this chapter and the variable-duration HMM set out by Murphy (2002), a form of hidden semi-Markov model.

[3] Note that it is straightforward to include dependence on past observations $p(y_t|h_t, c_t, y_{1:t-1})$ if desired since these do not change the structure of the recursions.

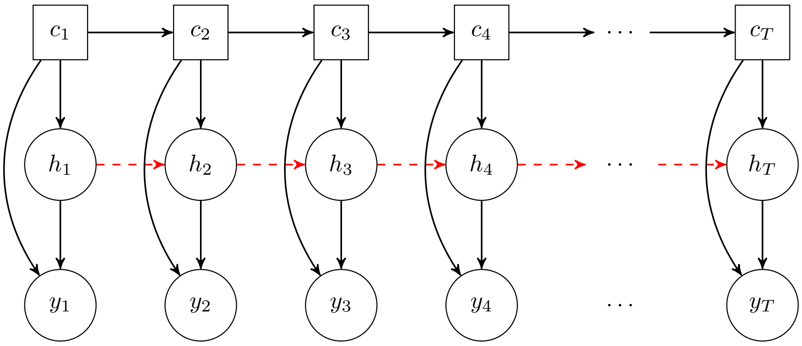


Figure 3.1: Conditional independence assumptions of a reset latent Markov model. The binary reset variable c_t indicates whether the standard dynamics continues, $p^0(h_t|h_{t-1})$ (corresponding to $c_t = 0$) or whether the latent variable h_t is redrawn from the reset distribution $p^1(h_t)$ (corresponding to $c_t = 1$).

3.2 Reset Model inference

A classical approach to deriving smoothing recursions for the reset model is based on the α-β recursion analogous to the updates given in equations (2.1) and (2.2), which has the advantage of being straightforward. However, for models such as the reset linear dynamical system set out in section 3.3, numerical stability issues are known to arise. In addition, it is unclear how best to form an approximation based on the α-β method since the backward-recursive messages are not distributions in the latent variables. To illustrate, I first review the α-β approach.

3.2.1 α-β Smoothing

By writing

$$p(h_t, c_t, y_{1:T}) = \underbrace{p(h_t, c_t, y_{1:t})}_{\alpha(h_t,c_t)}\underbrace{p(y_{t+1:T}|h_t, c_t, \cancel{y_{1:t}})}_{\beta(h_t,c_t)}$$

we can focus on calculating the two components. The forward α message is standard and recursively calculated using equation (2.1) upon replacing the latent variable $h_t \to (h_t, c_t)$. By defining[4]

$$\alpha(h_t, c_t) = \begin{cases} \alpha^0(h_t) & c_t = 0 \\ \alpha^1(h_t) & c_t = 1 \end{cases}$$

and $\alpha(c_t) = \int_{h_t} \alpha(h_t, c_t)$, I identify two cases of a reset and no-reset at t,

$$\alpha^0(h_t) = \sum_{c_{t-1}} \tau_{0|c_{t-1}} p^0(y_t|h_t) \int_{h_{t-1}} p^0(h_t|h_{t-1})\, \alpha^{c_{t-1}}(h_{t-1})$$
$$\alpha^1(h_t) = p^1(y_t|h_t)\, p^1(h_t) \sum_{c_{t-1}} \tau_{1|c_{t-1}} \alpha(c_{t-1})\,.$$

From these recursions, we see that the number of components in α grows linearly with time, making for an $O\left(t^2\right)$ computation for exact filtering.

[4] I will use component-conditional notation for these messages $\alpha(h_t, c_t) = \alpha(h_t|c_t)\,\alpha(c_t)$, which defines $\alpha(h_t|c_t) = p(h_t|c_t, y_{1:t})$ and $\alpha(c_t) = p(c_t, y_{1:t})$.

The backward β message $\beta(h_t, c_t) = p(y_{t+1:T}|h_t, c_t)$ is also calculated recursively using equation (2.2), which follows as

$$\beta(h_{t-1}, c_{t-1}) = \sum_{c_t} \tau_{c_t|c_{t-1}} \int_{h_t} p(y_t|h_t, c_t)\, p(h_t|h_{t-1}, c_t)\, \beta(h_t, c_t)$$

$$= \tau_{0|c_{t-1}} \underbrace{\int_{h_t} p^0(y_t|h_t)\, p^0(h_t|h_{t-1})\, \beta(h_t, c_t = 0)}_{\beta^0(h_{t-1})} + \tau_{1|c_{t-1}} \underbrace{\int_{h_t} p^1(y_t|h_t)\, p^1(h_t)\, \beta(h_t, c_t = 1)}_{\beta^1_{t-1}}$$

where I have written

$$\beta(h_{t-1}, c_{t-1}) = \tau_{0|c_{t-1}} \beta^0(h_{t-1}) + \tau_{1|c_{t-1}} \beta^1_{t-1}.$$

The recursions for these components are given as

$$\beta^0(h_{t-1}) = \int_{h_t} p^0(y_t|h_t)\, p^0(h_t|h_{t-1}) \left[\tau_{0|0}\beta^0(h_t) + \tau_{1|0}\beta^1_t\right], \tag{3.5}$$

$$\beta^1_{t-1} = \int_{h_t} p^1(y_t|h_t)\, p^1(h_t) \left[\tau_{0|1}\beta^0(h_t) + \tau_{1|1}\beta^1_t\right]. \tag{3.6}$$

The posterior $p(h_t, c_t|y_{1:T}) \propto \alpha(h_t, c_t)\, \beta(h_t, c_t)$ is then a mixture of $(t+1) \times (T - t + 1)$ components, and the algorithm scales as $O\left(T^3\right)$ to compute all the smoothed marginal posteriors. For large T, this can be expensive.

An obvious way to form an approximation is to drop components from either the α or β messages, or both. Dropping components from α is natural (since $\alpha(h_t, c_t)$ is a distribution in h_t and c_t up to a normalisation constant). It is less natural to form an approximation by dropping β components since the β messages are not distributions—usually it is only their interaction with the α message that is of ultimate interest (corresponding to the calculations of algorithm 2.3). I discuss ways to achieve this in section 3.4. Use of the β message approach is also known to cause numerical instability in important models of interest, in particular the linear dynamical system (Verhaegen and Van Dooren, 2002). This motivates the desire to find a γ, 'correction smoother' recursion.

3.2.2 α-γ Smoothing

Considering the standard γ correction smoother derivation, equation (2.4), we may begin

$$\gamma(h_{t-1}, c_{t-1}) = p(h_{t-1}, c_{t-1}|y_{1:T}) = \sum_{c_t} \int_{h_t} p(h_{t-1}, c_{t-1}|h_t, c_t, y_{1:t-1})\, \gamma(h_t, c_t)$$

The naïve approach is then to write the dynamics reversal term set out in equation (2.5),

$$p(h_{t-1}, c_{t-1}|h_t, c_t, y_{1:t-1}, \cancel{y_t}) = \frac{p(h_t, c_t, h_{t-1}, c_{t-1}|y_{1:t})}{p(h_t, c_t|y_{1:t})}.$$

However, the filtered distribution in the denominator $p(h_t, c_t|y_{1:t})$ is a mixture distribution. This is inconvenient since we cannot represent in closed form the result of the division of a mixture by another mixture, and this means that using a γ recursion is not directly accessible for this model. However, by considering an equivalent model, it is possible to perform γ smoothing.

3.2.3 $\tilde{\alpha}$-$\tilde{\gamma}$ Run-Length Smoothing

In this section, I first define an equivalent reset model based on the 'run length', $\rho_t \geq 0$, which counts the number of steps since the last reset (Adams and MacKay, 2007; Fearnhead and Liu, 2007),

$$\rho_t = \begin{cases} 0 & c_t = 1 \\ \rho_{t-1} + 1 & c_t = 0 \end{cases} \tag{3.7}$$

and $c_t = \mathbb{I}\left[\rho_t = 0\right]$. Formally, one can write a Markov transition on the run-length defined by

$$p(\rho_t|\rho_{t-1}) = \begin{cases} \tau_{1|1} & \rho_{t-1} = 0, \rho_t = 0 \\ \tau_{1|0} & \rho_{t-1} > 0, \rho_t = 0 \\ \tau_{0|1} & \rho_{t-1} = 0, \rho_t = 1 \\ \tau_{0|0} & \rho_{t-1} > 0, \rho_t = \rho_{t-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

and a corresponding latent Markov model

$$p(h_t|h_{t-1}, \rho_t) = \begin{cases} p^0(h_t|h_{t-1}) & \rho_t > 0 \\ p^1(h_t) & \rho_t = 0. \end{cases} \tag{3.9}$$

Finally

$$p(y_t|h_t, \rho_t) = \begin{cases} p^0(y_t|h_t) & \rho_t > 0 \\ p^1(y_t|h_t) & \rho_t = 0. \end{cases}$$

This model is then formally equivalent to the reset model defined by equations (3.2) to (3.4). Since this is a latent Markov model, we can apply the standard filtering and smoothing recursions. For filtering, we follow equation (2.1) with

$$\tilde{\alpha}(h_t, \rho_t) = p(y_t|h_t, \rho_t) \sum_{\rho_{t-1}} p(\rho_t|\rho_{t-1}) \int_{h_{t-1}} p(h_t|h_{t-1}, \rho_t)\, \tilde{\alpha}(h_{t-1}, \rho_{t-1})\,.$$

I again distinguish two cases,

$$\tilde{\alpha}(h_t, \rho_t = 0) = p^1(y_t|h_t)\, p^1(h_t) \sum_{\rho_{t-1}} p(\rho_t = 0|\rho_{t-1})\, \tilde{\alpha}(\rho_{t-1}) \tag{3.10}$$

$$\tilde{\alpha}(h_t, \rho_t > 0) = p^0(y_t|h_t)\, p(\rho_t|\rho_{t-1} = \rho_t - 1) \int\limits_{h_{t-1}} p^0(h_t|h_{t-1})\, \tilde{\alpha}(h_{t-1}, \rho_{t-1} = \rho_t - 1)\,. \tag{3.11}$$

In this case the $\tilde{\alpha}$ messages are therefore not mixtures, but single-component distributions. For completeness, the filtered posterior in the original reset model is obtained from

$$\alpha(h_t, c_t) = \begin{cases} \tilde{\alpha}(h_t, \rho_t = 0) & c_t = 1 \\ \sum_{\rho_t > 0} \tilde{\alpha}(h_t, \rho_t) & c_t = 0. \end{cases}$$

The run-length gives a natural interpretation of the components in the α message, namely that the components of the $\alpha(h_t, c_t)$ message are in fact simply the run-length components themselves.

Since the $\tilde{\alpha}$ messages are single components, one may implement the standard correction approach for smoothing on this redefined model,

$$\begin{aligned}\tilde{\gamma}(h_{t-1},\rho_{t-1}) &= \sum_{\rho_t}\int_{h_t} p(h_{t-1},\rho_{t-1}|h_t,\rho_t,y_{1:t-1})\,\tilde{\gamma}(h_t,\rho_t)\\ &= \sum_{\rho_t} p(\rho_{t-1}|\rho_t,y_{1:t-1})\underbrace{\int_{h_t}\frac{p(h_t|h_{t-1},\rho_t)\,\tilde{\alpha}(h_{t-1}|\rho_{t-1})}{p(h_t|\rho_t,y_{1:t-1})}}_{\text{dynamics reversal}}\tilde{\gamma}(h_t,\rho_t)\end{aligned}$$

where $p(\rho_{t-1}|\rho_t,y_{1:t-1}) \propto p(\rho_t|\rho_{t-1})\,\tilde{\alpha}(\rho_{t-1})$. Since $\tilde{\alpha}(h_{t-1}|\rho_{t-1})$ is a single component, the 'dynamics reversal' is a single component, and no numerical difficulties arise. As with the filtered posterior, the smoothed posterior for the original model is easily recovered since

$$\gamma(h_t,c_t) = \begin{cases}\tilde{\gamma}(h_t,\rho_t=0) & c_t=1\\ \sum_{\rho_t>0}\tilde{\gamma}(h_t,\rho_t) & c_t=0.\end{cases} \tag{3.12}$$

The resulting $\tilde{\alpha}$-$\tilde{\gamma}$ recursion provides a numerically stable way to perform smoothed inference in reset models since both the $\tilde{\alpha}$ and $\tilde{\gamma}$ messages are distributions.

Furthermore, a simple approximate smoothing algorithm is available based on dropping components from $\tilde{\alpha}$ and subsequently from $\tilde{\gamma}$. Simple schemes such as dropping low weight components can be very effective in this case since the weight of the component is directly related to its contribution to the posterior distribution. This is an example of a determinstic approximation scheme of the type set out in section 2.4.3 and corresponds to a form of Assumed Density Filtering, since the posterior is fir to a mixture distribution with a defined number of components.

3.2.4 Bracket Smoothing

Insight into the above $\tilde{\gamma}$ recursion can be obtained by introducing the index ς to correspond to the number of observation points until the next reset, the smoothing analogue of the forward-time run-length index ρ. I characterise this index as $\varsigma_t \in \{1,\ldots,T-t+1\}$, where $\varsigma_t = T-t+1$ corresponds to there being no reset in the sequence after observation point t. The forward run-length $\rho_t \in \{0,\ldots,t\}$ at observation t corresponds to the number of observation points since the last reset. I then index the smoothed posterior components[5] as $p(h_t,\rho_t,\varsigma_t|y_{1:T}) = p(h_t|\rho_t,\varsigma_t,y_{1:T})\,p(\rho_t,\varsigma_t|y_{1:T})$.

The smoothed partition posterior $p(\rho_t,\varsigma_t|y_{1:T})$ can then be calculated by a simple backward recursion, noting that in the no-reset case $\varsigma_t > 1$,

$$p(\rho_t,\varsigma_t|y_{1:T}) = p(\rho_{t+1}=\rho_t+1,\varsigma_{t+1}=\varsigma_t-1|y_{1:T})\,. \tag{3.13}$$

[5] The component $p(h_t|\rho_t,\varsigma_t,y_{1:T})$ describes the distribution of h_t given that (i) the previous reset occurred ρ_t time-steps ago (or there has not been a reset prior to time t if $\rho_t = t$); and (ii) the next reset occurs ς_t time-steps in the future (or there is no reset after time t if $\varsigma_t = T-t+1$). This is equivalent to asserting that the previous reset occurred at time $t-\rho_t$ (or there was no previous reset if $t-\rho_t < 1$) and that the next reset occurs at time $t+\varsigma_t$ (or there is no future reset if $t+\varsigma_t > T$).

In the reset case $\varsigma_t = 1 \Leftrightarrow \rho_{t+1} = 0$, so

$$\begin{aligned} p(\rho_t, \varsigma_t = 1|y_{1:T}) &= p(\rho_t, \rho_{t+1} = 0|y_{1:T}) \\ &= p(\rho_t|\rho_{t+1} = 0, y_{1:t}) \sum_{\varsigma_{t+1}} p(\rho_{t+1} = 0, \varsigma_{t+1}|y_{1:T}) \end{aligned} \quad (3.14)$$

since $\rho_t \perp\!\!\!\perp y_{t+1:T} \,|\, \rho_{t+1} = 0$. Then

$$p(\rho_t|\rho_{t+1} = 0, y_{1:t}) \propto p(\rho_{t+1} = 0|\rho_t)\, p(\rho_t|y_{1:t})$$

and $p(\rho_t|y_{1:t}) \propto \tilde{\alpha}(\rho_t)$. These recursions enable one to fully compute the discrete component $p(\rho_t, \varsigma_t|y_{1:T})$.

Reset points partition the sequence[6], so conditioning on ρ_t and ς_t simplifies the model to use only standard dynamics p^0 on the 'bracket' $y_{\rho_t,\varsigma_t} \equiv y_{t-\rho_t:t+\varsigma_t-1}$. Smoothing for the joint is then obtained using

$$\underbrace{p(h_t, \rho_t, \varsigma_t|y_{1:T})}_{\tilde{\gamma}(h_t,\rho_t,\varsigma_t)} = \underbrace{p(h_t|\rho_t, \varsigma_t, y_{\rho_t,\varsigma_t})}_{\tilde{\gamma}(h_t|\rho_t,\varsigma_t)} \underbrace{p(\rho_t, \varsigma_t|y_{1:T})}_{\tilde{\gamma}(\rho_t,\varsigma_t)}.$$

For the continuous component $p(h_t|\rho_t, \varsigma_t, y_{\rho_t,\varsigma_t})$ we may run any smoothing routine with the dynamics p^0 on the bracket y_{ρ_t,ς_t}, with

$$\tilde{\gamma}(h_t|\rho_t, \varsigma_t > 1) = \int_{h_{t+1}} p(h_t|h_{t+1}, c_{t+1} = 0)\, \tilde{\gamma}(h_{t+1}|\rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1) \quad (3.15)$$

noting that $\tilde{\gamma}(h_t|\rho_t, \varsigma_t = 1) = \tilde{\alpha}(h_t|\rho_t)$. This is the single component analogue of equation (2.4) for the case of the reset model.

Finally, these messages enable us to perform smoothing for the original reset model by appealing to equation (3.12) after marginalising the future reset index ς_t.

3.2.5 $\tilde{\beta}$ Recursion

It is also useful to index the β recursion with the ς indexing variable, and the recursions become

$$\begin{aligned} \tilde{\beta}(h_{t-1}, c_{t-1}, \varsigma_{t-1}) &= \begin{cases} \tau_{1|c_{t-1}} \tilde{\beta}^1_{t-1} & \varsigma_{t-1} = 1 \\ \tau_{0|c_{t-1}} \tilde{\beta}^0(h_{t-1}, \varsigma_{t-1}) & \varsigma_{t-1} > 1 \end{cases} \\ \tilde{\beta}^1_{t-1} &= \int_{h_t} p^1(y_t|h_t)\, p^1(h_t) \sum_{\varsigma_t} \tilde{\beta}(h_t, \rho_t = 1, \varsigma_t) \\ \tilde{\beta}^0(h_{t-1}, \varsigma_{t-1}) &= \int_{h_t} p^0(y_t|h_t)\, p^0(h_t|h_{t-1})\, \tilde{\beta}(h_t, \rho_t = 0, \varsigma_t = \varsigma_{t-1} - 1). \end{aligned}$$

We can then combine any combination of α or $\tilde{\alpha}$ with β or $\tilde{\beta}$; since upon indexing according to both indices ρ_t and ς_t each such message features a single component, deterministic approximations can be readily motivated.

[6] As set out in section 2.2.2, data in each segment are independent of other observations conditioned on the partition.

3.3 The Reset Linear Dynamical System

I have previously set out the structure and inference routines for the linear dynamical system in section 2.2.1, including the forward α filtering, and β and γ smooting recursions. These recursions are analytically tractable because the derivations as set out in appendix B show that manipulating Gaussians under linear transformation results in Gaussian posteriors.

In this section I define a reset LDS, incorporating the reset concept into the latent variable chain h_t with the effect of cutting the dynamics. The reset LDS is defined with transition distribution

$$p(\mathbf{h}_t|\mathbf{h}_{t-1}, c_t) = \begin{cases} \mathcal{N}\big(\mathbf{h}_t \big| \mathbf{A}^0\mathbf{h}_{t-1} + \bar{\mathbf{h}}^0, \mathbf{Q}^0\big) & c_t = 0 \\ \mathcal{N}\big(\mathbf{h}_t \big| \bar{\mathbf{h}}^1, \mathbf{Q}^1\big) & c_t = 1 \end{cases}$$

and emission

$$p(\mathbf{y}_t|\mathbf{h}_t, c_t) = \begin{cases} \mathcal{N}\big(\mathbf{y}_t \big| \mathbf{B}^0\mathbf{h}_t + \bar{\mathbf{y}}^0, \mathbf{R}^0\big) & c_t = 0 \\ \mathcal{N}\big(\mathbf{y}_t \big| \mathbf{B}^1\mathbf{h}_t + \bar{\mathbf{y}}^1, \mathbf{R}^1\big) & c_t = 1. \end{cases}$$

From which the effect of a reset ($c_t = 1$) to cause a break in the transition dynamics is clear. The belief network of section 3.1 remains applicable.

3.3.1 Marginal Inference

Filtering based on the run-length formalism for the reset LDS is straightforward to implement as I set out here. In this case each component of the filtered distribution is represented by

$$\tilde{\alpha}(\mathbf{h}_t|\rho_t) = \mathcal{N}(\mathbf{h}_t|\mathbf{f}_t(\rho_t), \mathbf{F}_t(\rho_t))$$

and since $\tilde{\alpha}(\rho_t) = p(\rho_t|\mathbf{y}_{1:t})\,p(\mathbf{y}_{1:t})$, we take $p(\rho_t|\mathbf{y}_{1:t}) \equiv w_t(\rho_t)$ and $p(\mathbf{y}_{1:t}) \equiv l_t$. Filtering then corresponds to sequentially updating the mean parameter $\mathbf{f}_t(\rho_t)$, covariance $\mathbf{F}_t(\rho_t)$, mixture weights $w_t(\rho_t)$, and likelihood l_t—the final routine is given in algorithm 3.1[7]. This form of filtering is particularly useful since, as explained in section 3.2.3, an exact correction smoother follows.

α-β Smoothing

For completeness and in order to compare approximate methods based on neglecting message components, I also implement the $\beta(\mathbf{h}_t, c_t)$ messages.

The β message can be used for smoothing in the reset LDS, utilising the canonical backward recursion of algorithm 2.2 and α-β combination of algorithm 2.3 to extract the smoothed posterior of interest.

The β message is given as a combination of quadratic-exponential canonical forms with a number of components that increments each iteration. In order to complete the β recursions for the reset model set out above, we need to calculate $\beta^0(\mathbf{h}_{t-1})$ and β^1_{t-1} from section 3.2.1 according to equations (3.5) and (3.6). Formally, one carries out the β recursion under the assumption of a mixture of canonical forms.

[7] Note that, for the general model, LDSFORWARD and LDSBACKWARD can be replaced with any update routine calculating sufficient statistics and corresponding likelihood.

Each $\beta^0(\mathbf{h}_{t-1})$ is calculated in canonical form using the standard backward information recursion of section 2.2.1; each β^0 message is of the form $\sum_j k_{tj} \exp -\frac{1}{2}\left(\mathbf{h}_t^\top \mathbf{P}_{tj}\mathbf{h}_t - 2\mathbf{h}_t^\top \mathbf{p}_{tj}\right)$, where the constant terms k_{tj} are necessary to compute the weights in the full posterior.

Fortunately, $\beta^0(\mathbf{h}_{t-1})$ can be easily calculated according to equation (3.5) by repeated application of the standard backwards information recursion given in algorithm 2.2 for each canonical component in the preceding message $\beta^0(\mathbf{h}_t)$. An additional component is added in each iteration from the β_t^1 term, which is simply written in quadratic-exponential form.

The constant value β_{t-1}^1 is also simple to calculate. To see this, we first note that the term in β_t^1 is trivially calculated by marginalising the variable $\mathbf{h}_t$. For each canonical component from the previous iteration in $\beta^0(\mathbf{h}_t)$ we note that the only remaining term is the constant term k_t from the canonical update of algorithm 2.2 since the break in dynamics can be considered with the standard update equations with the trivial transition matrix $\mathbf{A} = \mathbf{0}$, thus removing the exponential component.

The result is that whereas $\alpha(\mathbf{h}_t, c_t)$ is represented as a mixture of Gaussians in moment form, $\beta(\mathbf{h}_t, c_t)$ is a mixture of squared exponentials in canonical form. To compute the smoothed posterior $p(\mathbf{h}_t, c_t|\mathbf{y}_{1:T}) \propto \alpha(\mathbf{h}_t, c_t)\,\beta(\mathbf{h}_t, c_t)$, we need to multiply out both mixtures, converting the resulting mixture of moment-canonical interactions to a mixture of moments. To do this, we write

$$\begin{aligned} p(\mathbf{h}_t, c_t|\mathbf{y}_{1:T}) &\propto \alpha(\mathbf{h}_t, c_t)\,\beta(\mathbf{h}_t, c_t) = \alpha(\mathbf{h}_t, c_t)\left[\tau_{0|c_t}\beta^0(\mathbf{h}_t) + \tau_{1|c_t}\beta_t^1\right] \\ &\propto \left[\sum_{i(c_t)} w_i \mathcal{N}(\mathbf{f}_i, \mathbf{F}_i)\right]\left\{\tau_{0|c_t}\left[\sum_j k_{tj}\exp -\frac{1}{2}\left(\mathbf{h}_t^\top \mathbf{P}_{tj}\mathbf{h}_t - 2\mathbf{h}_t^\top \mathbf{p}_{tj}\right)\right] + \tau_{1|c_t}\beta_t^1\right\} \end{aligned}$$

where each Gaussian component from the α message is combined with each quadratic-exponential component from $\beta^0(\mathbf{h}_t)$ according to the standard case algorithm 2.3. Note that the frequent moment-to-canonical conversions required for the β message and forming the smoothed posterior mean that this procedure is computationally less stable and more expensive than correction based smoothing (Verhaegen and Van Dooren, 2002).

α-γ Smoothing

Since numerical stability is of such concern in the LDS, it is imperative to have a correction-based smoother for the reset LDS. There are two routes to achieve this: either we can use the run-length $\tilde{\alpha}$-$\tilde{\gamma}$ formalism, section 3.2.3, or apply the bracket smoother from section 3.2.4. Both are essentially equivalent and require that we have first computed the $\tilde{\alpha}$ messages. Deriving these smoothers is straightforward—the final bracket smoother recursion is given in algorithm 3.2.

3.4 Approximate Inference

Filtering has overall $O\left(T^2\right)$ complexity, meanwhile smoothing has $O\left(T^3\right)$ complexity. For long time-series $T \gg 1$, this can be prohibitively expensive, motivating a consideration of approximations.

Algorithm 3.1 RLDS Filtering for a model with parameters θ^0 (no reset) and θ^1 (reset).

```
$\{\mathbf{f}_1(\rho = 0), \mathbf{F}_1(\rho = 0), p_1\} \leftarrow \text{LDSForward}(\mathbf{0}, \mathbf{0}, \mathbf{y}_1; \theta^1)$    ▷ Initial reset case
$w_1(\rho = 0) \leftarrow p_1 \times p(c_1 = 1)$
$\{\mathbf{f}_1(\rho = 1), \mathbf{F}_1(\rho = 1), p_1\} \leftarrow \text{LDSForward}(\mathbf{0}, \mathbf{0}, \mathbf{y}_1; \theta^0)$    ▷ Initial non-reset case
$w_1(\rho = 1) \leftarrow p_1 \times p(c_1 = 0)$
$l_1 \leftarrow \sum w_1, w_1 \leftarrow w_1 / \sum w_1$    ▷ Likelihood, Normalise
for $t \leftarrow 2, T$ do
    $\{\mathbf{f}_t(\rho = 0), \mathbf{F}_t(\rho = 0), p_t\} \leftarrow \text{LDSForward}(\mathbf{0}, \mathbf{0}, \mathbf{y}_t; \theta^1)$    ▷ Reset case
    $w_t(\rho = 0) \leftarrow$
  $p_t \times \Big[p(c_{t+1} = 1 | c_t = 1)\, w_{t-1}(\rho_{t-1} = 0) + p(c_{t+1} = 1 | c_t = 0) \sum_{\rho_{t-1=1}}^{t-1} w_{t-1}(\rho_{t-1})\Big]$
    for $\rho \leftarrow 1, t$ do
        $\{\mathbf{f}_t(\rho), \mathbf{F}_t(\rho), p_t\} \leftarrow \text{LDSForward}(\mathbf{f}_{t-1}(\rho - 1), \mathbf{F}_{t-1}(\rho - 1), \mathbf{y}_t; \theta^0)$    ▷ Non-reset cases
        $w_t(\rho) \leftarrow p_t \times p(c_{t+1} = 0 | c_t = \mathbb{I}(\rho = 1))\, w_{t-1}(\rho_{t-1} = \rho - 1)$
    end for
    $l_t \leftarrow l_{t-1} \times \sum w_t, w_t \leftarrow w_t / \sum w_t$    ▷ Likelihood, Normalise
end for
```

3.4.1 Approximate Filtering

The α message (or equivalently, the $\tilde{\alpha}$ message) is constructed as a mixture of distributions—in the LDS case, a mixture of Gaussian components. It is therefore easy to motivate any reduced component mixture approximation technique (Titterington et al., 1985). My implementation uses the $\tilde{\alpha}$ formalism, and to approximate I simply retain the M components with largest weight, reducing the forward pass to $O(MT)$. That is, I rank the $\tilde{\alpha}(h_t, \rho_t)$ components by the weight $\tilde{\alpha}(\rho_t)$, and retain only the ρ_t with largest weight, a form of Assumed Density Filtering.

3.4.2 Approximate $\tilde{\alpha}$-$\tilde{\beta}$

For smoothing, a naïve approximate algorithm is to drop components from the β message (or equivalently, the $\tilde{\beta}$ message) according to the weights of the components in the β message mixture. However, the β message components in themselves are not of interest and dropping components based on low β weight gives generally poor performance. When the α (or $\tilde{\alpha}$) and β (or $\tilde{\beta}$) messages are combined, the result is the smoothed posterior. In general, the weights of these smoothed components are functions not just of the weights from the α and β messages, but of all parameters in the messages. The relationship between those parameters and the resulting component weights can be complex (the case of the linear dynamical system is shown in algorithm 2.3).

It is possible, however, to motivate an approximation by observing the bracket smoothing results of section 3.2.4. First, by noting that whatever algorithm we choose to implement (α-β, $\tilde{\alpha}$-$\tilde{\beta}$, α-γ, or $\tilde{\alpha}$-$\tilde{\gamma}$), we see that the resulting exact posterior has identical structure. In the bracket smoother, the pair (ρ_t, ς_t) specifies exactly when the previous and next resets occur, so this intuition can be applied to each

Algorithm 3.2 RLDS Bracket Correction Smoothing, with parameters θ^0 (no reset) and θ^1 (reset).

```
$x_T \leftarrow w_T$, $\mathbf{g}_T \leftarrow \mathbf{f}_T$, $\mathbf{G}_T \leftarrow \mathbf{F}_T$        ▷ Initialise to filtered posterior
for $t \leftarrow T-1, 1$ do
    $x_t(0:t, 2:T-t+1) \leftarrow x_{t+1}(1:t+1, 1:T-t)$        ▷ Non-reset cases
    for $\rho \leftarrow 0, t$ do
        $x_t(\rho, 1) \leftarrow p(c_{t+1} = 1 | c_t = \mathbb{I}(\rho = 0)) \times w_{t+1}(\rho)$        ▷ Reset cases
    end for
    $x_t(:, 1) \leftarrow x_t(:, 1) \times \sum x_{t+1}(0, :) / \sum x_t(:, 1)$        ▷ Normalise
    $\mathbf{g}_t(:, 1) \leftarrow \mathbf{f}_t$, $\mathbf{G}_t(:, 1) \leftarrow \mathbf{F}_t$        ▷ Copy filtered moments
    for $\rho \leftarrow 0, t; \varsigma \leftarrow 2, T-t+1$ do        ▷ Calculate moments
        $\{\mathbf{g}_t(\rho, \varsigma), \mathbf{G}_t(\rho, \varsigma)\} \leftarrow$
                $\text{LDSBACKWARD}(\mathbf{g}_{t+1}(\rho+1, \varsigma-1), \mathbf{G}_{t+1}(\rho+1, \varsigma-1), \mathbf{f}_t(\rho), \mathbf{F}_t(\rho); \theta^0)$
    end for
end for
```

smoothing algorithm. In equation (3.13), we observed that the posterior mass transfers directly through the backward recursion in the no-reset case.

After calculating a full (exact or approximate) $\tilde{\alpha}$ recursion, we can approximate the $\tilde{\alpha}$-$\tilde{\beta}$ algorithm as follows. First, calculate a full $\tilde{\beta}$ message. Second, calculate the components corresponding to $\varsigma_t = 1$ given as $\tau_{1|c_t(\rho_t)} \tilde{\beta}_t^1 \tilde{\alpha}(h_t, \rho_t)$. Third, combine the $\tilde{\alpha}$ and $\tilde{\beta}$ messages for those components we know to have significant mass according to equation (3.13), corresponding to $\varsigma_t > 1$. Finally, renormalise the weights to form the approximate posterior. In this way, we limit the number of posterior components to N. The requirement of a full $\tilde{\beta}$ message, however, means the algorithm has $O\left(T^2\right)$ complexity.

3.4.3 Approximate $\tilde{\alpha}$-$\tilde{\gamma}$

By taking a component-dropping approach with the $\tilde{\alpha}$-$\tilde{\gamma}$ recursion, it is possible to derive a linear-time algorithm. To do this, first calculate an approximate $\tilde{\alpha}$ message retaining a constant number of components on each iteration. Second, calculate the components corresponding to $\varsigma_t = 1$: the weights are given by equation (3.14), and the moments by the $\tilde{\alpha}$ message. Third, calculate the moments for those components we know to have significant mass according to equation (3.13) corresponding to $\varsigma_t > 1$, with the correction smoother update. Finally, renormalise the weights to form the approximate posterior. This is equivalent to dropping components from $\tilde{\gamma}$ and limits the number of components calculated at each point to N, resulting in an algorithm with overall linear time-complexity.

3.4.4 Example: Reset Linear-Dynamical System

I implemented a linear-time algorithm for the reset LDS by limiting the number of $\tilde{\alpha}$ and $\tilde{\gamma}$ components, dropping lowest-weight components when the limit is exceeded, and compared the results with the quadratic-complexity $\tilde{\alpha}$-$\tilde{\beta}$ approximate implementation. To aid a direct comparison of methods, I also ran approximate $\tilde{\gamma}$ smoothing based on the exact filtered posterior since this has overall quadratic complexity

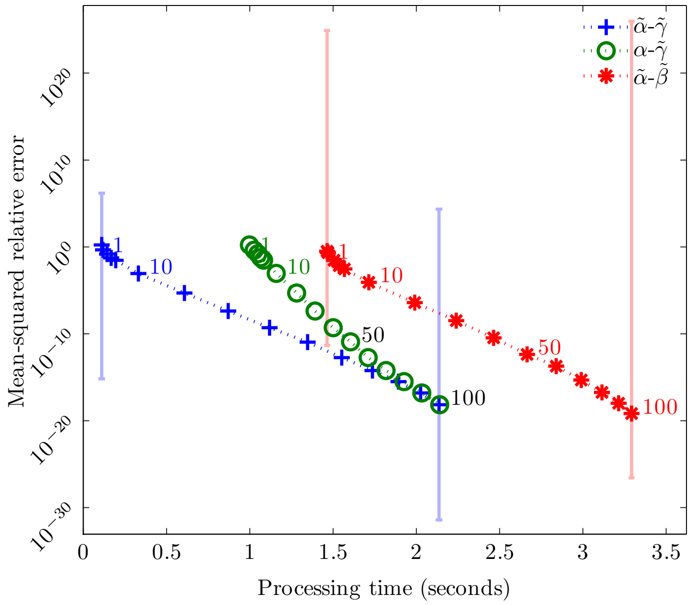


Figure 3.2: Comparison of approximation accuracy for the reset LDS. 1000 time-series ($T = 100$) were randomly generated using a single dimension for y_t and h_t. The graph shows the median error (compared with the exact correction-smoothed posterior) of the linear-time smoother based on approximate (blue) and exact filtering (green), and the quadratic-complexity $\tilde{\beta}$ smoother with approximate filtering (red), versus the mean running time. Error bars show max and min values. In each case, the points on the curve correspond to $N = 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ components in each approximate message. The error is calculated as $\text{mean}_t \left[\left(\langle h_t \rangle - \langle h'_t \rangle \right) / \langle h_t \rangle \right]^2$.

comparable with the $\tilde{\beta}$ routine. Results are shown in figure 3.2, in which I show how the runtimes and relative errors in the smoothed posteriors compare for different numbers of components.

I demonstrate the run-time reduction for different time-series lengths in figure 3.3.

3.4.5 Discussion

The mixture distributions comprising the messages need to be approximated to reduce the complexity of inference in a reset model. In the case of the reset LDS the messages are formed as Gaussian mixtures, and various approximation schemes are available, as some are set out in section 2.4. In this chapter, I simply dropped posterior components, which corresponds to a form of Assumed Density Filtering. This motivates a discussion of whether such scheme provides a stable approximation, and how to select the number of components to retain. Each retained posterior component corresponds, according to the bracket smoother, to a unique local partition of the time-series; in the worst case, each of the posterior components has equal mass. In this case, the discrete components of the filtering and smoothing messages correspond to little or no posterior belief about the probability of a reset at each point. Hence it may be fair to say that the model is not well suited to the data: a reparameterisation or different model may be appropriate. When considering the number of message components to retain, however, the 'cut-off' weight of the dropped components is known in each routine and can be used to conclude whether retaining more components

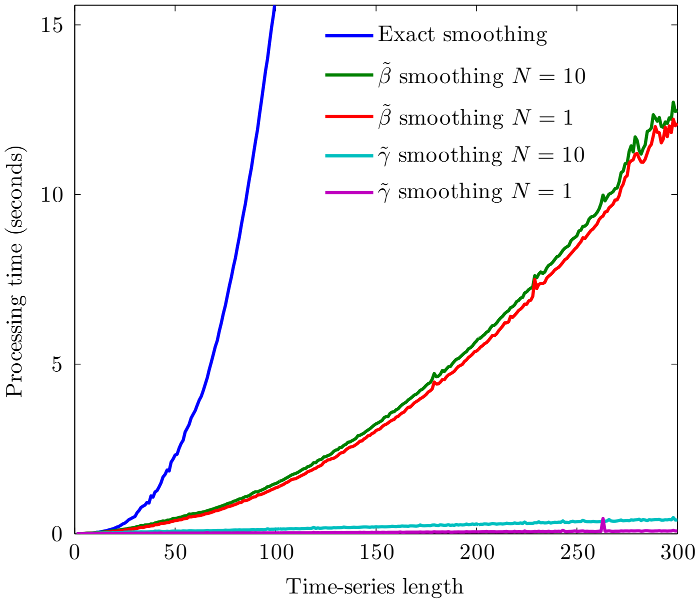


Figure 3.3: Median running time (10 iterations) for the reset LDS with variable time-series length.

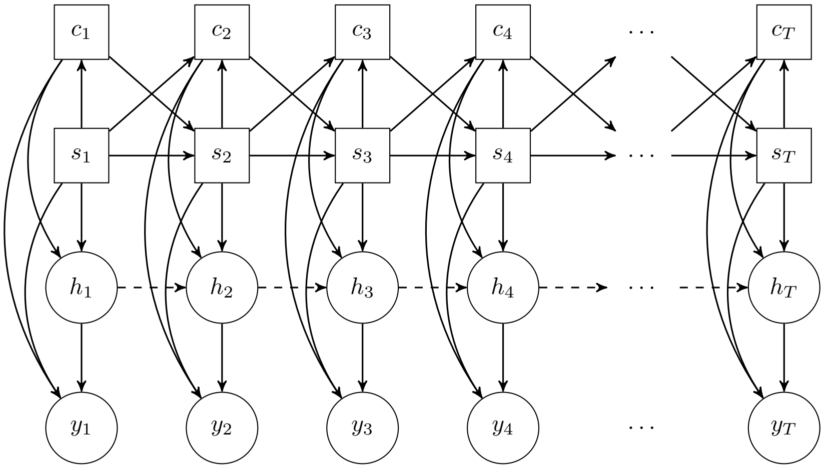


Figure 3.4: Switch-Reset model structure.

may be worth the computational expense.

The approximation routines are structured in a flexible way so as to allow different schemes to that used in my implementation. One example would be to only drop posterior components from messages when the mass of such components falls below a predetermined threshold, though this has the effect of increasing worst-case computational complexity. Finally, it should be noted that the smoothed posterior weights, calculated according to the bracket smoother and used in the $\tilde{\gamma}$ approximation, are calculated only from the filtered weights; so it is possible to conclude something about the number of smoothed components that may be reasonably dropped by filtering only.

3.5 Piecewise Reset Models

The recursions for the reset LDS are straightforward since the messages are closed within the space of the mixture of Gaussians (see appendix B). Other classes of model admit similar closure properties, and I briefly describe two such here based on the piecewise-constant assumption,

$$p(h_t|h_{t-1}, c_t) = \begin{cases} \delta(h_t - h_{t-1}) & c_t = 0 \\ p^1(h_t) & c_t = 1 \end{cases}$$

for which equation (3.11) is trivially rewritten

$$\tilde{\alpha}(h_t, \rho_t > 0) = p^0(y_t|h_t)\, p(\rho_t|\rho_{t-1} = \rho_t - 1)\, \tilde{\alpha}(h_{t-1} = h_t, \rho_{t-1} = \rho_t - 1) \tag{3.16}$$

and similarly for equation (3.15)

$$\tilde{\gamma}(h_t|\rho_t, \varsigma_t > 1) = \tilde{\gamma}(h_{t+1} = h_t|\rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1)\,.$$

Any model can be considered in which $p(y_t|h_t, c_t)$ and $p^1(h_t)$ are conjugate-exponential pairs. For example if we have a Gamma reset distribution $p^1(h_t) = \mathit{Gam}(h_t|\cdot, \cdot)$ and a Gaussian emission $p(y_t|h_t, c_t) = \mathcal{N}\big(y_t\big|0, h_t^{-1}\big)$, then the filtered and smoothed posteriors are mixtures of Gamma terms. Similarly, one can consider a piecewise-constant Poisson reset model in which the rate h_t is constant until reset from a Gamma distribution. The resulting posterior is a mixture of Gamma distributions (Barber and Cemgil, 2010). Bayesian priors over Gaussian mean and precision (for conjugacy, usually Gaussian and Gamma/Wishart respectively) fit readily into the piecewise-constant framework.

Example problems are well known for piecewise reset models, including the coal-mine disaster data of Jarrett (1979) and the well-logging data of Ó Ruanaidh and Fitzgerald (1996). I provide an example using the latter, using a Gaussian prior over the piecewise-constant mean of Gaussian data.

3.5.1 Piecewise-Constant Model: Well-Log Example

A piecewise reset model as described in this section, widely known simply as a change-point model, is implemented by specifying the reset-case latent prior $p^1(h_t)$, emission $p(y_t|h_t, c_t)$, and deriving the forward updates by appealing to equation (3.16) in the no-reset case and equation (3.10) in the reset case.

The well-logging data of Ó Ruanaidh and Fitzgerald (1996)[8] as described in section 2.2.2, form a noisy step function. Adams and MacKay (2007) used a Gaussian prior distribution over the piecewise-constant mean of the Gaussian-distributed data for filtered inference. I implemented such model with the smoothing approximation framework of this chapter, taking $h_t \to \mu_t$, and $p^0(\mu_t|\mu_{t-1}) = \delta(\mu_t - \mu_{t-1})$, using the parameters chosen by Adams and MacKay (2007): $p^1(\mu_t) = \mathcal{N}\big(\mu_t\big|1.15 \times 10^5, 10^8\big)$ and change-point probability $p(c_t = 1) = \frac{1}{250}$. Results are shown in figure 3.5.

[8] Obtained from Fearnhead and Clifford (2003).

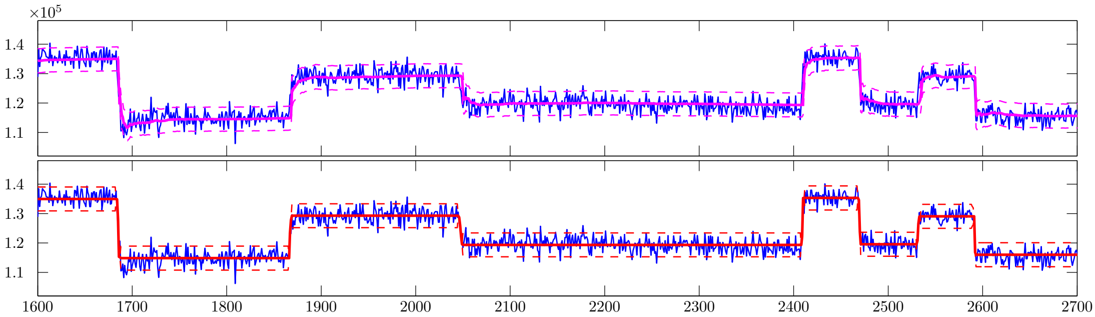


Figure 3.5: Window of the 4050-datum well-log data set, as shown by Adams and MacKay (2007). I show (i) the observed data overlaid with the filtered mean and *a posteriori* observation standard deviation; (ii) the smoothed equivalent. This example was run with $N = 10$ components comprising each approximate message.

3.6 Switch-Reset Models

The reset model defined above is limited to two kinds of dynamics—either continuation along the standard dynamics p^0 or the reset p^1. The switch-reset model enriches this by defining a set of S dynamical models,

$$p(h_t|h_{t-1}, c_t, s_t) = \begin{cases} p^0(h_t|h_{t-1}, s_t) & c_t = 0 \\ p^1(h_t|s_t) & c_t = 1 \end{cases}$$

$$p(y_t|h_t, c_t, s_t) = \begin{cases} p^0(y_t|h_t, s_t) & c_t = 0 \\ p^1(y_t|h_t, s_t) & c_t = 1 \end{cases}$$

with a Markov transition on the switch variables $p(s_t|s_{t-1}, c_{t-1})$. The reset is deterministically defined by $c_t = \mathbb{I}[s_t \neq s_{t-1}]$ and occurs with any change in the prevailing state s_t, see the belief network in figure 3.4. In the sense that the model incorporates multiple internal states selecting from different dynamics, the switch-reset model is more similar to a classic switching latent Markov model.

The intuition is that after a reset the model chooses from an available set of S dynamical models p^1. Another reset occurs if the state s_t changes from s_{t-1}. At that point the latent variable is reset, after which the dynamics continues. This is therefore a switching model, but with resets[9]. Inference in the class of switch-reset models is straightforward, and I give a derivation here. In the LDS case, the naïve correction approach runs into the same analytical difficulty as in section 3.2.2, and can be solved with a similar approach using run-length smoothing. The intuitive interpretation of the posterior components that was observed for the basic reset model transfers to the switching model, and the approximation schemes described above can be easily applied.

3.6.1 Switch-Reset Models: Inference

Marginal inference is straightforward based on extending the variable $h_t \to (h_t, s_t)$ and using the recursions in section 3.2. Briefly, I first define the equivalent run-length model, with transition

[9] Called a Reset-HMM in Barber and Cemgil (2010).

$p(\rho_t|s_t, s_{t-1}, \rho_{t-1})$. Then the filtered posterior is $p(h_t, s_t, \rho_t, y_{1:t}) = \tilde{\alpha}(h_t|s_t, \rho_t)\,\tilde{\alpha}(s_t, \rho_t)$. The discrete component updates according to

$$\tilde{\alpha}(s_t, \rho_t) = p(y_t|s_t, \rho_t) \sum_{s_{t-1}, \rho_{t-1}} p(\rho_t|s_t, s_{t-1}, \rho_{t-1})\, p(s_t|s_{t-1}, \rho_{t-1})\, \tilde{\alpha}(s_{t-1}, \rho_{t-1})$$

where we note $\rho_t \neq 0 \Rightarrow (\rho_{t-1} = \rho_t - 1) \wedge (s_{t-1} = s_t)$. This shows that discrete filtered distribution scales as $O\left(S^2T^2\right)$. The continuous component is calculated using standard forward propagation, conditioned on $\rho_{1:t}, s_{1:t}$. For smoothing, one may apply either the $\tilde{\alpha}$-$\tilde{\gamma}$ approach, or use bracketing. For bracketing, in the no-reset case the mass transfers directly,

$$p(s_t, \rho_t, \varsigma_t|y_{1:T}) = p(s_{t+1} = s_t, \rho_{t+1} = \rho_t + 1, \varsigma_{t+1} = \varsigma_t - 1|y_{1:T})$$

and for the reset case $\varsigma_t = 1 \Leftrightarrow s_{t+1} \neq s_t \Leftrightarrow \rho_{t+1} = 0$, so $p(s_t, \rho_t, \varsigma_t = 1|y_{1:T})$ is given by

$$p(s_t, \rho_t|y_{1:t}) \sum_{s_{t+1} \neq s_t} \frac{p(s_{t+1}|s_t, \rho_t)}{p(s_{t+1}, \rho_{t+1} = 0|y_{1:t})} \sum_{\varsigma_{t+1}} p(s_{t+1}, \rho_{t+1} = 0, \varsigma_{t+1}|y_{1:T})$$

with (for $s_{t+1} \neq s_t$)

$$p(s_{t+1}, \rho_{t+1} = 0|y_{1:t}) = \sum_{s_t \neq s_{t+1}, \rho_t} p(s_{t+1}|s_t, \rho_t)\, p(s_t, \rho_t|y_{1:t})\,.$$

This gives an overall $O\left(S^2T^3\right)$ complexity for smoothing. The continuous component of the smoothed posterior $p(h_t|s_t, \rho_t, \varsigma_t, y_{1:T})$ is calculated by standard smoothing on the bracket.

3.6.2 Switch-Reset LDS: Generated Example

I implemented a switch-reset LDS using the linear-time $\tilde{\alpha}$-$\tilde{\gamma}$ smoother. I first give a generated example of this model, as an experiment for which the truth of the state mass is known. The results are shown in figure 3.6, in which we observe that the approximate posterior tends to the exact posterior as the number of components increases. As can be seen, good results can be obtained based on using a very limited number of message components (10) compared to the number required to perform exact smoothing $(10, 100)$. Intuitively, the reason is that in the exact case, information is kept for filtering and smoothing time t from the whole sequence before and after t.

3.6.3 Switch-Reset LDS: Audio Example

In figure 3.7 I apply the model to a short speech audio signal[10] of $10,000$ observations. For these data, the latent variable $\mathbf{h}_t$ is used to model the coefficients of autoregressive lags, and I assume each observation $y_t = \sum_{m=1}^{6} h_t^m y_{t-m} + \epsilon_t$. Compared with a standard hidden Markov model in which a set of fixed autoregressive coefficients is used, this example provides a rich model in which the coefficients are free to evolve between state changes as set out in section 2.2.4.

Using my MATLAB code on a 2.4GHz machine, filtering took less than 400 seconds and subsequent smoothing less than 200 further; in the exact case, however, the problem is intractable needing the

[10] These data are from the TIMIT database.

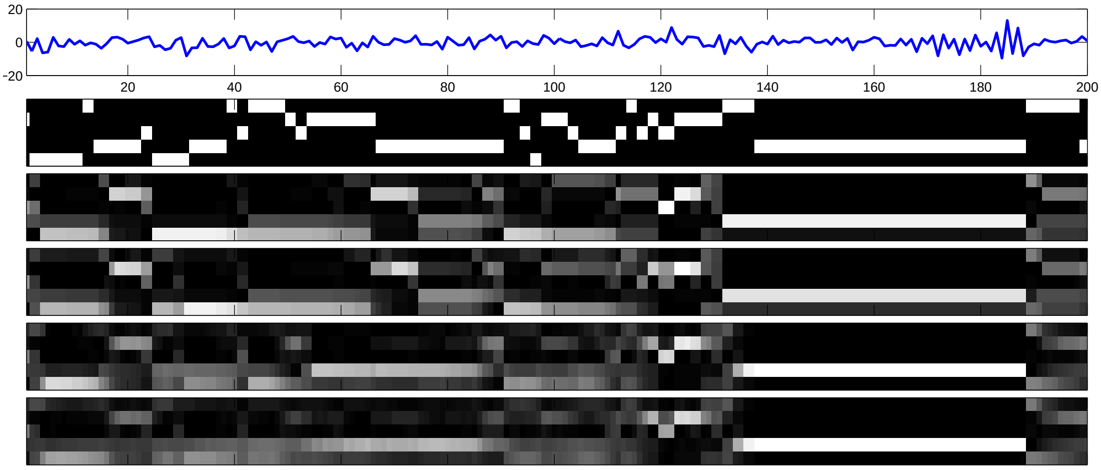


Figure 3.6: Switch-Reset LDS example. I generated a single-dimensional timeseries with $S = 5$ states, $T = 200$, and two-dimensional latent dynamics using random parameters. From top to bottom, I show (i) the generated signal; (ii) the generated state mass; (iii)-(v) the mass of the approximate smoothed posterior of each state using 1, 2 and 10 components; and (vi) the exact case, which contains a maximum of $10, 100$ components.

moments of some $O\left(10^{11}\right)$ Gaussian components (each of dimension 6) for the smoothed posterior in total, for each state. The model is highly sensitive to the state parameters, and I performed a very simple manual search of the parameter space and considered the likelihood (l_T) surface[11], with states broadly corresponding to 'growing', 'steady', and 'reducing' signal magnitudes by considering the sum of the autoregressive coefficients. The results show clear state switches between different phonemes, and each phoneme corresponds to a different (combination of) states in the smoothed posterior.

3.7 Conclusion

This chapter has considered inference in a class of dynamical model characterised by latent internal state resets. Whilst exact filtering in such a model scales as $O\left(T^2\right)$ for a sample of length T, exact smoothing has $O\left(T^3\right)$ complexity making exact inference costly for longer sequences.

To overcome this problem I employed a deterministic approximation strategy based on a form of Assumed Density Filtering, by dropping low-weight components from the resulting mixture distributions. The final algorithm shows linear time-complexity. The algorithm was demonstrated with example datasets and appears to perform well.

Finally, I extended the platform to a form of switching model in which internal state changes correspond to latent process resets, characterising the switch-reset class of models. Examples of the model were shown, appealing to the approximation schemes for inference set out earlier in the chapter.

[11] It is possible to use maximum likelihood learning techniques such as expectation maximisation.

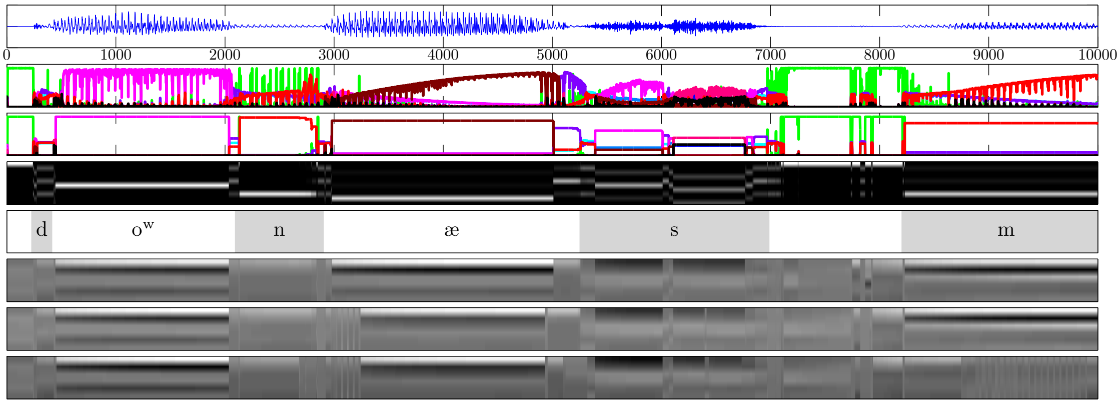


Figure 3.7: Switch-Reset LDS example with a short speech signal. I assumed that the signal is autoregressive of order 6, and used the switch-reset LDS to model the autoregressive coefficients with $S = 10$ different states, with $N = 10$ components retained in each message. From top to bottom, I show (i) the audio signal; (ii) the filtered posterior of each state; (iii)-(iv) the smoothed posterior state mass; (v) the main IPA phonemes comprising the sound; (vi) the mean value of the inferred autoregressive coefficients; (vii) also with $N = 5$; and (viii) with $N = 2$.

CHAPTER 4

Heteroskedastic Linear-Dynamical System

Switching models require many parameters to be specified, and in a switching linear dynamical system, the state posterior may be particularly sensitive to the selection of variance parameter. In this chapter, I show how to detect and model variance regimes in an enhanced linear dynamical system by application of reset model inference. I first show how to take a Bayesian approach to the variance of a linear-dynamical system by placing a Gamma prior over the precision of the sequence, and consider the inference recursions. Second, by placing the prior over piecewise-constant variance segments, I show a form of switching linear dynamical system, but without the need to specify multiple variance parameters a priori. I illustrate with bee-tracking data and finance data. The linear-complexity algorithm of chapter 3 is applicable.

4.1 Introduction

The linear-dynamical system, like many models used in practice, is based on the assumption of Gaussian-distributed variables. The Gaussian distribution is commonplace for a number of reasons, including the result of the central limit theorem that states that under certain conditions, statistics over sufficiently-large samples are approximately Gaussian-distributed. For the linear-dynamical system, there is also the convenient result that inference based on linear-Gaussian dynamics is possible cheaply in closed form by updating the sufficient statistics of a Gaussian message, as set out in section 2.2.1.

The linear-Gaussian dynamics of the linear-dynamical system are specified with the parameters of the transition and emission distributions; these comprise the transition and emission matrices, along with parameters for the additive mean and variance of the Gaussian noise. This is a significant number of parameters, and when one is interested in a switching linear-dynamical system with S states, these parameters need to be specified in respect of each of the S states.

4.1.1 Parameter Estimation

Finding values for the parameters is a key problem for any model, and it is common to use likelihood-maximisation techniques as discussed in section 2.4.4. For the linear-dynamical system and the switching linear-dynamical system, the Expectation Maximisation algorithm is applicable[1]. An alternative method based on the properties of a deterministic linear-dynamical model known as 'subspace identification' (Barber, 2012) may also be used, though this technique may only be used to initialise the parameters before starting the EM algorithm. The likelihood function, in particular for switching models in which states must be specified for each of the S internal states, is high-dimensional and there may be a high number of local maxima—it is in general a difficult problem to find the global maximum and EM may not always converge to the same estimates for different initialisations.

Because the parameter space is high-dimensional, it is interesting to consider how states with unknown characteristics can be defined *a priori*, limiting the number of parameters needed without removing the key benefits of the switching model.

In particular, the posterior distributions (particularly in respect of which of the S system states prevails at each point) can be overwhelmingly influenced by the magnitude of noise parameters of the predefined states. Attempts have been made to place Bayesian priors over the switching model parameters including those of Fox et al. (2008) and Barber and Chiappa (2006), although extending the switching model in such ways can add significantly to the complexity of the overall approach. It is desirable, therefore, to consider ways in which the need to specify the emission noise for each state can be relaxed. An obvious approach is to consider a Bayesian prior for the emission noise.

4.1.2 Variance Modelling

Initially, I seek to place a prior distribution over the variance, or precision, of a Gaussian distribution. Fortunately, a conjugate prior is available for the precision of a Gaussian distribution in the form of a Gamma distribution. In the following section, I derive an important known result showing conjugacy of the Gamma distribution for the precision of a Gaussian variable, and also show that the marginal of the variable after integrating away the variance forms a student-t distributed variable.

The student's t distribution normally arises as the distribution for an average of observations in a sample of Gaussian-distributed values. However, we may also consider occasions when the student's t distribution may arise in data, and a good example comes in the field of finance—Fergusson and Platen (2006) showed that the student's t distribution fits very well to the returns of stock prices. Tipping and Lawrence (2005) describe the use of student-t noise models to overcome the drawbacks of Gaussian modelling where the data have high kurtosis (a heavy-tailed distribution) due to outliers; the work relies on variational

[1] Note that due to the complexity of inference in the switching linear-dynamical system only approximate posteriors may be available, restricting the efficacy of the EM algorithm. Variational learning techniques may be used.

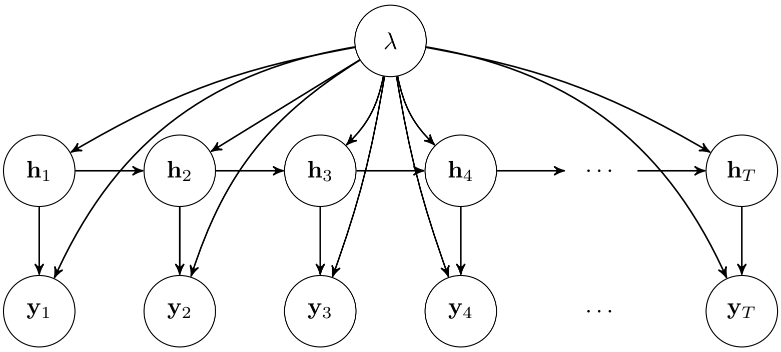


Figure 4.1: Belief network for the LDS with Bayesian precision.

approximation to characterise the posteriors, but provides a clear rationale for student-t noise models and describes the characterisations of the distribution.

There are analytical difficulties in modelling a transition for Gaussian variance with closed-form inference, in so-called switching stochastic volatility models. Carvalho and Lopes (2007) explain some of the recent approaches and note that solutions range from deterministic approximations to (sequential) sampling methods. For example, Dikmen and Cemgil (2008) describe a scheme of Markov chains in inverse Gamma-distributed random variables for variance modelling, and noting that exact inference is infeasible, appeal to sampling approximations.

The contribution of this chapter is to develop an augmented linear-dynamical model in which precision is given a prior distribution. I show that inference in the model can be made analytically tractable, and by working with the reset/change-point framework of chapter 3, I show how the model can be extended to detect variance regimes in observed data. The contribution of this chapter is not directly comparable with Tipping and Lawrence (2005), which does not model a latent dynamic state, and Dikmen and Cemgil (2008), which does not consider variance regimes, rather preferring dynamical variance. By contrast to previous works, this chapter contributes a combined model for inference with a latent state and piecewise-constant variance regimes, and consistent with the rest of the thesis relies on the mildest analytical approximation to the mathematics of exact inference when exact inference is not possible.

4.2 Model inference

I first derive analytical results relating to the conjugacy of the Gamma distribution for the precision of Gaussian variables. Thereafter, I move on to describe a linear-dynamical model with Bayesian precision.

4.2.1 Bayesian Precision

It is well-known that the Gamma distribution provides a conjugate prior for precision λ in a Gaussian variable $\mathbf{x}$ with $p(\mathbf{x}|\lambda) = \mathcal{N}\big(\mathbf{x}\big|\boldsymbol{\mu}, \lambda^{-1}\boldsymbol{\Sigma}\big)$. This result follows by setting a prior on the precision

$p(\lambda) = \mathit{Gam}(\lambda|a,b)$ and considering the joint density function $(\dim \mathbf{x} = d)$

$$\begin{aligned}p(\mathbf{x}|\lambda)\,p(\lambda) &= \mathcal{N}\big(\mathbf{x}\big|\boldsymbol{\mu},\lambda^{-1}\boldsymbol{\Sigma}\big)\,\mathit{Gam}(\lambda|a,b)\\ &= \frac{\sqrt{\lambda^d}}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}}\exp -\frac{\lambda}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\times\frac{b^a\lambda^{a-1}\exp -b\lambda}{\Gamma(a)}\\ &\qquad\qquad = \frac{b^a\lambda^{a+\frac{d}{2}-1}}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}\,\Gamma(a)}\exp -\left[b+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]\lambda\end{aligned}$$

and by collating terms in λ, we see a Gamma posterior for λ,

$$\begin{aligned}p(\mathbf{x}|\lambda)\,p(\lambda) &= \mathit{Gam}\left(\lambda\middle|a+\frac{d}{2},b+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)\\ &\qquad\times\frac{\Gamma\big(a+\frac{d}{2}\big)}{\left[b+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{a+\frac{d}{2}}}\times\frac{b^a}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}\,\Gamma(a)}.\end{aligned}$$

The remaining terms form the marginal for $\mathbf{x}$,

$$\begin{aligned}&\frac{\Gamma\big(a+\frac{d}{2}\big)}{\left[b+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{a+\frac{d}{2}}}\times\frac{b^a}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}\,\Gamma(a)}\\ &= \frac{\Gamma\big(\frac{2a+d}{2}\big)}{\Gamma(a)}\frac{1}{\sqrt{\det 2\pi b^{-1}\boldsymbol{\Sigma}}}\left[1+\frac{1}{2b}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{-\frac{2a+d}{2}}\\ &\qquad\qquad = \mathit{Student}\left(\mathbf{x}\middle|\boldsymbol{\mu},\frac{b}{a}\boldsymbol{\Sigma},2a\right).\end{aligned}$$

The derivation therefore shows the result of conditioning λ on $\mathbf{x}$,

$$\begin{aligned}p(\mathbf{x}|\lambda)\,p(\lambda) &= \mathcal{N}\big(\mathbf{x}\big|\boldsymbol{\mu},\lambda^{-1}\boldsymbol{\Sigma}\big)\,\mathit{Gam}(\lambda|a,b)\\ &= \underbrace{\mathit{Gam}\left(\lambda\middle|a+\frac{d}{2},b+\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right)}_{p(\lambda|\mathbf{x})}\underbrace{\mathit{Student}\left(\mathbf{x}\middle|\boldsymbol{\mu},\frac{b}{a}\boldsymbol{\Sigma},2a\right)}_{p(\mathbf{x})}.\end{aligned}\tag{4.1}$$

From this we can see that the student's t distribution can be considered as an infinite average of Gaussian distributions with unknown variance.

4.2.2 Linear Dynamical System with Bayesian Precision

The linear-dynamical system is described in section 2.2.1 as a latent Markov model with transition and emission distributions set out respectively in equations (2.6) and (2.7). I aim to place a prior over the precision of the observations in the linear-dynamical system. Ideally, this would involve simply rewriting equation (2.7) as

$$p(\mathbf{y}_t|\mathbf{h}_t,\lambda) = \mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{h}_t,\lambda^{-1}\mathbf{R}\big).$$

However, taking this step causes the derivations for closed form inference to break down since the product of Gaussians in the conditioning step equation (2.8) has a λ term buried only in the marginal variance of $\mathbf{y}_t$, and we cannot make use of the conditioning result equation (4.1). However if we additionally rewrite equation (2.6) as

$$p(\mathbf{h}_t|\mathbf{h}_{t-1},\lambda) = \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{h}_{t-1},\lambda^{-1}\mathbf{Q}\big)$$

the covariance of joint $p(\mathbf{h}_t, \mathbf{y}_t)$ given as in equation (2.8) factorises nicely with respect to λ since each component of the covariance has scaled precision.

In this model λ acts to scale the variance of the transition and emission of the whole LDS sequence—a belief network is shown in figure 4.1.

Inference

This redefined linear-dynamical model should therefore permit closed-form inference routines.

Filtering. For this new model, the Gauss-Gamma distribution provides a conjugate form for $\alpha(\mathbf{h}_t, \lambda) \equiv p(\mathbf{h}_t, \lambda, \mathbf{y}_{1:t}) = p(\mathbf{h}_t|\lambda, \mathbf{y}_{1:t})\, p(\lambda|\mathbf{y}_{1:t})\, p(\mathbf{y}_{1:t}) \equiv \alpha(\mathbf{h}_t|\lambda)\, \alpha_t(\lambda)\, p(\mathbf{y}_{1:t})$. This follows as in equation (2.1) after replacing $h_t \to (\mathbf{h}_t, \lambda)$,

$$\begin{aligned}\alpha(\mathbf{h}_t, \lambda) &= \int_{\mathbf{h}_{t-1}} p(\mathbf{h}_t, \mathbf{h}_{t-1}, \mathbf{y}_t, \lambda, \mathbf{y}_{1:t-1})\\ &= \int_{\mathbf{h}_{t-1}} p(\mathbf{y}_t|\mathbf{h}_t, \lambda)\, p(\mathbf{h}_t|\mathbf{h}_{t-1}, \lambda)\, \alpha(\mathbf{h}_{t-1}, \lambda)\end{aligned}$$

In this model, assume $\alpha(\mathbf{h}_{t-1}, \lambda) = \mathcal{N}\big(\mathbf{h}_{t-1}\big|\mathbf{f}_{t-1}, \lambda^{-1}\mathbf{F}_{t-1}\big)\, \mathit{Gam}(\lambda|a_{t-1}, b_{t-1})\, p(\mathbf{y}_{1:t-1})$, and then we can integrate $\mathbf{h}_{t-1}$ in the dynamics update step,

$$\begin{aligned}\frac{\alpha(\mathbf{h}_t, \lambda)}{p(\mathbf{y}_{1:t-1})} &= \mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{h}_t, \lambda^{-1}\mathbf{R}\big)\, \mathit{Gam}(\lambda|a_{t-1}, b_{t-1})\\ &\quad\times \int_{\mathbf{h}_{t-1}} \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{h}_{t-1}, \lambda^{-1}\mathbf{Q}\big)\, \mathcal{N}\big(\mathbf{h}_{t-1}\big|\mathbf{f}_{t-1}, \lambda^{-1}\mathbf{F}_{t-1}\big)\\ &= \mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{h}_t, \lambda^{-1}\mathbf{R}\big)\, \mathit{Gam}(\lambda|a_{t-1}, b_{t-1})\, \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{f}_{t-1}, \lambda^{-1}\boldsymbol{\Sigma}_{\mathbf{h}}\big).\end{aligned}$$

The Gaussian terms can be conditioned as equation (2.8),

$$\begin{aligned}&\mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{h}_t, \lambda^{-1}\mathbf{R}\big)\, \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{f}_{t-1}, \lambda^{-1}\boldsymbol{\Sigma}_{\mathbf{h}}\big)\\ &= \underbrace{\mathcal{N}\Big(\mathbf{h}_t\Big|\mathbf{M}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right) + \mathbf{A}\mathbf{f}_{t-1}, \lambda^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{h}} - \mathbf{M}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{h}}^{\top}\right)\Big)}_{\alpha(\mathbf{h}_t|\lambda) = \mathcal{N}(\mathbf{h}_t|\mathbf{f}_t, \lambda^{-1}\mathbf{F}_t)} \underbrace{\mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \lambda^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\big)}_{p(\mathbf{y}_t|\lambda, \mathbf{y}_{1:t-1})}.\end{aligned} \tag{4.2}$$

Now we use the Gamma conditioning of equation (4.1) for the observation $\mathbf{y}_t$,

$$\begin{aligned}\frac{\alpha(\mathbf{h}_t, \lambda)}{p(\mathbf{y}_{1:t-1})} &= \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{f}_t, \lambda^{-1}\mathbf{F}_t\big)\, \mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \lambda^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\big)\, \mathit{Gam}(\lambda|a_{t-1}, b_{t-1})\\ &= \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{f}_t, \lambda^{-1}\mathbf{F}_t\big)\, \mathit{Student}\Big(\mathbf{y}_t\Big|\mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \tfrac{b_{t-1}}{a_{t-1}}\boldsymbol{\Sigma}_{\mathbf{y}}, 2a_{t-1}\Big)\\ &\quad\times \mathit{Gam}\Big(\lambda\Big|a_{t-1} + \tfrac{\dim \mathbf{y}_t}{2}, b_{t-1} + \tfrac{1}{2}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)\Big).\end{aligned}$$

This completes the filtering update derivations.

To summarise, in addition to the recursive updates for $\mathbf{f}_t$ and $\mathbf{F}_t$ as for the Kalman update given in equation (2.9) and equation (2.10), we have further updates for the Gamma parameters a_t and b_t for the posterior $\alpha_t(\lambda) \equiv \mathit{Gam}(\lambda|a_t, b_t)$ given by

$$a_t = a_{t-1} + \tfrac{\dim(\mathbf{y}_t)}{2} \tag{4.3}$$

$$b_t = b_{t-1} + \tfrac{1}{2}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right) \tag{4.4}$$

Algorithm 4.1 Gamma forward

```
function GammaUpdate($a$, $b$, $\mathbf{f}$, $\mathbf{F}$, $\mathbf{y}$)
    $\boldsymbol{\mu}_{\mathbf{h}} \leftarrow \mathbf{A}\mathbf{f} + \bar{\mathbf{h}}$, $\boldsymbol{\mu}_{\mathbf{y}} \leftarrow \mathbf{B}\boldsymbol{\mu}_{\mathbf{h}} + \bar{\mathbf{y}}$, $\boldsymbol{\Sigma}_{\mathbf{h}} \leftarrow \mathbf{A}\mathbf{F}\mathbf{A}^{\top} + \mathbf{Q}$, $\boldsymbol{\Sigma}_{\mathbf{y}} \leftarrow \mathbf{B}\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{B}^{\top} + \mathbf{R}$
    $a' \leftarrow a + \frac{1}{2}\dim \mathbf{y}$                                          ▷ Gamma parameters
    $b' \leftarrow b + \frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right)^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right)$
    $p' \leftarrow \frac{\Gamma(a')}{\Gamma(a)} \frac{1}{\sqrt{\det 2\pi b \boldsymbol{\Sigma}_{\mathbf{y}}}} \left[\frac{b'}{b}\right]^{-a'}$          ▷ Compute likelihood
    return $a'$, $b'$, $p'$
end function
```

and the likelihood

$$p(\mathbf{y}_{1:t}) = \mathit{Student}\Big(\mathbf{y}_t \Big| \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \tfrac{b_{t-1}}{a_{t-1}} \boldsymbol{\Sigma}_{\mathbf{y}}, 2a_{t-1}\Big)\, p(\mathbf{y}_{1:t-1})\,.$$

These updates define the filtering recursion for the linear-dynamical system augmented with Bayesian precision λ, in particular the Gamma update shown in algorithm 4.1 (the standard update algorithm 2.1 applies to the Gaussian component). This is an analytical result for exact inference in the augmented model.

Smoothing. In chapter 3, I noted the importance of the direct correction smoothing approach for the linear-dynamical system in order to maintain numerical stability: I therefore continue this approach here.

For the reverse recursion $\gamma(\mathbf{h}_t, \lambda) = p(\mathbf{h}_t, \lambda | \mathbf{y}_{1:T})$ as equation (2.4) we have

$$\begin{aligned}
\gamma(\mathbf{h}_t, \lambda) &= \int_{\mathbf{h}_{t+1}} p(\mathbf{h}_t, \mathbf{h}_{t+1}, \lambda | \mathbf{y}_{1:T}) = \int_{\mathbf{h}_{t+1}} p(\mathbf{h}_t | \mathbf{h}_{t+1}, \lambda, \mathbf{y}_{1:t})\, p(\mathbf{h}_{t+1}, \lambda | \mathbf{y}_{1:T}) \\
&= \int_{\mathbf{h}_{t+1}} \frac{p(\mathbf{h}_{t+1} | \lambda, \mathbf{h}_t)\, p(\mathbf{h}_t | \lambda, \mathbf{y}_{1:t})}{p(\mathbf{h}_{t+1} | \lambda, \mathbf{y}_{1:t})} \gamma(\mathbf{h}_{t+1}, \lambda)
\end{aligned}$$

and in this case, the 'dynamics reversal' term $p(\mathbf{h}_t | \mathbf{h}_{t+1}, \lambda, \mathbf{y}_{1:t})$ is conditioned on the value of λ; as a result the algebra follows exactly as for the simple linear-dynamical system. The Gauss-Gamma is therefore a conjugate form for the smoothing messages $\gamma(\mathbf{h}_t, \lambda)$, with the exception that there are in fact no further calculations to be done in respect of the Gamma component $p(\lambda | \mathbf{y}_{1:T}) = \mathit{Gam}(\lambda | a_T, b_T)$ since this is known from filtering only. The resulting recursion is exactly the same as for the original linear dynamical model given in algorithm 2.4.

4.2.3 Variance Regimes

The above derivations show that exact filtering and smoothing inference routines are applicable in the augmented linear-dynamical system with a Bayesian approach to modelling precision of both the transition and emission distributions. The variance scale parameter scales the precision of the whole sequence, and for inference routines, messages passed forward and backward across the series have the form of a Gauss-Gamma component for each time step.

The key concept in this chapter is the requirement to model regimes of different variance (corresponding to precision) across the series, in particular, where those different regimes are unknown *a priori*.

A conceptually simple way to achieve this is to place the Bayesian-variance linear-dynamical system into the framework of reset models set out in chapter 3. By doing this, one would permit regimes of variance across the observed sequence, with breaks in the variance parameter λ corresponding also to breaks (resets) in the latent Markov chain $\mathbf{h}_t$. A belief network is shown for this model in figure 4.2(a).

To implement this model, one simply has to take the routine for reset model filtering of section 3.2.3 and bracket smoothing set out in section 3.2.4 and apply the dynamics inference updates set out above. Each likelihood $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ is given as a student distribution, used to inform the posterior of a reset at t.

Since we may consider the problem of inference in such a heteroskedastic model by application of the change-point framework of chapter 3, I do not focus on the model any further in this chapter. Instead, I consider a related but formally different problem: permitting resets in the latent precision scale λ, but maintaining continuous dynamics in the latent variable $\mathbf{h}_t$.

Such a model appeals more closely to the aims of this chapter and to the characteristics of the switching linear dynamical system, in which the continuity in the dynamics of the latent chain $\mathbf{h}_{1:T}$ remains unbroken.

For this model, I assume piecewise-constant precision, and define a variable for the precision scale in respect of each observation given by λ_t. At each point, either the precision scale remains constant $\lambda_t = \lambda_{t-1}$, or λ_t is redrawn afresh from the prior distribution specified *a priori*, *Gam*$(\lambda_t|a^\star, b^\star)$.

I place this variance-regime model into the logical framework of the reset model set out in chapter 3 by defining binary reset variables at each point $c_t \in \{0, 1\}$, with $\lambda_t = \lambda_{t-1}$ if and only if $c_t = 0$. A belief network is shown in figure 4.2(b), which shows the contrast with the standard reset model shown in figure 4.2(a). As can be seen in the diagram, I seek to maintain continuity of dynamics in the latent variable chain $\mathbf{h}_{1:T}$.

Inference

I proceed to consider the problem of filtering and smoothing in the model with resets in the latent precision λ_t. As with chapter 3, I replace the reset indicator c_t with a forward run-length ρ_t,

$$\rho_t = \begin{cases} 0 & c_t = 1 \\ \rho_{t-1} + 1 & c_t = 0 \end{cases}$$

and make the model formally equivalent based on the run-length with transition

$$p(\rho_t|\rho_{t-1}) = \begin{cases} p(c_t = 1|c_{t-1} = 1) & \rho_{t-1} = 0, \rho_t = 0 \\ p(c_t = 1|c_{t-1} = 0) & \rho_{t-1} > 0, \rho_t = 0 \\ p(c_t = 0|c_{t-1} = 1) & \rho_{t-1} = 0, \rho_t = 1 \\ p(c_t = 0|c_{t-1} = 0) & \rho_{t-1} > 0, \rho_t = \rho_{t-1} + 1 \\ 0 & \text{otherwise.} \end{cases}$$

The precision scale has transition

$$p(\lambda_t|\lambda_{t-1},\rho_t) = \begin{cases} \delta\left(\lambda_t - \lambda_{t-1}\right) & \rho_t > 0 \\ \mathit{Gam}(\lambda_t|a^\star, b^\star) & \rho_t = 0 \end{cases}$$

and the precision-scale versions of the linear-dynamical transition and emission distributions given above are applicable,

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{h}_t,\lambda_t) &= \mathcal{N}\big(\mathbf{y}_t\big|\mathbf{B}\mathbf{h}_t, \lambda_t^{-1}\mathbf{R}\big) \\ p(\mathbf{h}_t|\mathbf{h}_{t-1},\lambda_t) &= \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{h}_{t-1}, \lambda_t^{-1}\mathbf{Q}\big)\,. \end{aligned}$$

Filtering. The formulation of the model follows the reset framework of chapter 3 with resets in the latent variable chain for λ_t. As with the filtering recursions for the standard reset model, an additional component is contributed to the message with each iteration so the number of components grows linearly with t, and these components are again indexed with the run-length variable ρ_t.

Considering the filtering recursion given in equation (2.1), we may write

$$\begin{aligned} \alpha(\mathbf{h}_t,\lambda_t,\rho_t) = p(\mathbf{y}_t|\mathbf{h}_t,\lambda_t) \int_{\mathbf{h}_{t-1}} & p(\mathbf{h}_t|\mathbf{h}_{t-1},\lambda_t) \\ & \times \sum_{\rho_{t-1}} p(\rho_t|\rho_{t-1}) \int_{\lambda_{t-1}} p(\lambda_t|\lambda_{t-1},\rho_t)\,\alpha(\mathbf{h}_{t-1},\lambda_{t-1},\rho_{t-1}) \end{aligned}$$

and by writing $\alpha(\mathbf{h}_{t-1},\lambda_{t-1},\rho_{t-1}) \equiv \alpha(\mathbf{h}_{t-1}|\lambda_{t-1})\,\alpha(\lambda_{t-1}|\rho_{t-1})\,\alpha(\rho_{t-1})$ we identify the two cases of a reset and no reset at t. When no reset occurs at t, indicated by index $\rho_t > 0$, we may integrate the previous value of the precision scale λ_{t-1} to write

$$\begin{aligned} \alpha(\mathbf{h}_t,\lambda_t,\rho_t > 0) = p(\mathbf{y}_t|\mathbf{h}_t,\lambda_t) \int_{\mathbf{h}_{t-1}} & p(\mathbf{h}_t|\mathbf{h}_{t-1},\lambda_t) \\ & \times p(\rho_t|\rho_{t-1} = \rho_t - 1)\,\alpha(\mathbf{h}_{t-1}|\lambda_{t-1} = \lambda_t)\,\alpha(\lambda_{t-1} = \lambda_t|\rho_{t-1})\,\alpha(\rho_{t-1}) \end{aligned}$$

and upon choosing conjugate forms for the components $\alpha(\mathbf{h}_t|\lambda_t)$ as Gaussian, $\alpha(\lambda_t|\rho_t)$ as Gamma, and $\alpha(\rho_t)$ a constant, the algebra follows as in the case of constant λ given in section 4.2.2,

$$\begin{aligned} \alpha(\mathbf{h}_t,\lambda_t,\rho_t > 0) = \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{f}_t, \lambda_t^{-1}\mathbf{F}_t\big)\, & \mathit{Student}\Big(\mathbf{y}_t\Big|\mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \tfrac{b_{t-1}}{a_{t-1}}\boldsymbol{\Sigma}_{\mathbf{y}}, 2a_{t-1}\Big) \\ \times\, \mathit{Gam}\Big(\lambda_t\Big|a_{t-1} + \tfrac{\dim \mathbf{y}_t}{2}, & b_{t-1} + \tfrac{1}{2}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)^\top \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)\Big) \\ & \times p(\rho_t|\rho_{t-1} = \rho_t - 1)\,\alpha(\rho_{t-1})\,. \end{aligned}$$

and the updates for each component can be easily understood.

However in the case of a reset at t, indicated by $\rho_t = 0$, we have

$$\begin{aligned} \alpha(\mathbf{h}_t,\lambda_t,\rho_t = 0) = p(\mathbf{y}_t|\mathbf{h}_t,\lambda_t) & \int_{\mathbf{h}_{t-1}} p(\mathbf{h}_t|\mathbf{h}_{t-1},\lambda_t) \\ \times \sum_{\rho_{t-1}} p(\rho_t = 0|\rho_{t-1})\, \mathit{Gam}(\lambda_t|a^\star, b^\star)\,\alpha(\rho_{t-1}) & \int_{\lambda_{t-1}} \alpha(\mathbf{h}_{t-1}|\lambda_{t-1})\,\alpha(\lambda_{t-1}|\rho_{t-1}) \end{aligned} \tag{4.5}$$

and the component in the previous precision scale λ_{t-1} remains. Unfortunately, because there is no way to remove the dependency on λ_{t-1} a closed-form inference update recursion is not available with conjugate α messages. I therefore seek a simple approximation by rewriting equation (4.5) as

$$\begin{aligned}\alpha(\mathbf{h}_t, \lambda_t, \rho_t = 0) &= p(\mathbf{y}_t|\mathbf{h}_t, \lambda_t) \int_{\mathbf{h}_{t-1}} p(\mathbf{h}_t|\mathbf{h}_{t-1}, \lambda_t) \\ &\times \sum_{\rho_{t-1}} p(\rho_t = 0|\rho_{t-1})\, \mathit{Gam}(\lambda_t|a^\star, b^\star)\, \alpha(\rho_{t-1})\, \alpha(\mathbf{h}_{t-1}|\lambda_{t-1} \equiv \lambda_t) \int_{\lambda_{t-1}} \alpha(\lambda_{t-1}|\rho_{t-1}) \end{aligned} \tag{4.6}$$

which effectively assumes for the purpose of the transition in $\mathbf{h}_t$ that the variance is scaled approximately by the new value of λ_t; the term in λ_{t-1} can then be integrated away.

On the assumption that the filtered component in $\mathbf{h}_t$ is a Gaussian, this approximation is equivalent to defining a transition $p(\mathbf{h}_t|\mathbf{h}_{t-1}, \lambda_{t-1}, \lambda_t)$ such that in the reset case,

$$\begin{aligned}p(\mathbf{h}_t|\lambda_{t-1}, \lambda_t, \mathbf{y}_{1:t-1}) &= \int_{\mathbf{h}_{t-1}} p(\mathbf{h}_t|\mathbf{h}_{t-1}, \lambda_{t-1}, \lambda_t)\, \mathcal{N}\big(\mathbf{h}_{t-1}\big|\mathbf{f}_{t-1}, \lambda_{t-1}^{-1}\mathbf{F}_{t-1}\big) \\ &= \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{A}\mathbf{f}_{t-1}, \lambda_t^{-1}\left(\mathbf{A}\mathbf{F}_{t-1}\mathbf{A}^\top + \mathbf{Q}\right)\big)\,.\end{aligned}$$

Of course other approximation approaches are possible, though for the rest of this chapter I focus on this method alone since as I shall now go on to show, the resulting recursions have a single common Gaussian component, and this is important for the case of higher dimension latent variable states.

We can complete the recursion by following equation (4.6),

$$\begin{aligned}\alpha(\mathbf{h}_t, \lambda_t, \rho_t = 0) &= \mathcal{N}\big(\mathbf{h}_t\big|\mathbf{f}_t, \lambda_t^{-1}\mathbf{F}_t\big)\, \mathit{Student}\Big(\mathbf{y}_t\Big|\mathbf{B}\mathbf{A}\mathbf{f}_{t-1}, \tfrac{b^\star}{a^\star}\mathbf{\Sigma}_\mathbf{y}, 2a^\star\Big) \\ &\times \mathit{Gam}\Big(\lambda_t\Big|a^\star + \tfrac{\dim \mathbf{y}_t}{2}, b^\star + \tfrac{1}{2}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)^\top \mathbf{\Sigma}_\mathbf{y}^{-1}\left(\mathbf{y}_t - \mathbf{B}\mathbf{A}\mathbf{f}_{t-1}\right)\Big) \\ &\times \sum_{\rho_{t-1}} p(\rho_t = 0|\rho_{t-1})\, \alpha(\rho_{t-1})\,.\end{aligned}$$

from which the updates for each component are clear.

Smoothing. For smoothing in this model, we write equation (2.4) as

$$\gamma(\mathbf{h}_t, \lambda_t, \rho_t) = \int_{\mathbf{h}_{t+1},\lambda_{t+1}} \sum_{\rho_{t+1}} \frac{p(\mathbf{h}_{t+1}, \lambda_{t+1}, \rho_{t+1}|\mathbf{h}_t, \lambda_t, \rho_t)\, \alpha(\mathbf{h}_t, \lambda_t, \rho_t)}{p(\mathbf{h}_{t+1}, \lambda_{t+1}, \rho_{t+1}|\mathbf{y}_{1:t})} \gamma(\mathbf{h}_{t+1}, \lambda_{t+1}, \rho_{t+1})$$

where the terms in the dynamics reversal are expanded to

$$\begin{aligned}&p(\mathbf{h}_{t+1}, \lambda_{t+1}, \rho_{t+1}|\mathbf{h}_t, \lambda_t, \rho_t, \mathbf{y}_{1:t}) \\ &\qquad \propto p(\rho_{t+1}|\rho_t)\, p(\lambda_{t+1}|\lambda_t, \rho_t)\, p(\mathbf{h}_{t+1}|\lambda_t, \mathbf{h}_t)\, \alpha(\mathbf{h}_t|\lambda_t)\, \alpha(\lambda_t|\rho_t)\, \alpha(\rho_t)\,.\end{aligned}$$

The non-reset case $\rho_{t+1} = \rho_t + 1$ follows intuitively from this as the terms in the dynamics reversal collapse since there is only one possible value of ρ_t in the denominator. In this case the dynamics matches exactly the case of the simple linear-dynamical system with Bayesian precision set out in section 4.2.2.

However, as with the filtering recursion, an approximation is needed for the reset case $\rho_{t+1} = 0$ of the smoothing derivation. For smoothing, this materialises as two distinct replacements of λ terms in the density: first in the dynamics reversal, and second in the smoothed posterior from $t+1$. For the posterior from $t+1$, I assume that

$$\int_{\lambda_{t+1}} \gamma(\mathbf{h}_{t+1}, \lambda_{t+1}, \rho_{t+1}) = \gamma(\mathbf{h}_{t+1} | \lambda_{t+1} = \lambda_t, \cancel{\rho_{t+1}})\, \gamma(\rho_{t+1})$$

and for the more complex case of the dynamics reversal, that for the denominator the dependency of the transition and smoothed posterior $\alpha(\mathbf{h}_t | \lambda_t)$ on the ρ_t-specific λ_t can be approximated by the individual λ_t of interest. On these assumptions, I write for the dynamics reversal

$$\frac{p(\rho_{t+1} = 0 | \rho_t)\, p(\mathbf{h}_t | \lambda_t, \mathbf{h}_{t+1})\, \alpha(\lambda_t | \rho_t)\, \alpha(\rho_t)}{\sum_{\rho'_t} p(\rho_{t+1} = 0 | \rho'_t)\, \alpha(\rho'_t)}.$$

Finally, I combine these derivations to write the final posterior in the form of section 2.2.1 with the conjugate forms for the components $\gamma(\mathbf{h}_t | \lambda_t)$ as a single Gaussian, $\gamma(\lambda_t | \rho_t)$ as a mixture of Gamma distributions, and $\gamma(\rho_t)$ a constant as

$$\gamma(\mathbf{h}_t, \lambda_t, \rho_t) = \mathcal{N}\Big(\mathbf{h}_t \Big| \overleftarrow{\mathbf{A}}\,(\mathbf{g}_{t+1} - \mathbf{A}\mathbf{f}_t) + \mathbf{f}_t, \overleftarrow{\mathbf{A}}\mathbf{G}_{t+1}\overleftarrow{\mathbf{A}}^\top + \mathbf{F}_t - \overleftarrow{\mathbf{A}}\mathbf{A}\mathbf{F}_t^\top\Big)$$
$$\times \left[\gamma(\lambda_t | \rho_{t+1} = \rho_t + 1)\, \gamma(\rho_{t+1} = \rho_t + 1) + \frac{p(\rho_{t+1} = 0 | \rho_t)\, \alpha(\rho_t)}{\sum_{\rho'_t} p(\rho_{t+1} = 0 | \rho'_t)\, \alpha(\rho'_t)} \alpha(\lambda_t | \rho_t)\, \gamma(\rho_{t+1} = 0)\right]$$

to complete the recursions.

In situations when exact inference does not scale easily such as for the model described above, a broad range of approximation techniques may be applied. For switching dynamical models such as the heteroskedastic linear-dynamical model considered in this chapter, often it can be difficult to devise schemes which are simple to implement and representative of the posterior—Barber (2006) discusses the complexity of approximation schemes for the switching linear-dynamical system in particular, and makes the key observation that sampling may suffer in high-dimensional latent spaces. This is my primary motivation for pursuing non-sampling methods, and to maintain the approach of the thesis of finite mixture approximation methods similar to that used in chapter 3; since so many approximation schemes are possible it is infeasible compare all such methods and so I am guided by the principles laid out by Barber (2006), in which it is argued that projection to finite mixture distributions is a suitable approach. This allows the application of the mildest analytical approximation to render the problem tractable, and to build on the insight gained earlier in the thesis.

The ‘fuzzy equality’ approach to inference here relies on the idea that even in the case of a reset, λ_t is a good approximate substitute for the value of λ_{t-1} and vice-versa, so far as the transition in $\mathbf{h}_t$ is concerned. I consider this approximation, however undesirable, to be reasonable on the assumption that (i) it encourages some degree of continuity in the value of λ_t in the event of a reset consistent with continually-evolving processes, and (ii) the approximation affects only the variance of the transition in $\mathbf{h}_t$, whilst the most obvious difference in variance is visible from the magnitude of the observations $\mathbf{y}_t$.

Approximate inference. In forming the variant reset model set out above in which the latent precision λ can reset at any time point, we have defined a change-point model. As is set out in full in chapter 3, inference in this model scales as $O\left(T^2\right)$ on the forward (filtering) pass and $O\left(T^3\right)$ for a full smoothed

Algorithm 4.2 Heteroskedastic LDS filtering

```
$(\mathbf{f}_1, \mathbf{F}_1) \leftarrow \text{LDSFORWARD}(\mathbf{0}, \mathbf{0}, \mathbf{y}_1)$    ▷ Initial Kalman update
$(\mathbf{a}_1[\rho = 0], \mathbf{b}_1[\rho = 0], p_1^\star) \leftarrow \text{GAMMAUPDATE}(a^\star, b^\star, \mathbf{f}_1, \mathbf{F}_1, \mathbf{y}_1)$    ▷ Reset Gamma
$(\mathbf{a}_1[\rho = 1], \mathbf{b}_1[\rho = 1], p_1) \leftarrow \text{GAMMAUPDATE}(a_0, b_0, \mathbf{f}_1, \mathbf{F}_1, \mathbf{y}_1)$    ▷ No reset
$\mathbf{w}_1[\rho = 0] \leftarrow p_1^\star \times p(c_1 = 1), \mathbf{w}_1[\rho = 1] \leftarrow p_1 \times p(c_1 = 0)$
$l_1 \leftarrow \sum \mathbf{w}_1, \mathbf{w}_1 \leftarrow \text{COLLAPSE}(\mathbf{w}_1)$    ▷ Likelihood; Approximate and normalise
for $t \leftarrow 2, T$ do
    $(\mathbf{f}_t, \mathbf{F}_t) \leftarrow \text{LDSFORWARD}(\mathbf{f}_{t-1}, \mathbf{F}_{t-1}, \mathbf{y}_t)$    ▷ Kalman update
    $(\mathbf{a}_t[\rho = 0], \mathbf{b}_t[\rho = 0], p_t^\star) \leftarrow \text{GAMMAUPDATE}(a^\star, b^\star, \mathbf{f}_t, \mathbf{F}_t, \mathbf{y}_t)$    ▷ Reset
    $\mathbf{w}_t[\rho = 0] \leftarrow p_t^\star \times \sum_{\rho_{t-1}} p(c_{t+1}{=}1|c_t{=}\mathbb{I}[\rho_{t-1} = 0])\, \mathbf{w}_{t-1}[\rho_{t-1}]$
    for $\rho \leftarrow$ find $(\mathbf{w}_{t-1}) + 1$ do    ▷ Non-reset cases
        $(\mathbf{a}_t[\rho], \mathbf{b}_t[\rho], p_t) \leftarrow \text{GAMMAUPDATE}(\mathbf{a}_{t-1}[\rho - 1], \mathbf{b}_{t-1}[\rho - 1], \mathbf{f}_t, \mathbf{F}_t, \mathbf{y}_t)$
        $\mathbf{w}_t[\rho] \leftarrow p_t \times p(c_{t+1} = 0|c_t = \mathbb{I}[\rho = 1])\, \mathbf{w}_{t-1}[\rho_{t-1} = \rho - 1]$
    end for
    $l_t \leftarrow l_{t-1} \times \sum \mathbf{w}_t, \mathbf{w}_t \leftarrow \text{COLLAPSE}(\mathbf{w}_t)$    ▷ Likelihood; Approximate and normalise
end for
```

posterior. For the heteroskedastic linear-dynamical system set out in this chapter, however, smoothing in fact scales as $O\left(T^2\right)$, which follows from the observation that no further calculations need to be done in respect of the Gamma components in the reverse pass, whilst there is only a single Gaussian component. In order to perform the inference routines over datasets of significant sizes as I show in the experiments section below, I apply the approximation scheme of chapter 3 to this model and retain only a fixed number of posterior components in each recursive step, reducing the complexity of the combined algorithm to linear in T.

The combined filtering algorithm for the heteroskedastic linear-dynamical system is given in algorithm 4.2, and the smoothing routine is given in algorithm 4.3.

4.3 Experiments

In this section I apply the heteroskedastic linear-dynamical system to two example datasets to demonstrate the applicability of the model with unknown variance regimes to real-world problems. These examples model multi-dimensional series as individual autoregressive processes of order 1 according to the formulation set out in section 2.2.4, which shows how the autoregressive coefficients can be modelled with the latent variable of a (linear-) dynamical model. In this approach, the autoregressive coefficients are free to evolve according to the dynamics of the model; in these examples, the heteroskedastic linear-dynamical system is used to allow the evolution of the autoregressive coefficients whilst modelling variance regimes in the data.

Algorithm 4.3 Heteroskedastic LDS smoothing

```
$\mathbf{X}_T \leftarrow \mathbf{w}_T$, $\mathbf{g}_T \leftarrow \mathbf{f}_T$, $\mathbf{G}_T \leftarrow \mathbf{F}_T$, $\mathbf{A}_T \leftarrow \mathbf{a}_T$, $\mathbf{B}_T \leftarrow \mathbf{b}_T$    ▷ Initialise to filtered posterior
for $t \leftarrow T-1, 1$ do
    $(\mathbf{g}_t, \mathbf{G}_t) \leftarrow \text{LDSBACKWARD}(\mathbf{g}_{t+1}, \mathbf{G}_{t+1}, \mathbf{f}_t, \mathbf{F}_t)$    ▷ RTS update
    $\mathbf{X}_t[0:t, 2:T-t+1] \leftarrow \mathbf{X}_{t+1}[1:t+1, 1:T-t]$    ▷ Non-reset weights
    $\mathbf{A}_t[0:t, 2:T-t+1] \leftarrow \mathbf{A}_{t+1}[1:t+1, 1:T-t]$    ▷ Copy Gamma params, no reset
    $\mathbf{B}_t[0:t, 2:T-t+1] \leftarrow \mathbf{B}_{t+1}[1:t+1, 1:T-t]$
    $\mathbf{A}_t[0:t, 1] \leftarrow \mathbf{a}_t$    ▷ Copy Gamma params, reset
    $\mathbf{B}_t[0:t, 1] \leftarrow \mathbf{b}_t$
    for $\rho \leftarrow \text{find}\,(\mathbf{w}_{t+1})$ do
        $\mathbf{X}_t[\rho, 1] \leftarrow p(c_{t+1} = 1 | c_t = \mathbb{I}[\rho = 0]) \times \mathbf{w}_{t+1}[\rho]$    ▷ Reset case weights
    end for
    $\mathbf{X}_t[\rho, 1] \leftarrow \text{NORMALISE}(\mathbf{X}_t[\rho, 1])$
    $\mathbf{X}_t \leftarrow \text{COLLAPSE}(\mathbf{X}_t)$    ▷ Approximate and normalise
end for
```

4.3.1 Bee Data

I first applied the heteroskedastic linear-dynamical system with variance regimes to the Bee 'waggle dance' data[2] first published by Oh et al. (2006), in which a switching linear-dynamical system was used. The data describe the motion characteristics of honey bees, which perform a 'waggle dance'—the key problem is to segment the observed data into regimes of turning and 'waggle' phases. These data consist of the position (x, y) of a bee, along with its orientation θ, automatically found from a video. As with earlier analysis of these data, I first pre-processed by replacing the angle θ with the pair $(\sin\theta, \cos\theta)$, and centred and scaled the data. By modelling the data as an autoregressive process order 1, I aimed to detect regions in which the bee performed the 'waggle' phase by using the heteroskedastic linear-dynamical system to model the 4-dimensional autoregression coefficient. Results from this experiment, shown in figure 4.3, compare favourably with the change-point analysis of the same data by Xuan and Murphy (2007), who sought to detect points of changing dependency structure in data by considering covariance—by contrast, the generative model here combines the change-point analysis with a generative scheme for the data, and models the latent state representing autoregressive coefficients. Change-points correspond to significant changes in variance regime, which appear to correspond to changing from the 'waggle dance' phase to a turning phase and vice-versa.

4.3.2 Finance Data

I also applied the model to 30 finance portfolios[3] also studied by Xuan and Murphy (2007). These data represent stocks from the NYSE, AMEX, and NASDAQ exchanges over the period 1st July 1963 to 29th July 2011, separated into 30 industry portfolios. I took the daily returns data and fitted an

[2] Downloaded from `http://www.cc.gatech.edu/~borg/ijcv_psslds`

[3] Obtained from `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

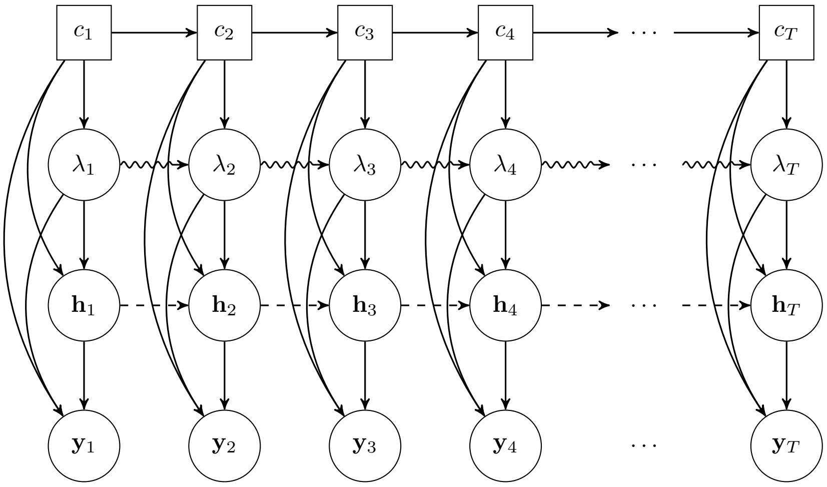
autoregressive model of order 1 using a 30-dimensional latent variable to model the autoregressive coefficients. In order to facilitate such analysis I first fitted a simple linear-dynamical system to these data using Expectation Maximisation, and used the resulting parameter estimates in the heteroskedastic linear-dynamical system—the results are shown in figure 4.4. Some insight into characteristics of financial markets can be considered in the results shown in figure 4.4(v). First of all, one can see the significance of the financial crises of 1973-1974 (worldwide stock market downturn following changes to monetary management systems, compounded by an oil crisis), 1998 (Russian debt default leads to collapse of Long-Term Capital Management), and 2008 (global financial crisis and recession). These collapses were accompanied by significant periods of high variance in market returns as shown by regimes of low precision λ_t. In addition the development of liquidity and integration in world-wide markets can be seen in the smooth and cyclical nature of variance regimes in price returns for more recent years, in particular in the posterior for λ_t over the period 1994-2011.

4.4 Conclusion

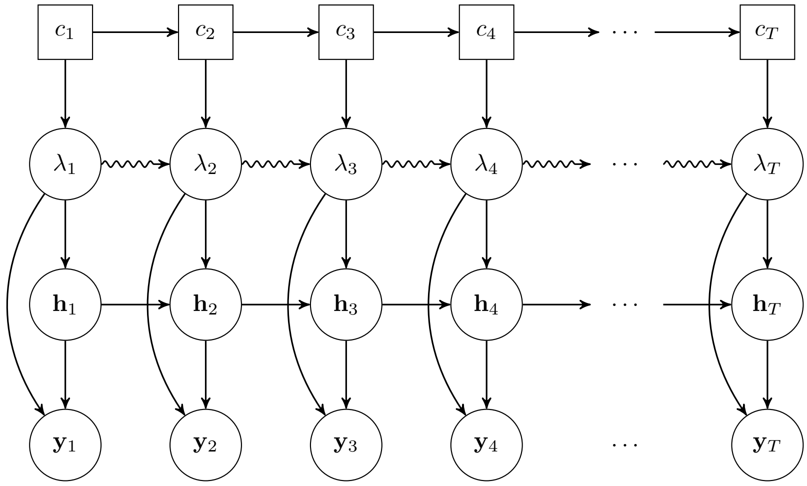
This chapter has introduced an approach to Bayesian modelling of the precision of observations in a linear-dynamical system. Exact inference in such a model remains tractable in closed form, and recursive update algorithms are given.

By appealing to the insights of the reset model framework developed in chapter 3, I went on to develop a heteroskedastic linear-dynamical model which permits resets in a process characterising an unknown precision variable, but seeks to maintain continuity in the dynamics of the latent state variable. Due to this latter feature, comparisons are drawn with the exponentially-complex switching linear-dynamical system.

Inference for the example applications was performed with the linear-time approximation scheme described in detail in chapter 3.



(a) Simple heteroskedastic linear-dynamical system based on the reset model framework of chapter 3. Here c_t also indicates whether the standard dynamics continues, $p^0(\mathbf{h}_t|\mathbf{h}_{t-1})$ ($c_t = 0$) or whether $\mathbf{h}_t$ is redrawn from the reset distribution $p^1(\mathbf{h}_t)$ ($c_t = 1$).



(b) Without resets in the latent Markov chain for $\mathbf{h}_t$.

Figure 4.2: Belief networks for the alternate versions of the heteroskedastic linear-dynamical system. The binary reset $c_t = 1$ indicates that precision scale λ_t is redrawn from the reset distribution $Gam(\lambda_t|a^\star, b^\star)$, otherwise it remains constant—indicated by a wavy edge.

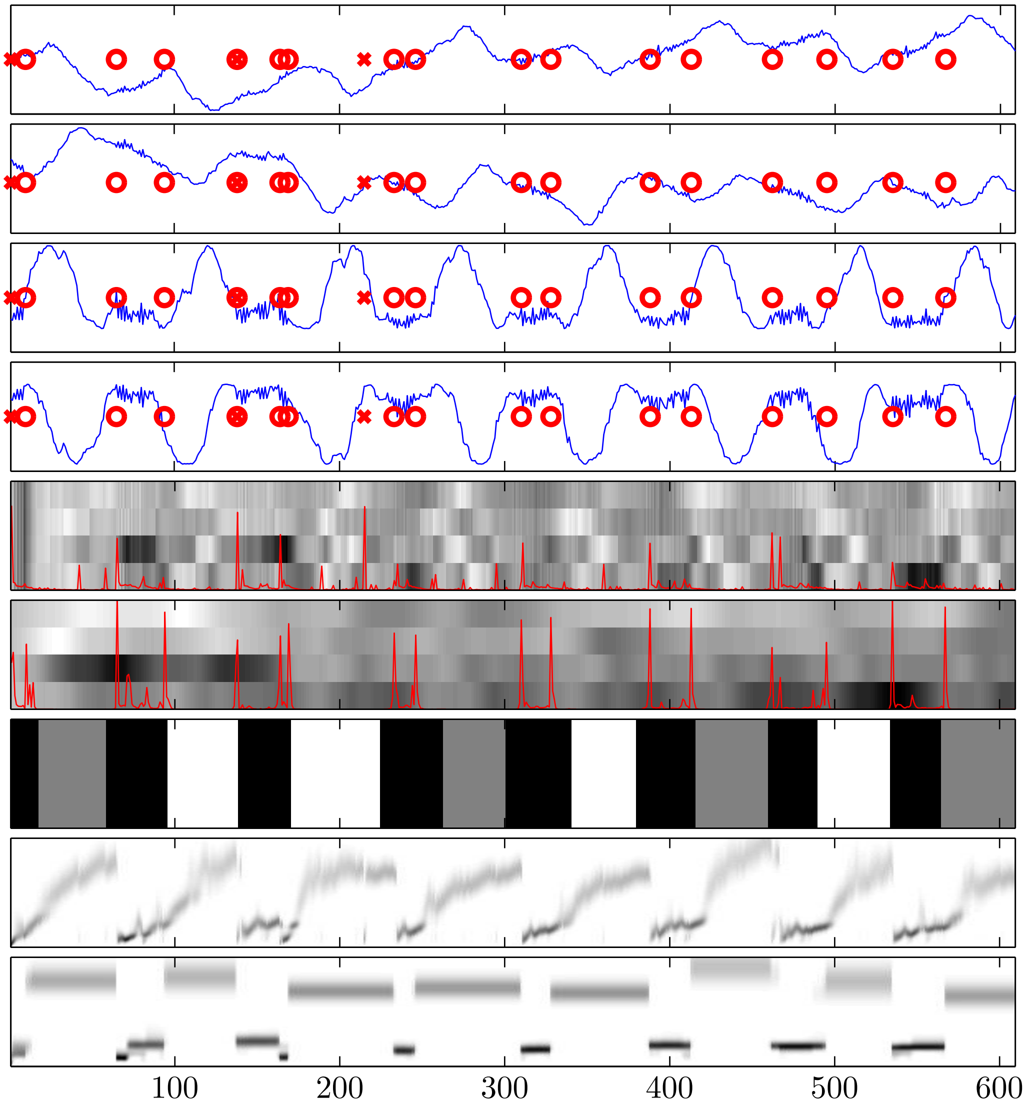


Figure 4.3: Results using the bee dance dataset, based on retaining 50 posterior components in each inference routine. I show (i)-(iv) the four data series, overlaid with x and o to represent the filtered and smoothed reset posteriors respectively; (v)-(vi) the mean autoregressive coefficients from filtering and smoothing respectively, overlaid with the reset posterior; (vii) the (hand-labelled) ground truth of the bee actions with the 'waggle' phase in black; and (viii)-(ix) the density of the mixture of Gammas for λ_t in the filtered and smoothed posteriors respectively—high λ_t corresponds to low variance.

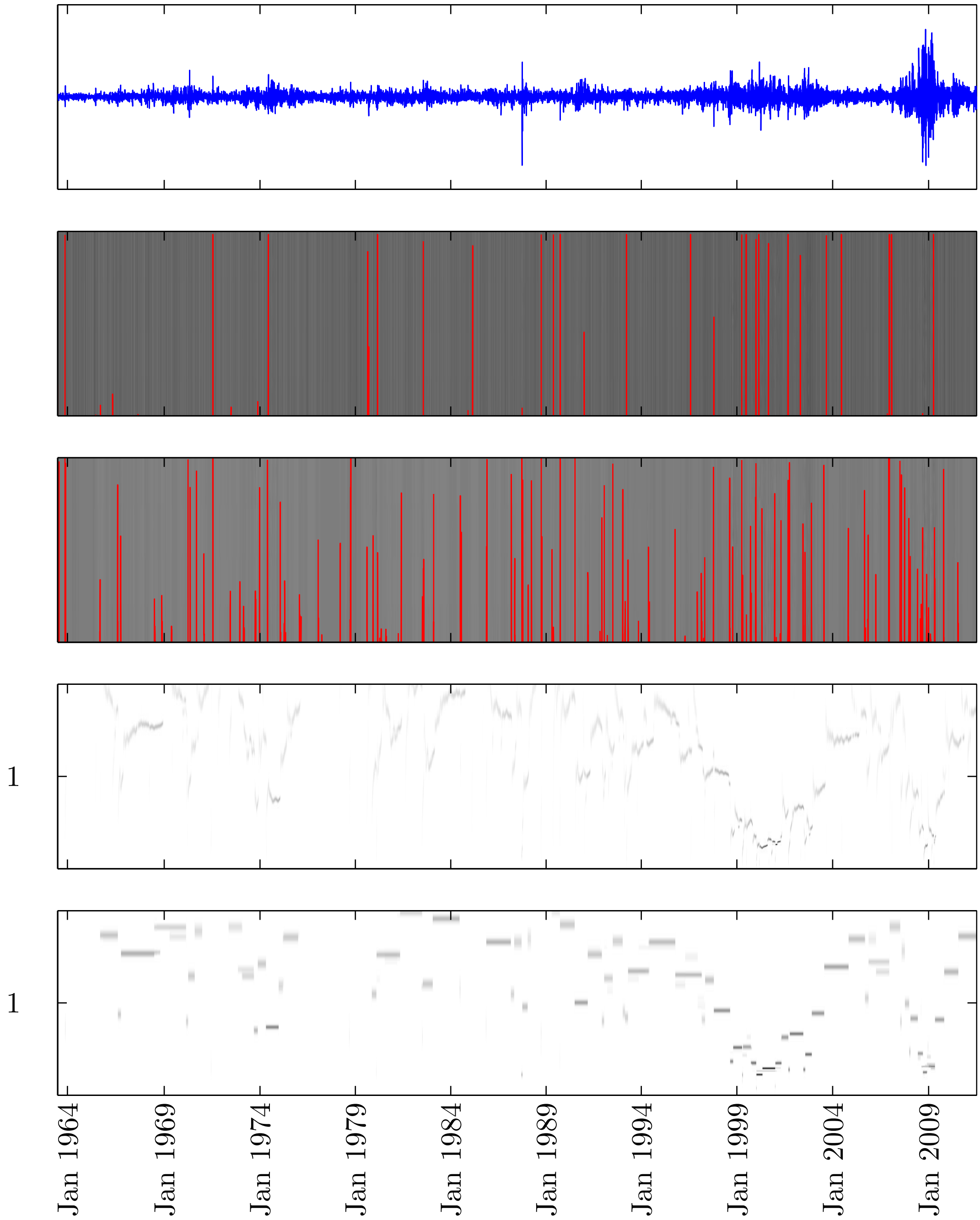


Figure 4.4: Results using French 30-portfolio finance dataset (this is 30-dimensional data with some $12,104$ observations), based on retaining 50 posterior components in each inference routine. I show top-to-bottom (i) an example of one of the 30 time series from the dataset representing the daily returns of a portfolio of finance stocks over the period 1st July 1963 to 29th July 2011; (ii-iii) the mean autoregressive coefficients from filtering and smoothing respectively overlaid with the posterior distribution of a reset; and (iv)-(v) the filtered and smoothed posterior densities of the precision scale λ_t (large λ_t corresponds to low observation variance).

CHAPTER 5

Bayesian Conditional Cointegration

Cointegration is an important topic for time-series, and describes a relationship between two series in which a linear combination is stationary. Classically, the test for cointegration is based on a two stage process in which first the linear relation between the series is estimated by Ordinary Least Squares. Subsequently a unit root test is performed on the residuals. A well-known deficiency of this classical approach is that it can lead to erroneous conclusions about the presence of cointegration. As an alternative, I present a coherent framework for estimating whether cointegration exists using Bayesian inference which is empirically superior to the classical approach. Finally, I apply the technique to model intermittent cointegration in which cointegration may exist only for limited time. In contrast to previous approaches this model makes no restriction on the number of possible cointegration segments.

The contribution of this chapter was published in Bracegirdle and Barber (2012).

5.1 Introduction

Cointegration, introduced in section 2.3, is an important concept in time-series analysis and relates to the fact that the differences between two time-series may be more predictable than any individual time-series itself. This chapter makes three contributions to this area by formulating a cointegration relationship as an instance of Bayesian inference in an associated probabilistic model, which provides a useful conceptual framework to describe cointegration. First, I show how a cointegration relationship may be estimated in the Bayesian model, and second, show how we may seek to determine whether such a relationship yields stationary residuals. The resulting regression estimation algorithm is shown to be more robust to spurious results than the classical approach of least-squares estimation set out in section 2.3. Thirdly, in practice two series may only be intermittently cointegrated as introduced in section 2.3.4—that is, they are only cointegrated over shorter segments of time. To date the identification of these segments has been attempted with rather limiting assumptions such as the presence of either only two or three segments (Gregory and Hansen, 1996; Hatemi-J, 2008). To address this I phrase the problem as inference in a corresponding changepoint model, which places no limitation on the number of segments and apply an inference scheme informed by the reset model framework of chapter 3.

5.1.1 Another Look At Cointegration

The classical two-stage approach of ordinary least-squares estimation and Dickey-Fuller testing makes potentially conflicting assumptions about the data: it is known that the regression estimation is a case of spurious regression if the residuals have a unit root, but this may be deemed the case by the subsequent unit root test; under $\mathcal{H}_0$, the regression is spurious. In practice, the estimation of a spurious regression may generally be avoided by appealing to economic theory in order to find a sensible rationale for the existence of any such relationship—this point is explained with some examples by Wooldridge (2009), but this ad-hoc approach is far from bulletproof.

As explained in section 2.2.3, OLS is known to provide a consistent estimator for the linear relationship even if the residuals are serially correlated. Watson and Teelucksingh (2002) discuss the strengths and weaknesses of the Engle-Granger two-step procedure, and note that whilst OLS is a consistent estimator, for small samples the bias can be significant since the strong exogeneity condition does not hold in the presence of serially-correlated residuals.

Furthermore, the notion of cointegration makes no statement about the relationship between x_t and ϵ_t, and whilst this does not affect the consistency of the OLS estimator, it is a problem for the unit root test: for the test statistic of the Dickey-Fuller step to have a t-distribution, x_t is required to be strictly exogenous—as explained by Wooldridge (2009), this assumption is too strong in general.

It is therefore of significant interest to consider alternative approaches to the problem of detecting cointegration between series.

The approach taken in this chapter seeks to model the conditional distribution $p(y_{1:T}|x_{1:T})$ while making no assumption about the underlying generating process of x_t. This approach renders the contribution of this chapter distinct from other Bayesian approaches, in particular those based on the vector error-correction representation of cointegration as set out in section 2.3.3 which assumes the processes can be modelled with an autoregressive model for $I(1)$ random walks.

5.2 Modelling Cointegration

In contrast to classical approaches, the aim of this chapter is to make a single cointegration model for which inference and learning can be carried out in a unified way, without the need for making potentially conflicting assumptions. The approach is to form a generative model of observations $p(y_{1:T}|x_{1:T}, \theta)$, where $\theta = \left\{\alpha, \beta, \sigma^2, \phi\right\}$. First, I develop intuition for the model by placing the cointegration equations

$$y_t = \alpha + \beta x_t + \epsilon_t$$

$$\epsilon_t = \phi\epsilon_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}\left(0, \sigma^2\right)$$

into a generative model based on observations $y_{1:T}$ and latent variables $\epsilon_{1:T}$ with density given by

$$p(y_{1:T}, \epsilon_{1:T} | x_{1:T}) = \prod_t p(y_t | x_t, \epsilon_t)\, p(\epsilon_t | \epsilon_{t-1})\,, \quad \epsilon_0 = \emptyset$$

where the emission is specified by a deterministic distribution

$$p(y_t | x_t, \epsilon_t) = \delta(y_t - \alpha - \beta x_t - \epsilon_t)$$

and the transition for ϵ_t is given as

$$p(\epsilon_t | \epsilon_{t-1}) = \mathcal{N}\big(\epsilon_t \big| \phi \epsilon_{t-1}, \sigma^2\big)\,.$$

The belief network for this model is given in figure 5.1. The marginal likelihood on the observations $y_{1:T}$ in this model is given by

$$p(y_{1:T} | x_{1:T}) = \int_{\epsilon_{1:T}} p(y_{1:T}, \epsilon_{1:T} | x_{1:T})\,.$$

The integration distributes,

$$p(y_{1:T} | x_{1:T}) = \int_{\epsilon_{1:T-1}} \prod_{t=1}^{T-1} p(y_t | x_t, \epsilon_t)\, p(\epsilon_t | \epsilon_{t-1}) \int_{\epsilon_T} p(y_T | x_T, \epsilon_T)\, p(\epsilon_T | \epsilon_{T-1})$$

and since we have a delta function,

$$\int_{\epsilon_T} p(y_T | x_T, \epsilon_T)\, p(\epsilon_T | \epsilon_{T-1}) = \mathcal{N}\big(y_T - \alpha - \beta x_T \big| \phi \epsilon_{T-1}, \sigma^2\big)\,.$$

Iterating this yields a product of Gaussian terms for the likelihood,

$$p(y_{1:T} | x_{1:T}) = \prod_{t=2}^{T} \mathcal{N}\big(y_t - \alpha - \beta x_t \big| \phi\,(y_{t-1} - \alpha - \beta x_{t-1})\,, \sigma^2\big)\,.$$

Finally, applying labels according to $\epsilon_t = y_t - \alpha - \beta x_t$,

$$p(y_{1:T} | x_{1:T}) = p(\epsilon_1) \prod_{t=2}^{T} \mathcal{N}\big(\epsilon_t \big| \phi \epsilon_{t-1}, \sigma^2\big) = p(\epsilon_{1:T})\,.$$

Hence mathematically the value of the likelihood $p(y_{1:T} | x_{1:T})$ is equivalent to the likelihood on the Markov chain with 'observations' ϵ_t shown in figure 5.2.

Ordinary least-squares estimation for the regression parameters α and β corresponds to the maximum likelihood solution of this model, on the assumption that the residuals are independently normally distributed. We may see this by considering the log likelihood which is given as

$$\log p(\epsilon_{1:T}) = \log p(\epsilon_1) - \frac{1}{2\sigma^2} \sum_{t=2}^{T} (\epsilon_t - \phi \epsilon_{t-1})^2 - \frac{T-1}{2} \log 2\pi\sigma^2$$

and when $\phi = 0$, maximising this degenerates to minimising the sum of squared residual terms

$$\sum_{t=2}^{T} \epsilon_t^2 = \sum_{t=2}^{T} (y_t - \alpha - \beta x_t)^2$$

which is exactly the expression to be minimised upon estimating the values of the coefficients α and β by ordinary least-squares, set out in section 2.2.3. The condition that $\phi = 0$ corresponds to independent normally-distributed error terms since the model assumes $\epsilon_t = \phi \epsilon_{t-1} + \eta_t$. In the case that the true value of the latent variable ϕ generating the data is non-zero, however, a different solution may be optimal in the maximum likelihood sense.

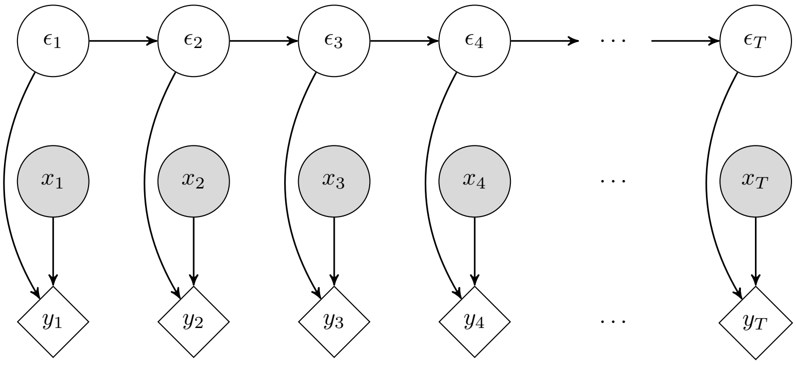


Figure 5.1: Belief network for the natural model for cointegration. Diamonds represent delta functions, shaded variables are in the conditioning set.

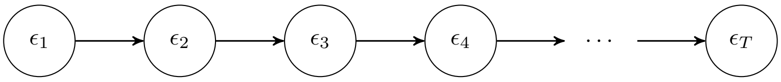


Figure 5.2: Belief network for the Markov chain model for ϵ_t.

5.3 Bayesian Cointegration

The genesis of concern with the classical approach to cointegration relates to the potentially inconsistent treatment of the unobserved parameter ϕ. In this model, I seek to be more consistent by avoiding the assumption that $\phi = 0$ for estimation, while testing the hypothesis that $\phi = 1$. In order to do this, I first construct a generative probabilistic model and think about marginalising ϕ.

We can construct a Bayesian model for estimating both ϕ and the parameters α, β, σ^2 by considering ϕ to be a latent variable and providing a prior distribution. The belief network shown in figure 5.3 applies; the parameters α, β, σ^2 shown at the bottom of the figure are used to determine the value of each y_t and the latent variable ϕ appears at the top. As I have already shown, OLS corresponds to maximum likelihood estimation in this model with the condition that the latent variable $\phi = 0$.

In this case, as set out in section 2.3.1, for cointegration we require $|\phi| < 1$, and we can encode this with a uniform distribution $p(\phi) = \mathcal{U}(\phi|\,(-1,1)) = \frac{1}{2}\,[\phi \in (-1,1)]$. A specification for the distribution of ϵ_1 is then required to complete the model, and the joint density for the model is then given by

$$p(y_{1:T}, \epsilon_{1:T}, \phi|x_{1:T}) = p(y_{1:T}|\epsilon_{1:T}, x_{1:T})\, p(\epsilon_{1:T}|\phi)\, p(\phi)$$

from which the marginal model

$$p(y_{1:T}, \phi|x_{1:T}) = p(\phi) \int_{\epsilon_{1:T}} p(y_{1:T}|\epsilon_{1:T}, x_{1:T})\, p(\epsilon_{1:T}|\phi)$$

is formed by integration. The posterior $p(\phi|x_{1:T}, y_{1:T})$ is then equivalent to $p(\phi|\epsilon_{1:T}) \propto p(\phi)\, p(\epsilon_{1:T}|\phi)$.

5.3.1 Inference of ϕ

I present two inference schemes for the latent variable ϕ. First, I present an iterative scheme based on sequentially updating the belief, and second, a direct method. The sequential inference method is

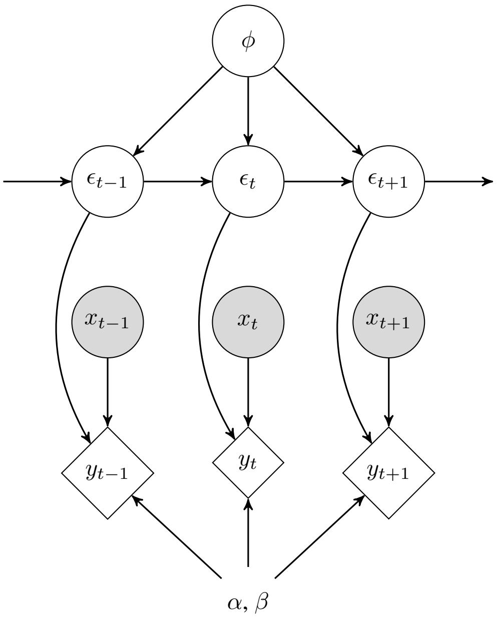


Figure 5.3: Belief network for the combined Bayesian cointegration model. Shaded variables are in the conditioning set.

particularly useful for the intermittent cointegration model set out in section 5.5, and uses an improper uniform prior for ϵ_1. In contrast, the direct method relies on a more intuitive prior for ϵ_1 and is the main scheme for cointegration model inference.

Sequential method

Given a sequence of 'observations' $\epsilon_{1:T}$ our interest is to infer the posterior distribution $p(\phi|\epsilon_{1:T})$. In this section, I simply place an improper uniform prior $p(\epsilon_1) = \mathcal{U}(\epsilon_1|\mathbb{R})$, representing the belief that ϵ_1 may take any real value. Inference in this model is similar to Kalman inference in the well-known 'state-space representation of the Vector Error Correction model' (Seong et al., 2007). However, the uniform prior for ϕ on the interval $(-1, 1)$ complicates the likelihood calculation when compared with the simple linear-Gaussian case. To address this, I cast inference of ϕ as a problem in sequentially updating the belief about the distribution of ϕ for each observation ϵ_t. Initially, we begin with the prior $p(\phi) = \mathcal{U}(\phi|\,(-1, 1))$, and then update the distribution by finding the posterior after observing each ϵ_t. For ϵ_1, there is no update to make since $p(\phi|\epsilon_1) \propto p(\phi)\, p(\epsilon_1)$. Thereafter, $p(\phi|\epsilon_{1:t}) \propto p(\epsilon_t|\epsilon_{t-1}, \phi)\, p(\phi|\epsilon_{1:t-1})$. This can be calculated analytically since, if $p(\phi|\epsilon_{t-1}) \propto p(\phi)\, \mathcal{N}(\phi|f_{t-1}, F_{t-1})$ then the product of Gaussians can be conditioned according to Corollary B.6,

$$\mathcal{N}(\phi|f_{t-1}, F_{t-1})\, \mathcal{N}\big(\epsilon_t\big|\phi\epsilon_{t-1}, \sigma^2\big) = \mathcal{N}\big(\epsilon_t\big|\epsilon_{t-1}f_{t-1}, \sigma^2 + \epsilon_{t-1}^2 F_{t-1}\big)\, \mathcal{N}(\phi|f_t, F_t)$$

where

$$f_t = \frac{f_{t-1}\sigma^2 + \epsilon_t\epsilon_{t-1}F_{t-1}}{\sigma^2 + \epsilon_{t-1}^2 F_{t-1}}, \quad F_t = \frac{\sigma^2 F_{t-1}}{\sigma^2 + \epsilon_{t-1}^2 F_{t-1}} \tag{5.1}$$

Algorithm 5.1 Bayesian cointegration inference: Kalman method

```
$f_2 \leftarrow \frac{\epsilon_2}{\epsilon_1},\quad F_2 \leftarrow \frac{\sigma^2}{\epsilon_1^2},\quad l \leftarrow \frac{1}{|\epsilon_1|}$    ▷ Initialise
for $t = 3$ to $T$ do
    $f_t \leftarrow \frac{f_{t-1}\sigma^2 + \epsilon_t\epsilon_{t-1}F_{t-1}}{\sigma^2 + \epsilon_{t-1}^2 F_{t-1}},\ F_t \leftarrow \frac{\sigma^2 F_{t-1}}{\sigma^2 + \epsilon_{t-1}^2 F_{t-1}}$    ▷ Kalman updates
    $l \leftarrow l \times \mathcal{N}\big(\epsilon_t \big| \epsilon_{t-1} f_{t-1}, \sigma^2 + \epsilon_{t-1}^2 F_{t-1}\big)$    ▷ Likelihood update
end for
$l \leftarrow l \times \frac{1}{2}\int_{-1}^{1} \mathcal{N}(\phi | f_T, F_T)$
return $\{l, \langle\phi\rangle, \langle\phi^2\rangle\}$    ▷ Return posterior moments
```

Algorithm 5.2 Bayesian cointegration inference: Direct method

```
$\hat{e}_{12} \leftarrow \sum_{t=2}^{T} \epsilon_t\epsilon_{t-1},\quad \hat{e}_1 \leftarrow \sum_{t=2}^{T-1} \epsilon_t^2$    ▷ Summarise data
$f_T \leftarrow \frac{\hat{e}_{12}}{\hat{e}_1},\quad F_T \leftarrow \frac{\sigma^2}{\hat{e}_1}$
$l \leftarrow \frac{\left(2\pi\sigma^2\right)^{\frac{1-T}{2}}}{\sqrt{\hat{e}_1}} \left[\underset{\phi}{\int_{-1}^{1}} \frac{1}{2}\sqrt{1-\phi^2}\,\mathcal{N}(\phi | f_T, F_T)\right] \exp -\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T} \epsilon_t^2 - \frac{\left(\hat{e}_{12}\right)^2}{\hat{e}_1}\right)$
return $\{l, \langle\phi\rangle, \langle\phi^2\rangle\}$    ▷ Return posterior moments
```

then the posterior update is given by

$$p(\phi | \epsilon_{1:t}) \propto p(\phi)\, \mathcal{N}(\phi | f_t, F_t)$$

which we identify as updates similar to the Kalman updates of the linear dynamical system set out in section 2.2.1. On setting the initial[1] f_2, F_2 from the emission

$$\mathcal{N}\big(\epsilon_2 \big| \phi\epsilon_1, \sigma^2\big) = \frac{1}{|\epsilon_1|}\mathcal{N}\left(\phi \left| \frac{\epsilon_2}{\epsilon_1}, \frac{\sigma^2}{\epsilon_1^2}\right.\right) \tag{5.2}$$

this defines a recursion for the parameters f_t, F_t.

After completing the updates, the required posterior $p(\phi | \epsilon_{1:T})$ is given by a Gaussian distribution with mean f_T and variance F_T truncated to the interval $(-1, 1)$, see algorithm 5.1.

Likelihood. The likelihood in this model is given by

$$p(\epsilon_{1:T}) = p(\epsilon_1) \prod_{t=2}^{T} \int_{\phi} p(\epsilon_t | \epsilon_{t-1}, \phi)\, p(\phi | \epsilon_{1:t-1})$$

where the integral in respect of $t = 2$ is given by

$$p(\epsilon_2 | \epsilon_1) = \frac{1}{2\,|\epsilon_1|} \underset{\phi}{\int_{-1}^{1}} \mathcal{N}\left(\phi \left| \frac{\epsilon_2}{\epsilon_1}, \frac{\sigma^2}{\epsilon_1^2}\right.\right)$$

and thereafter each $p(\epsilon_t | \epsilon_{1:t-1})$ is given by

$$\mathcal{N}\big(\epsilon_t \big| \epsilon_{t-1} f_{t-1}, \sigma^2 + \epsilon_{t-1}^2 F_{t-1}\big)\, \frac{\int_{-1}^{1} \mathcal{N}(\phi | f_t, F_t)}{\int_{-1}^{1} \mathcal{N}(\phi | f_{t-1}, F_{t-1})}.$$

[1] In the unlikely event that $\epsilon_1 = 0$, the Gaussian term arises from the first non-zero ϵ_{t-1}.

Therefore the likelihood is given as

$$p(\epsilon_{1:T}) = \frac{1}{2\,|\epsilon_1|}\left[\prod_{t=2}^{T-1}\mathcal{N}\big(\epsilon_{t+1}\big|\epsilon_t f_t, \sigma^2 + \epsilon_t^2 F_t\big)\right]\int_{-1}^{1}\mathcal{N}(\phi|f_T, F_T)\,.$$
$$\phi$$

Direct inference

The recursive inference scheme given above is useful to develop intuition of the Bayesian cointegration model, and is particularly useful in the development of a switching model for intermittent cointegration, as I describe later in the chapter.

For the purpose of simple cointegration, however, this calculation scheme can be quite onerous, and the assumption of a flat prior on ϵ_1 is not ideal to ensure that the residuals process is stationary. As set out in section 2.2.3, the process $\epsilon_{1:T}$ is stationary when each $\langle\epsilon_t\rangle = 0$ and $\left\langle\epsilon_t^2\right\rangle$ is constant, independent of t, and $p(\epsilon_t, \epsilon_s)$ depends on $|t-s|$ not t or s. According to the recurrence relation for ϵ_t, the variance satisfies $\left\langle\epsilon_{t+1}^2\right\rangle = \phi^2\left\langle\epsilon_t^2\right\rangle + \sigma^2$ and we can therefore ensure stationary of $\epsilon_{1:T}$ if $\left\langle\epsilon_1^2\right\rangle = \sigma^2/\left(1-\phi^2\right)$. So I consider $p(\epsilon_1|\phi) = \mathcal{N}\big(\epsilon_1\big|0, \sigma^2/\left(1-\phi^2\right)\big)$ to be a natural prior for ϵ_1. Indeed, with this prior we can show the sequence is stationary by considering the final condition on the joint distribution of any $p(\epsilon_t, \epsilon_s)$. First of all, I note that with this prior on ϵ_1, we have each marginal

$$p(\epsilon_t) = \mathcal{N}\left(\epsilon_t\middle|0, \frac{\sigma^2}{(1-\phi^2)}\right).$$

Then assuming $t > s$ (without loss of generality) by appealing to the result of Gaussian linear transforms Corollary B.3 the joint

$$\begin{aligned}p(\epsilon_t, \epsilon_s) &= \int_{\epsilon_{s+1:t-1}}\mathcal{N}\big(\epsilon_t\big|\phi\epsilon_{t-1}, \sigma^2\big)\,\mathcal{N}\big(\epsilon_{t-1}\big|\phi\epsilon_{t-2}, \sigma^2\big)\\ &\qquad\times\ldots\times\mathcal{N}\big(\epsilon_{s+1}\big|\phi\epsilon_s, \sigma^2\big)\,\mathcal{N}\left(\epsilon_s\middle|0, \frac{\sigma^2}{(1-\phi^2)}\right)\\ &= \mathcal{N}\left(\epsilon_t\middle|\phi^{(t-s)}\epsilon_s, \sigma^2\left(\sum_{i=1}^{t-s}\phi^{2(i-1)}\right)\right)\mathcal{N}\left(\epsilon_s\middle|0, \frac{\sigma^2}{(1-\phi^2)}\right)\end{aligned}$$

from which we can see that the pairwise joint distribution relies on the value only of the difference $|t-s|$ not t or s. Using the prior $p(\epsilon_1|\phi) = \mathcal{N}\big(\epsilon_1\big|0, \sigma^2/\left(1-\phi^2\right)\big)$ is therefore consistent with all of the conditions for a stationary sequence $\epsilon_{1:T}$; fortunately it is possible to derive a direct formulation of inference based on this prior.

First, writing

$$\begin{aligned}p(\phi, \epsilon_{1:T}) &= p(\phi)\,p(\epsilon_1)\prod_{t=2}^{T}\mathcal{N}\big(\epsilon_t\big|\phi\epsilon_{t-1}, \sigma^2\big)\\ &= p(\phi)\,\frac{\sqrt{1-\phi^2}}{\left(2\pi\sigma^2\right)^{\frac{T}{2}}}\exp -\frac{1}{2\sigma^2}\left[\epsilon_1^2\left(1-\phi^2\right) + \sum_{t=2}^{T}\left(\epsilon_t - \phi\epsilon_{t-1}\right)^2\right]\end{aligned}$$

we can write the posterior for ϕ and complete the square,

$$\begin{aligned}p(\phi|\epsilon_{1:T}) &\propto p(\phi)\,\sqrt{1-\phi^2}\exp -\frac{1}{2\sigma^2}\left[\phi^2\left(\sum_{t=2}^{T-1}\epsilon_t^2\right) - 2\phi\left(\sum_{t=2}^{T}\epsilon_t\epsilon_{t-1}\right) + \left(\sum_{t=1}^{T}\epsilon_t^2\right)\right]\\ &\propto p(\phi)\,\sqrt{1-\phi^2}\,\mathcal{N}\left(\phi\middle|\frac{\hat{e}_{12}}{\hat{e}_1}, \frac{\sigma^2}{\hat{e}_1}\right)\end{aligned}$$

where

$$\hat{e}_{12} \equiv \sum_{t=2}^{T} \epsilon_t \epsilon_{t-1}, \qquad \hat{e}_1 \equiv \sum_{t=2}^{T-1} \epsilon_t^2.$$

For the data likelihood, we write for $p(\epsilon_{1:T})$,

$$p(\epsilon_{1:T}) = \frac{\left(2\pi\sigma^2\right)^{\frac{1-T}{2}}}{\sqrt{\hat{e}_1}} \left[\int_{-1}^{1}_{\phi} \tfrac{1}{2}\sqrt{1-\phi^2}\,\mathcal{N}(\phi|f_T, F_T)\right] \exp -\frac{1}{2\sigma^2}\left(\sum_{t=1}^{T}\epsilon_t^2 - \frac{\left(\hat{e}_{12}\right)^2}{\hat{e}_1}\right)$$

5.3.2 Estimating α, β, σ^2

In order to estimate the cointegration relation, I note that the likelihood for the cointegration model above is given as a function of the data $x_{1:T}$, $y_{1:T}$ and the parameters of the linear cointegration relationship α, β. We can therefore take an estimate of the cointegration relationship by setting the parameters $\theta = \left\{\alpha, \beta, \sigma^2\right\}$ based on maximising the likelihood

$$p(y_{1:T}|x_{1:T}, \theta) = \int_{\phi} p(\phi)\, p(\epsilon_{1:T}|\phi)\,.$$

Since ϕ is a latent variable, it is convenient to approach this using the Expectation Maximisation (EM) algorithm set out in section 2.4.4. For the second stage of maximising the lower bound, since q is fixed it suffices to maximise the energy term; in this model, we replace $h \to \phi$, $v \to \epsilon_{1:T}$.

The energy term is the log of the model joint, and since we are interested in maximising with respect to the regression parameters θ the relevant terms are

$$\left\langle \log p(\epsilon_1|\phi)\right\rangle_{q(\phi|\epsilon_{1:T})} + \sum_{t=2}^{T}\left\langle \log p(\epsilon_t|\epsilon_{t-1},\phi)\right\rangle_{q(\phi|\epsilon_{1:T})}$$

and in the case of the direct inference model given above, this (up to a constant) equals

$$-\frac{1}{2\sigma^2}\left[\epsilon_1^2\left\langle 1-\phi^2\right\rangle_{q(\phi|\epsilon_{1:T})} + \sum_{t=2}^{T}\left\langle\left(\epsilon_t - \phi\epsilon_{t-1}\right)^2\right\rangle_{q(\phi|\epsilon_{1:T})}\right] - \frac{T}{2}\log 2\pi\sigma^2$$

where $q(\phi|\epsilon_{1:T}) = p\big(\phi|\epsilon_{1:T}, \theta^{\text{old}}\big)$ is the posterior from the previous estimates. The energy can be optimised by finding the stationary point by differentiating by α and β. Since

$$\left\langle\left(\epsilon_t - \phi\epsilon_{t-1}\right)^2\right\rangle = \epsilon_t^2 - 2\epsilon_t\epsilon_{t-1}\left\langle\phi\right\rangle + \epsilon_{t-1}^2\left\langle\phi^2\right\rangle$$

the result is a system of linear equations for α and β involving the first and second (non-central) moments of ϕ from the posterior $q(\phi|\epsilon_{1:T})$ that yield a unique solution. In the case of the sequential approach to inference set out above, these moments can be found analytically by appealing to the derivations shown in appendix C. For the direct inference method, the integrals cannot be evaluated analytically and it is necessary to appeal to numerical integration techniques[2].

By differentiating the above energy term with respect to σ^2, we find that optimally

$$\widehat{\sigma}^2 = \frac{1}{T}\left[\epsilon_1^2\left\langle 1-\phi^2\right\rangle_{q(\phi|\epsilon_{1:T})} + \sum_{t=2}^{T}\left\langle\left(\epsilon_t - \phi\epsilon_{t-1}\right)^2\right\rangle_{q(\phi|\epsilon_{1:T})}\right] \tag{5.3}$$

so the variance can be estimated once the new regression estimates α and β have been found.

[2] My implementation uses Gauss-Kronrod quadrature since the posterior has singularities at the end-points of the definite integral, $\phi = \pm 1$.

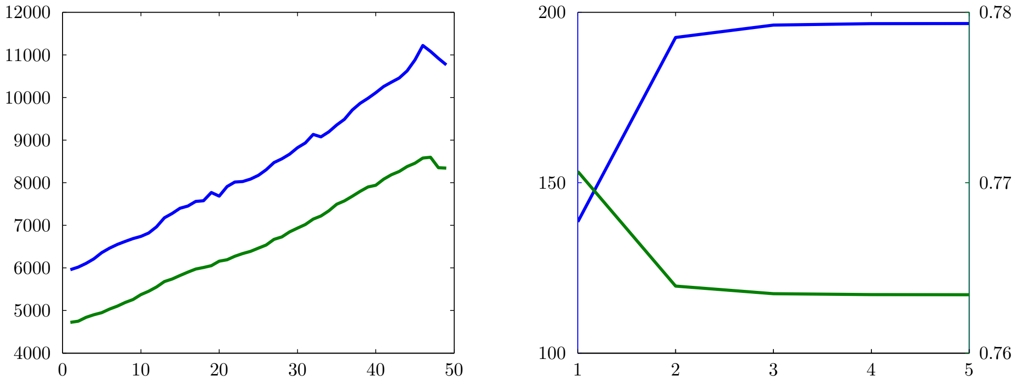


Figure 5.4: Estimation in the Bayesian cointegration model for US income and consumption. I show on the left the raw data for personal income (blue line, \$millions) and personal consumption expenditure (green line, \$millions), quarterly data for the period 1997-2009. On the right, I show the evolution of α (blue line, left-hand axis) and β (green line, right-hand axis) as the parameters are updated by EM—the initial estimates are given by OLS. The inference routine gives the posterior $p(\phi|\epsilon_{1:T})$ shown in figure 5.5, indicating correlated residuals. The Dickey-Fuller test shows p-value 0.0026 for the OLS estimate and 0.0033 for the EM estimate, highly likely to reject the null hypothesis of a random walk and showing strong evidence of cointegration in both cases.

5.3.3 Example

To illustrate the effectiveness of parameter estimation in the Bayesian model based on the direct scheme of inference of ϕ, I give an application to a classical problem: the link between personal income and personal consumption. Such data were analysed by Davidson et al. (1978) for the UK; here, I take a look at similar data for the US[3]. The results are shown in figure 5.4 in which I draw attention to the difference between the OLS estimate for the parameters and the estimate resulting from EM in the cointegration model—most striking for the intercept parameter α. I also give the posterior for ϕ in figure 5.5 which shows a mode at 0.62, accounting for the difference in estimations of α and β compared with OLS.

5.3.4 Estimator Convergence

It is known that OLS is a consistent estimator for the parameters α, β, and further, that OLS is the maximum likelihood estimator (MLE) in this model when $\phi = 0$. As noted in section 5.2, the MLE found by EM is in general different when $\phi \neq 0$, and it is known that OLS can show considerable bias for small sample sizes. I present a comparison of the convergence properties of OLS and MLE in this model for increasing sample size T compared with the true values based on generated data; the results are shown in figure 5.6. In this experiment, I generated $10,000$ series each of length $2,000$ and estimated the parameters with the first T samples from each series for increasing T—in order to show the convergence properties of the estimation schemes. Figure 5.7 shows a further comparison of the accuracy of the two methods. Interestingly, the EM estimates for α and β are in general more accurate and converge more

[3] US income and consumption figures were obtained from the US Department of Commerce: Bureau of Economic Analysis.

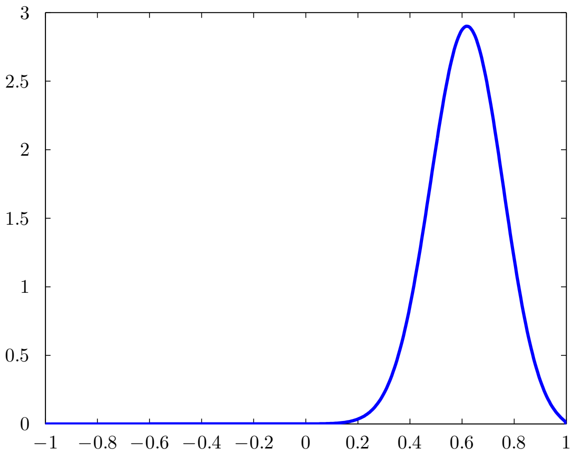


Figure 5.5: Posterior for ϕ from income-consumption estimation of figure 5.4. The distribution shows significant evidence that the true value of ϕ is greater than zero, corresponding to correlated residual error terms.

quickly than OLS.

5.4 Bayesian Cointegration Testing

So far in this chapter I have shown a novel method of estimating a cointegration relationship between the series x_t and y_t. In the analysis, there has been no assumption made about the underlying generating process by focussing only on the conditional relationship $p(y_{1:T}|x_{1:T})$. As set out in section 2.3, in order to conclude that the series are in fact cointegrated, we require an estimate of the relationship and evidence to support the conclusion that the residuals process ϵ_t is stationary, rather than a random walk[4]. In the classical approach, once the relationship has been estimated with ordinary least squares the second step is to test for a unit root in the residuals process, which would indicate a random walk.

The Bayesian approach of this chapter has so far only delivered an estimate of the relationship; an important second step is to consider the confidence that the residuals are not a random walk. In this section, I propose a simple analogy to the Dickey-Fuller testing step of the classical approach.

At the beginning of this section, I therefore assume there is already a known or estimated value for the relationship coefficients α, β and focus on checking for stationarity in the residuals process ϵ_t. I return to estimation of the coefficients later in our discussion and show how the ‘testing’ scheme set out in this section can be combined with the estimation scheme of section 5.3.

Section 2.1.2 set out the notion of Bayesian model selection, and in particular, the idea that comparing the likelihood of models allows one to determine the most appropriate model, by appealing to the Bayes factor or otherwise. For our purposes, if we take as $\mathcal{M}_2$ the cointegration model set out above with $|\phi| < 1$, we

[4] The series are cointegrated if there *exists* a linear combination which renders the residuals of the process stationary. We are hence required to pick a candidate relationship and check whether the corresponding residuals are most likely to be stationary.

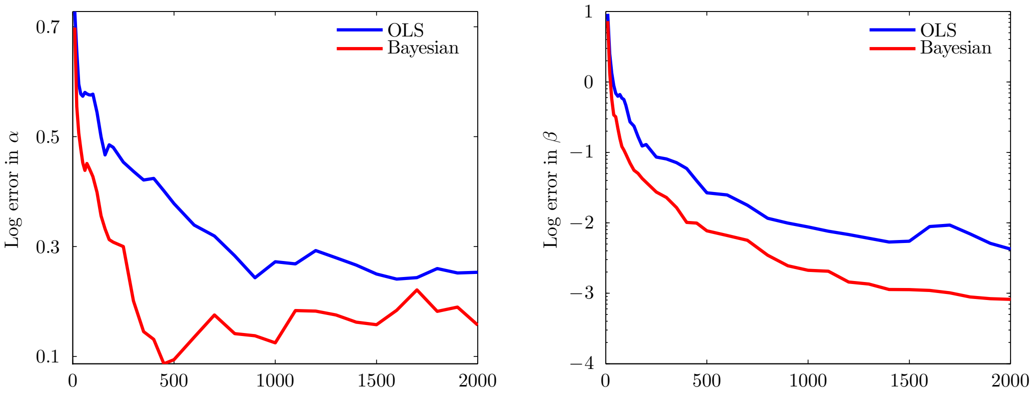


Figure 5.6: Comparison of squared error in OLS and Bayesian estimators for α, β. I generated $10,000$ cointegrated series of length $2,000$ and for each, estimated the relationship for increasing sample size T with both OLS and EM. Here I show the average error in the estimated values for the coefficients compared with the true value; the results show that in general EM converges more quickly than OLS.

wish to compare the fit with a random walk model $\mathcal{M}_1$. In this section, I show how to perform inference in such a model.

5.4.1 Random Walk Likelihood

We consider ϵ_t a random walk when each $\epsilon_t \sim \mathcal{N}\big(\epsilon_{t-1}, \sigma^2\big)$. Based on the modelling framework already discussed, this corresponds to the case $\phi = 1$. The likelihood for ϵ_t a random walk is calculated as

$$p(\epsilon_{1:T}|\phi = 1) = p(\epsilon_1)\prod_{t=2}^{T}\mathcal{N}\big(\epsilon_t\big|\epsilon_{t-1}, \sigma^2\big)\,.$$

For the prior $p(\epsilon_1)$ I choose a wide-interval uniform distribution.

The cointegration model and random walk model can then be compared according to the ratio of the data likelihoods,

$$\frac{p(y_{1:T}|x_{1:T}; \phi = 1)}{p(y_{1:T}|x_{1:T}; |\phi| < 1)} = \frac{p(\epsilon_{1:T}|\phi = 1)}{p(\epsilon_{1:T}|\,|\phi| < 1)} \equiv \frac{l_{\mathrm{RW}}}{l_{\mathrm{C}}}$$

where the numerator represents the ‘null hypothesis’ of a unit root in the residuals process. The denominator is simply the marginal likelihood for the Bayesian cointegration model given in section 5.3.

5.4.2 Estimation and Testing

Both the numerator (random walk model) and denominator (cointegration model) of the Bayes factor are functions in the variance σ^2, and I set the value in each of the two models by maximum likelihood[5]. For the random walk model, the estimate of the variance is given as

$$\sigma^2_{\mathrm{RW}} = \frac{1}{T-1}\sum_t \left(\epsilon_t - \epsilon_{t-1}\right)^2 .$$

[5] An alternative approach would take a prior for the variance and marginalise the parameter. I mention a tractable prior for σ^2 in section 7.3.

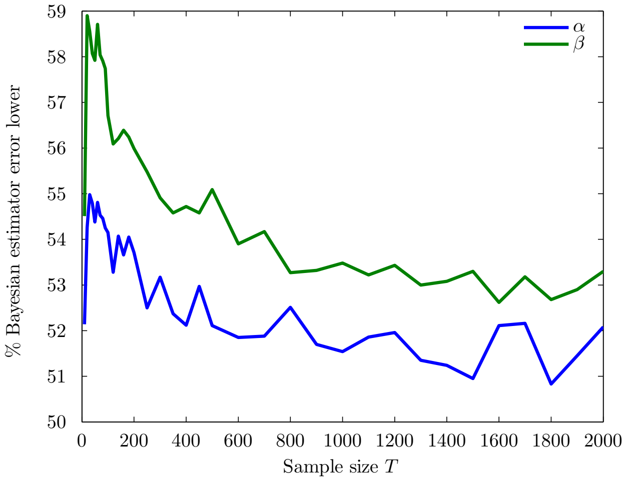


Figure 5.7: Comparison of the relative error for the estimator convergence data of figure 5.6. Here I show the proportion of cases with the error in the estimator obtained by EM lower than the error in the OLS estimate.

Note that each $\epsilon_t = y_t - \alpha - \beta x_t$ is a function in the data x_t, y_t and the cointegration parameters α, β. Prior to comparing the likelihoods, it is necessary to consider how best to estimate the parameters α and β for the comparison. It is possible to perform maximum likelihood estimation for the parameters α, β in the random walk model set out in this section. The likelihood function of the random walk model is not however dependent on α, and the maximum likelihood value for β is given as

$$\beta = \frac{\sum_t (x_t - x_{t-1})(y_t - y_{t-1})}{\sum_t (x_t - x_{t-1})^2}.$$

There are three primary options for choice of α, β: (i) find the parameters by OLS and use these values for both cointegration and random walk; (ii) find the parameters by maximum likelihood (expectation maximisation) in the cointegration model and use the same parameters for the random walk model; and (iii) find the parameters for each of the cointegration and random walk models by maximum likelihood. Conceptually, our interest is to first *estimate* a cointegration relationship, and second to *test* whether such relationship is more likely to be a random walk—and for this reason, I consider the third option conceptually undesirable. Option (i) is computationally the simplest but suffers from the implicit assumption of OLS estimation that $\phi = 0$; option (ii) is the preferred option, using the estimation scheme set out in section 5.3. I come to a comparison of the effectiveness of the different methods later in figure 5.9.

When the parameters α, β are estimated with the cointegration model of section 5.3 as with my preferred method and not with the random walk model of this section it can be tricky to describe the proposed method in the framework of Bayesian model selection, which would naturally place priors over the parameters α, β for each model and marginalise them for the purposes of the comparison. For the purposes of this thesis, I have sought to build a method directly comparable with the classic Engle-Granger cointegration estimation and testing approach. Bayesian analysis of the cointegrating parameters

Algorithm 5.3 Bayesian cointegration testing

```
$\left\{\alpha, \beta, \sigma^2\right\} \leftarrow \textsc{LinearRegression}\left(x_{1:T}, y_{1:T}\right)$    ▷ Initialise to OLS estimate
**repeat**
    $\epsilon_{1:T} \leftarrow y_{1:T} - \alpha - \beta x_{1:T}$    ▷ Find residual
    $\left\{l_{\mathrm{C}}, \langle\phi\rangle, \langle\phi^2\rangle\right\} \leftarrow \textsc{CointInference}\left(\epsilon_{1:T}, \sigma^2\right)$    ▷ Inference of $\phi$
    $\left\{\alpha, \beta, \sigma^2\right\} \leftarrow \mathrm{EM}\left(x_{1:T}, y_{1:T}, \langle\phi\rangle, \langle\phi^2\rangle\right)$    ▷ Update parameter estimates
**until** convergence
$\sigma^2_{\mathrm{RW}} \leftarrow \sum_t \left(\epsilon_t - \epsilon_{t-1}\right)^2 / \left(T-1\right)$    ▷ Random walk variance
$l_{\mathrm{RW}} \leftarrow \prod_{t=2}^{T} \mathcal{N}\left(\epsilon_t \middle| \epsilon_{t-1}, \sigma^2_{\mathrm{RW}}\right)$    ▷ Random walk likelihood
**return** cointegrated $\leftarrow l_{\mathrm{RW}}/l_{\mathrm{C}} <$ threshold    ▷ Compare likelihoods
```

is a natural extension to this model.

Note that combining estimation in the Bayesian cointegration model and the likelihood comparison of this section does not make the conflicting assumptions of the classical approach.

As mentioned in section 2.1.2, the magnitude of the Bayes factor informs the strength of support for one or other model. To draw a conclusion a decision threshold is required, analogous to the significance level of a classical statistical test such as the Dickey-Fuller test. Differing decision thresholds for the Bayes factor correspond to different widths of uniform prior for ϵ_1 in the RW model since

$$\frac{l_{\mathrm{RW}}}{l_{\mathrm{C}}} < C \Leftrightarrow \frac{\tilde{l}_{\mathrm{RW}}}{l_{\mathrm{C}}} < 1, \quad \tilde{l}_{\mathrm{RW}} \equiv \frac{l_{\mathrm{RW}}}{C}.$$

5.4.3 Robust Regression Estimation

I give a comparison of the success rate of the algorithm with the classical cointegration test in figure 5.8. The results in figure 5.8 show that, compared with OLS estimation and unit root testing, the Bayesian technique is less likely to result in a spurious relationship for series of length $T > 20$. Figure 5.9 shows a plot of the sensitivity of the Bayesian cointegration estimation and testing approach compared with the classical method, for differing values of the decision thresholds. The results show that for these generated series, the combined Bayesian approach is more robust than OLS and Dickey-Fuller unit root testing.

5.5 Intermittent Cointegration

As set out in section 2.3.4, previous approaches to intermittent cointegration generally assume there are only two or three segments. In order to permit more regimes, ideally an unknown number, a more complex model and inference scheme are required. In this section, I show that inference in a model with an unknown number of cointegration regimes is achievable efficiently by appealing to the reset model framework of chapter 3 and the sequential inference scheme described in section 5.3.1. The following model seeks to detect a cointegration relationship between two series, while allowing for regions when the relationship is in fact a random walk.

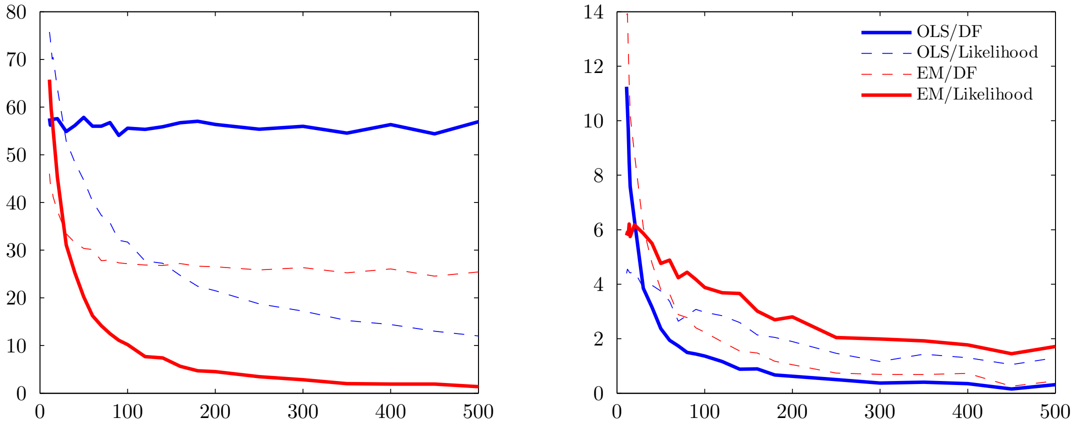


Figure 5.8: Plot of rate of type I errors (false positives, left-hand plot) and type II errors (false negatives, right-hand plot) for the different cointegration estimation/testing combinations against time-series length T (x-axis). OLS refers to estimation using least squares; EM refers to estimation in the Bayesian cointegration model with expectation maximisation. DF refers to a simple Dickey-Fuller test at 5% significance; Likelihood refers to comparing the likelihood of the Bayesian cointegration model and random walk model. $10,000$ series of each length were simulated according to the generative model, uniformly split between $|\phi| < 1$ in the cointegrated case and $\phi = 1$ in the random walk case. The high rate of false positives for the classical test is striking: here cointegration was detected when the data were generated with $\phi = 1$. I attribute this to the fact that OLS is known to produce invalid estimates when the processes are non-stationary: so called 'spurious regression'.

I structure the segmented cointegration problem as a regime-switching Bayesian model and appeal to the piecewise-constant reset model of section 3.5. Whereas the simple cointegration model assumes $|\phi| < 1$ and tests the belief that the residuals are stationary, for this switching model I allow time-varying ϕ_t arranged to be piecewise-constant with no *a priori* restriction on the number of piecewise-constant regions. The model will switch between the cointegrated case $|\phi_t| < 1$ and the alternative $\phi_t = 1$, corresponding to a random walk. The model therefore permits resets in the latent variable ϕ at change-points.

The binary state switch i_t denotes whether there is a cointegration relationship at time t: $i_t = 1$ denotes regions of ϵ_t corresponding to a random walk; alternatively when $i_t = 0$, ϵ_t follows a stationary cointegration relation. The model is given by the joint probability distribution

$$p(\epsilon_{1:T}, i_{2:T}, \phi_{2:T}) = p(\epsilon_1) \prod_{t=2}^{T} p(\epsilon_t | \epsilon_{t-1}, \phi_t)\, p(i_t | i_{t-1})\, p(\phi_t | \phi_{t-1}, i_t, i_{t-1})$$

where $i_1, \phi_1 = \emptyset$. A belief network showing this distribution is given in figure 5.10.

The emission term is given as with the simple cointegration case, $p(\epsilon_t | \epsilon_{t-1}, \phi_t) = \mathcal{N}(\epsilon_t | \phi_t \epsilon_{t-1}, \sigma^2)$. The transition distribution in ϕ_t depends on the prevailing state i_t; if there has been a change of state, the value for ϕ_t is redrawn from a prior distribution. In the random walk regions, we have $\phi_t = 1$, which is

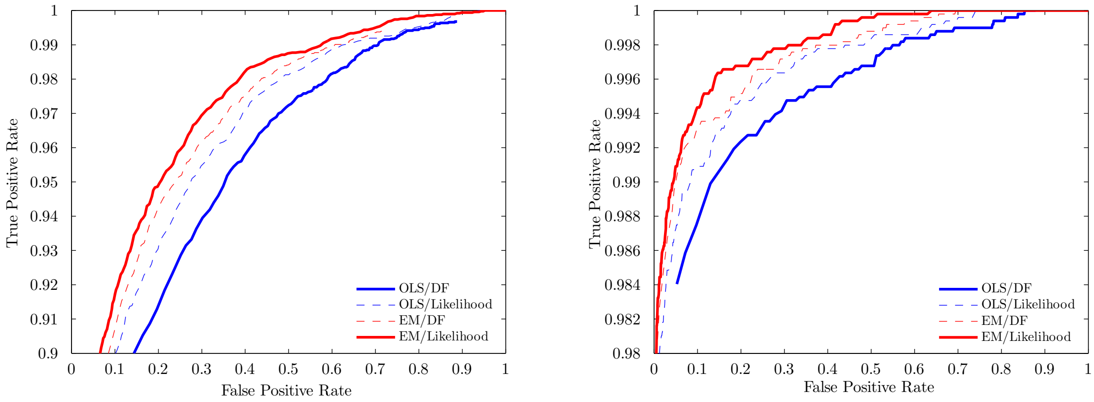


Figure 5.9: Plot of the receiver operating characteristic (ROC) curve for the cointegration estimation/testing combinations. Left, I show the ROC for $T = 50$, $10,000$ generated cointegrated series; on the right, the same with $T = 500$. EM refers to estimation in the Bayesian cointegration model by maximum likelihood; DF refers to a Dickey-Fuller test; Likelihood refers to comparing the likelihood of the cointegration model with the random walk model $\phi = 1$. Optimal classification occurs with true positive rate 100%, false positive rate 0%; the results show that OLS/DF is dominated by EM/Likelihood.

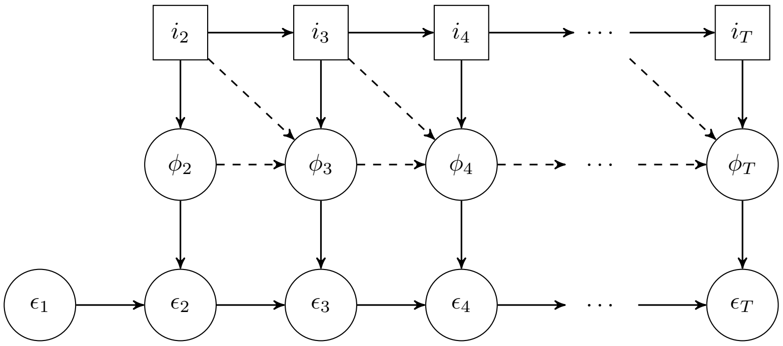


Figure 5.10: Belief network for the partial cointegration model. Dashed edges to ϕ_t are absent when $i_t = 1$.

encoded in the reset distribution $p^1(\phi_t) = \delta(\phi_t - 1)$. Therefore the initial distribution for ϕ_2 is

$$p(\phi_2|i_2) = \begin{cases} p^1(\phi_2) = \delta(\phi_2 - 1) & i_2 = 1 \\ p^0(\phi_2) = \mathcal{U}(\phi_2|\,(-1,1)) & i_2 = 0 \end{cases}$$

and for the piecewise-constant transition,

$$p(\phi_t|\phi_{t-1}, i_t, i_{t-1}) = \begin{cases} p^1(\phi_t) & i_t = 1 \\ \delta(\phi_t - \phi_{t-1}) & i_t = 0, i_{t-1} = 0 \\ p^0(\phi_t) & i_t = 0, i_{t-1} = 1. \end{cases}$$

The transition distribution for the state i_t is specified as a constant *a priori*. This means a geometric prior is effectively placed over the length of each regime, as with the reset of model of chapter 3.

5.5.1 Inference of $i_{2:T}$, $\phi_{2:T}$

I briefly describe the inference routines for the cointegration reset model in this section. As can be seen, the inference is an application of insight into reset model inference developed in chapter 3.

5.5.2 Filtering

The filtered distribution[6] $\alpha(\phi_t) \equiv p(\phi_t|\epsilon_{1:t})$ will be written as

$$\begin{aligned}\alpha(\phi_t) &= p(\phi_t, i_t = 1|\epsilon_{1:t}) + p(\phi_t, i_t = 0|\epsilon_{1:t}) \\ &= \alpha(\phi_t|i_t = 1)\,\alpha(i_t = 1) + \alpha(\phi_t|i_t = 0)\,\alpha(i_t = 0)\end{aligned}$$

Non-cointegrated case

The easiest case corresponds to $i_t = 1$ so I tackle this first.

$$\begin{aligned}p(\phi_t, i_t = 1|\epsilon_{1:t}) &= \frac{p(\phi_t, i_t = 1, \epsilon_t|\epsilon_{1:t-1})}{p(\epsilon_t|\epsilon_{1:t-1})} \\ &= \frac{1}{Z_t} p(\epsilon_t|\epsilon_{t-1}, \phi_t)\,p(\phi_t|i_t = 1)\,p(i_t = 1|\epsilon_{1:t-1})\end{aligned}$$

where $Z_t = p(\epsilon_t|\epsilon_{1:t-1})$ will be calculated later. Then

$$\begin{aligned}\alpha(i_t = 1) &= \int_{\phi_t} \frac{1}{Z_t} p(\epsilon_t|\epsilon_{t-1}, \phi_t)\,p^1(\phi_t)\,p(i_t|\epsilon_{1:t-1}) \\ &= \frac{1}{Z_t} \mathcal{N}\big(\epsilon_t|\epsilon_{t-1}, \sigma^2\big) \sum_{i_{t-1}} p(i_t = 1|i_{t-1})\,\alpha(i_{t-1}) \qquad (5.4)\end{aligned}$$

and the continuous component is trivial,

$$\alpha(\phi_t|i_t = 1) \propto p(\epsilon_t|\epsilon_{t-1}, \phi_t)\,p(\phi_t|i_t = 1) \quad \Rightarrow \quad \alpha(\phi_t|i_t = 1) = p^1(\phi_t)$$

Cointegrated case

When $i_t = 0$, the inference is more complicated since the posterior for ϕ_t is non-trivial.

$$\begin{aligned}p(\phi_t, i_t = 0|\epsilon_{1:t}) &= \frac{1}{Z_t} \sum_{i_{t-1}} p(\phi_t, i_t, i_{t-1}, \epsilon_t|\epsilon_{1:t-1}) \\ &= \frac{1}{Z_t} \sum_{i_{t-1}} p(\epsilon_t|\epsilon_{t-1}, \phi_t)\,p(i_t = 0|i_{t-1})\,\alpha(i_{t-1}) \underbrace{\int_{\phi_{t-1}} p(\phi_t|i_t = 0, \phi_{t-1}, i_{t-1})\,\alpha(\phi_{t-1}|i_{t-1})}_{=\begin{cases} p^0(\phi_t) & i_{t-1} = 1 \\ \alpha(\phi_{t-1} = \phi_t|i_{t-1} = 0) & i_{t-1} = 0 \end{cases}}\end{aligned}$$

and the emission term

$$p(\epsilon_t|\epsilon_{t-1}, \phi_t) = \mathcal{N}\big(\epsilon_t|\phi_t\epsilon_{t-1}, \sigma^2\big) = \begin{cases} \mathcal{N}\big(\epsilon_t|0, \sigma^2\big) & \epsilon_{t-1} = 0 \\ \frac{1}{|\epsilon_{t-1}|}\mathcal{N}\left(\phi_t \middle| \frac{\epsilon_t}{\epsilon_{t-1}}, \frac{\sigma^2}{\epsilon_{t-1}^2}\right) & \epsilon_{t-1} \neq 0. \end{cases} \qquad (5.5)$$

[6] In contrast with earlier application of α notation for filtering, the messages here are assumed to be in normalised form.

Combining these last two results, we see that according to chapter 3 an additional component is contributed to the filtering recursion for $i_t = 0$ at each iteration. By further noting that a product of Gaussians is a Gaussian, we see that the posterior is given by a truncated mixture of Gaussians; and therefore we derive a recursion by writing each $\alpha(\phi_t | i_t = 0) = p^0(\phi_t) \sum_{\rho_t} w_{\rho_t} \mathcal{N}(\phi_t | f_{\rho_t}, F_{\rho_t})$. The index ρ_t corresponds to the run-length parameter as set out in chapter 3—giving the number of timesteps since a switch to cointegration.

Ignoring, for brevity, the simpler cases when $\epsilon_{t-1} = 0$, the new regime $\rho_t = 0$ has the component given in equation (5.5); for $\rho_t > 0$ we have a product of Gaussians given by the following (with $\rho_{t-1} = \rho_t - 1$), which we condition

$$\begin{aligned}\mathcal{N}\big(\phi_t \big| f_{\rho_{t-1}}, F_{\rho_{t-1}}\big)\, \mathcal{N}\big(\epsilon_t \big| \phi_t \epsilon_{t-1}, \sigma^2\big) &= \\ &\mathcal{N}\big(\epsilon_t \big| \epsilon_{t-1} f_{\rho_{t-1}}, \sigma^2 + \epsilon_{t-1}^2 F_{\rho_{t-1}}\big)\, \mathcal{N}\left(\phi_t \,\middle|\, \frac{f_{\rho_{t-1}}\sigma^2 + \epsilon_t \epsilon_{t-1} F_{\rho_{t-1}}}{\sigma^2 + \epsilon_{t-1}^2 F_{\rho_{t-1}}}, \frac{\sigma^2 F_{\rho_{t-1}}}{\sigma^2 + \epsilon_{t-1}^2 F_{\rho_{t-1}}}\right)\end{aligned}$$

according to equation (5.1).

After calculating the new Gaussian components for $\alpha(\phi_t | i_t = 0)$, we finish the recursion derivations with

$$\alpha(i_t = 0) = \int_{\phi_t} p(\phi_t, i_t = 0 | \epsilon_{1:t})\,. \tag{5.6}$$

5.5.3 Likelihood

The likelihood of the data $p(\epsilon_{1:T})$ is given by

$$p(\epsilon_{1:T}) = p(\epsilon_1) \prod_{t=2}^{T} p(\epsilon_t | \epsilon_{1:t-1}) = p(\epsilon_1) \prod_{t=2}^{T} Z_t$$

where the normalisation Z_t is calculated during the filtering recursion from equation (5.4) and equation (5.6) since

$$\alpha(i_t = 1) + \alpha(i_t = 0) = 1,$$

hence each likelihood is given by

$$\begin{aligned}Z_t = \mathcal{N}\big(\epsilon_t \big| \epsilon_{t-1}, \sigma^2\big) &\left(\sum_{i_{t-1}} p(i_t = 1 | i_{t-1})\, \alpha(i_{t-1}) \right) \\ + p(i_t = 0 | i_{t-1} = 0)\, \alpha(i_{t-1} = 0) &\int_{\phi_t} \mathcal{N}\big(\epsilon_t \big| \phi_t \epsilon_{t-1}, \sigma^2\big)\, \alpha(\phi_{t-1} = \phi_t | i_{t-1} = 0) \\ &+ p(i_t = 0 | i_{t-1} = 1)\, \alpha(i_{t-1} = 1) \int_{\phi_t} \mathcal{N}\big(\epsilon_t \big| \phi_t \epsilon_{t-1}, \sigma^2\big)\, p^0(\phi_t)\,.\end{aligned}$$

Whilst it is not necessary to calculate this likelihood, it does permit a common-sense stopping criterion for the EM recursion. The likelihood in practice quickly becomes very small, and it is useful to work in log space to avoid numerical underflow, and for the same reason, the weights w_{ρ_t} are also calculated in log space in my implementation.

5.5.4 Smoothing

As with the filtered distribution, I write the smoothed posterior $\gamma(\phi_t) \equiv p(\phi_t|\epsilon_{1:T})$ as

$$\gamma(\phi_t) = p(\phi_t, i_t = 1|\epsilon_{1:T}) + p(\phi_t, i_t = 0|\epsilon_{1:T})$$
$$= \gamma(\phi_t|i_t = 1)\,\gamma(i_t = 1) + \gamma(\phi_t|i_t = 0)\,\gamma(i_t = 0)\,.$$

The naïve approach, which follows as

$$\gamma(\phi_t, i_t) = \int_{\phi_{t+1}} \sum_{i_{t+1}} p(i_t, \phi_t|\phi_{t+1}, i_{t+1}, \epsilon_{1:t})\,\gamma(\phi_{t+1}, i_{t+1})$$

fails because the first term 'dynamics reversal'

$$\frac{p(\phi_{t+1}, i_{t+1}|i_t, \phi_t)\,\alpha(\phi_t, i_t)}{\int_{\phi_t} \sum_{i_t} p(\phi_{t+1}, i_{t+1}|i_t, \phi_t)\,\alpha(\phi_t, i_t)}$$

has a the mixture of components in the denominator for the case $i_{t+1} = 0$. However, exact smoothing can be derived by utilising the interpretation of the run-length index ρ_t from the filtering recursion, as set out in chapter 3.

The filtered distribution was finally characterised as

$$\alpha(\phi_t) = \alpha(\phi_t|i_t = 1)\,\alpha(i_t = 1) + \sum_{\rho_t} \alpha(\phi_t|\rho_t)\,\alpha(\rho_t|i_t = 0)\,\alpha(i_t = 0)$$

where $\alpha(\phi_t|\rho_t)$ is given by a single truncated Gaussian and $\alpha(\rho_t|i_t = 0)$ is proportional to the weight w_{ρ_t}. The index ρ_t indicates the *number of time-steps since the current regime started.* That is, for fixed ρ_t, we know $i_{t-\rho_t-1} = 1$ and $i_{t-\rho_t:t} = 0$. Between time-steps, $\rho_t = \rho_{t+1} - 1$. Conditioning on ρ_t is equivalent to conditioning on $i_{t-\rho_t-1:t}$, and this extra information already encoded into the components enables us to write a simple recursion for the smoothed posterior.

Non-cointegrated case

We start by first considering $\gamma(\phi_t|i_t = 1) = p^1(\phi_t)$; this can be seen intuitively, or algebraically as

$$\gamma(\phi_t|i_t = 1) \propto p(\phi_t, \epsilon_{t+1:T}|i_t = 1, \epsilon_{1:t}) \propto \alpha(\phi_t|i_t = 1) = p^1(\phi_t)$$

since $\phi_t \perp\!\!\!\perp \phi_{t+1}\,|\,i_t = 1$.

The discrete component

$$\gamma(i_t = 1) = \int_{\phi_{t+1}} p(\phi_{t+1}, \rho_{t+1} = 0, i_{t+1} = 0|\epsilon_{1:T}) + p(i_t = 1|i_{t+1} = 1, \epsilon_{1:t})\,\gamma(i_{t+1} = 1)$$

where the first term is found as the integral of the subset of the components from $t + 1$ indexed by $\rho_{t+1} = 0$, and the second term

$$p(i_t = 1|i_{t+1} = 1, \epsilon_{1:t}) \propto p(i_{t+1} = 1|i_t = 1)\,\alpha(i_t = 1)\,.$$

Cointegrated case

To complete the recursion, it suffices to find the components $\gamma(\phi_t, \rho_t, i_t = 0)$.

$$\gamma(\phi_t, i_t = 0) = \int_{\phi_{t+1}} p(\phi_t, i_t = 0, \phi_{t+1}, i_{t+1} = 0|\epsilon_{1:T}) \\ + p(\phi_t, i_t = 0|i_{t+1} = 1, \epsilon_{1:t})\, \gamma(i_{t+1} = 1) \quad (5.7)$$

and the latter term is easily calculated since

$$p(\phi_t, i_t = 0|i_{t+1} = 1, \epsilon_{1:t}) = \alpha(\phi_t|i_t = 0)\, p(i_t = 0|i_{t+1} = 1, \epsilon_{1:t})$$
$$p(i_t = 0|i_{t+1} = 1, \epsilon_{1:t}) \propto p(i_{t+1} = 1|i_t = 0)\, \alpha(i_t = 0)\,.$$

Finally, the first term in equation (5.7) collapses,

$$\int_{\phi_{t+1}} p(\phi_t, i_t = 0, \phi_{t+1}, i_{t+1} = 0|\epsilon_{1:T}) \\ = \sum_{\rho_{t+1}>0} \int_{\phi_{t+1}} p(\phi_t, i_t = 0, \phi_{t+1}, i_{t+1} = 0, \rho_{t+1}|\epsilon_{1:T}) \\ = \sum_{\rho_{t+1}>0} \gamma(\phi_{t+1} = \phi_t, \rho_{t+1}, i_{t+1} = 0)\,.$$

We see that all of the previous components from $\gamma(\phi_{t+1} = \phi_t, \rho_{t+1} > 0, i_{t+1} = 0)$ survive the recursion without change, and that all of the components from $\alpha(\phi_t, i_t = 0)$ are contributed with a prefactor.

The result is an algorithm for exact inference in this switching model that scales as $O\left(T^2\right)$, from which the calculated posteriors $p(\phi_t, i_t|\epsilon_{1:T})$ are each given as a mixture of Gaussian distributions truncated to the interval $(-1, 1)$.

5.5.5 Learning

As with the simple cointegration model of section 5.3.2, EM is used for parameter estimation in this reset model for intermittent cointegration. For this regime-switching problem, the latent variables are $h \to \{i_{2:T}, \phi_{2:T}\}$, and the observations remain as before, $v \to \epsilon_{1:T}$.

Terms relevant to the optimal solution are the sum of quadratic forms derived in section 5.3.2 with varying ϕ_t,

$$\left\langle (\epsilon_t - \phi_t\epsilon_{t-1})^2 \right\rangle = \epsilon_t^2 - 2\epsilon_t\epsilon_{t-1} \langle \phi_t \rangle + \epsilon_{t-1}^2 \left\langle \phi_t^2 \right\rangle .$$

By retaining the first and second (non-central) moments of each component found while filtering[7], the linear system of equations for the regression parameters can be solved exactly in this switching model, and the variance estimate can be updated accordingly.

Whilst the state transition $p(i_t|i_{t-1})$ and $p(i_2)$ can also in principle be learned on the basis of maximum likelihood with EM, in the examples in this chapter, these quantities are left to be user specified.

[7] Since the sequential inference updates of section 5.3.1 is used for this model, the moments of each component can be calculated analytically as set out in appendix C.

For completeness, the relevant term from the energy is

$$\sum_t \langle \log p(i_t | i_{t-1}) \rangle_{q(i_t, i_{t-1} | \epsilon_{1:T})}$$

and optimally,

$$p(i_t | i_{t-1}) \propto \sum_t q(i_t, i_{t-1} | \epsilon_{1:T})$$

from which it follows that

$$p(i_t = 1 | i_{t-1}) = \frac{\sum_t q(i_{t-1} | i_t = 1, \epsilon_{1:t-1})\, q(i_t = 1 | \epsilon_{1:T})}{\sum_t q(i_{t-1} | \epsilon_{1:T})}$$

$$q(i_{t-1} | i_t = 1, \epsilon_{1:t-1}) \propto q(i_t = 1 | i_{t-1})\, q(i_{t-1} | \epsilon_{1:t-1})$$

and by normalisation,

$$p(i_t = 0 | i_{t-1}) = 1 - p(i_t = 1 | i_{t-1})$$

The initial case is easily shown to have optimum

$$p(i_2) = q(i_2 | \epsilon_{1:T})\,.$$

5.5.6 Point Estimates

For data experiments, we may wish to compare the 'results' of the algorithm with other algorithms for partial cointegration. In contrast with traditional algorithms developed in the econometrics literature, the algorithm in this section takes a Bayesian approach to intermittent cointegration and results in a posterior over all variables, rather than point estimates. For comparison, I take as point estimates the maximum marginal values—that is, after complete smoothed inference, I take each

$$\widehat{i}_t = \arg\max_{i_t} p(i_t | \epsilon_{1:T})$$

and then calculate the posteriors for ϕ_t conditioned on this partition,

$$p\left(\phi_t \middle| \widehat{i}_{2:T}, \epsilon_{1:T}\right)$$

using a simple case of the filtering and smoothing recursions, to find point estimates

$$\widehat{\phi}_t = \arg\max_{\phi_t} p\left(\phi_t \middle| \widehat{i}_{2:T}, \epsilon_{1:T}\right).$$

5.5.7 Experiments

I now consider real-world applications for the partial cointegration model. For these experiments, I ran the inference and learning algorithm for the parameters $\theta = \left\{\alpha, \beta, \sigma^2\right\}$ initialised by OLS, and use the state transition distributions fixed at values designed to match the *a priori* belief about the cointegration regimes.

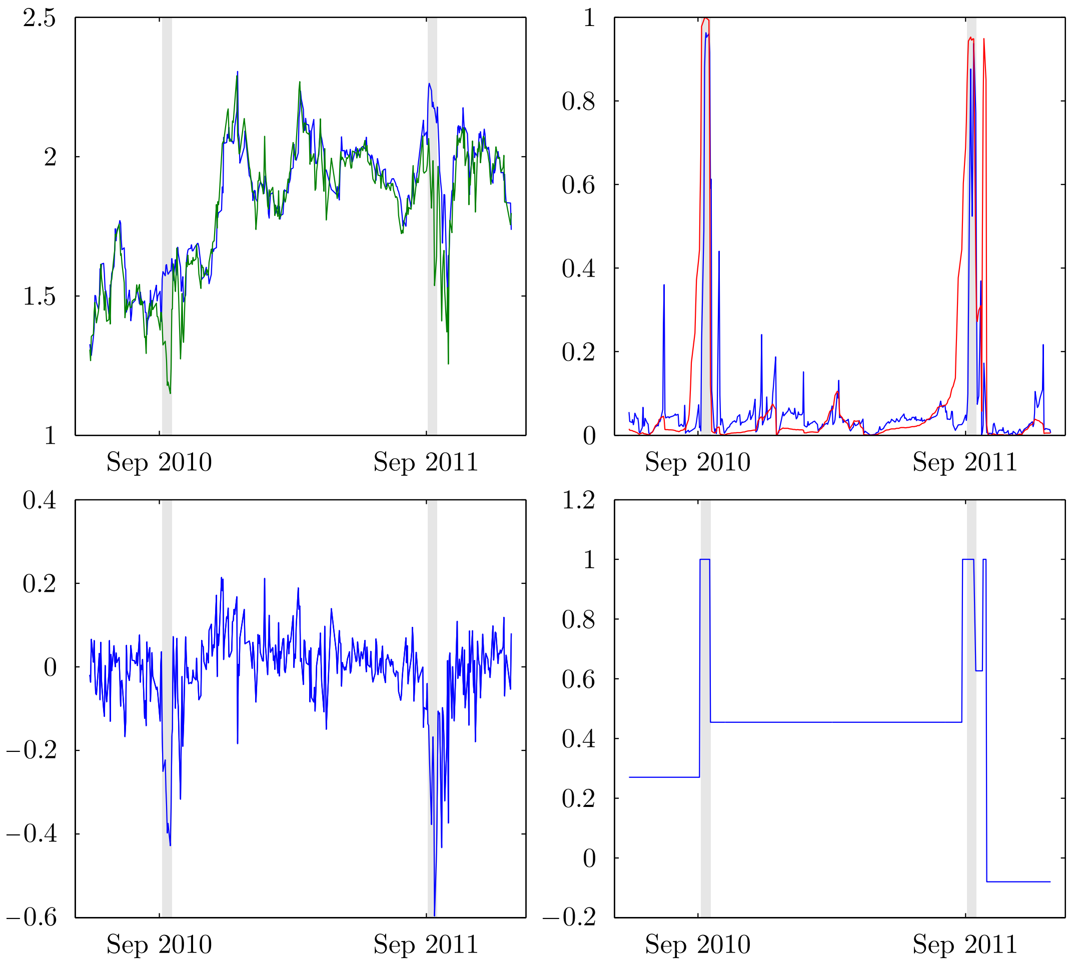


Figure 5.11: Results of learning for the Interconnector gas prices—maintenance occurs each September during the shaded regions. I show (NW) the price series x_t, y_t; (SW) residuals ϵ_t calculated with the final estimates for α, β; (NE) filtered (blue) and smoothed (red) posterior for random walk $i_t = 1$; and (SE) the maximum marginal result for $\widehat{\phi}_t$ conditioned on the chosen partition $\widehat{i}_{2:T}$ as set out in section 5.5.6.

Gas Prices

The Interconnector, as noted in section 5.5, is a sub-sea natural gas pipeline: I took for y UK gas prices[8] and for x the Zeebrugge prices[9]. There are approximately 245 pricing days per year, and the pipeline closes annually for two weeks, so I specified the state transition distribution accordingly, $p(i_t = 1|i_{t-1} = 0) = \frac{1}{230}$ and $p(i_t = 0|t_{t-1} = 1) = \frac{1}{15}$. The algorithm reaches convergence with parameters α, β that show cointegration between the annual maintenance period, see figure 5.11.

Euro-area Bond Yields

The yields on Greek and German 10-year benchmark bonds followed a similar path prior to the Euro-area financial crisis of 2008 onwards. In this section, I use the cointegration switching model to find regions

[8] UK SAP natural gas prices were downloaded from `www.nationalgrid.com/uk/Gas/Data`.

[9] Prices for Zeebrugge are from `www.apxendex.com` and were converted into kWh using 1 kWh=29.3072222 therm.

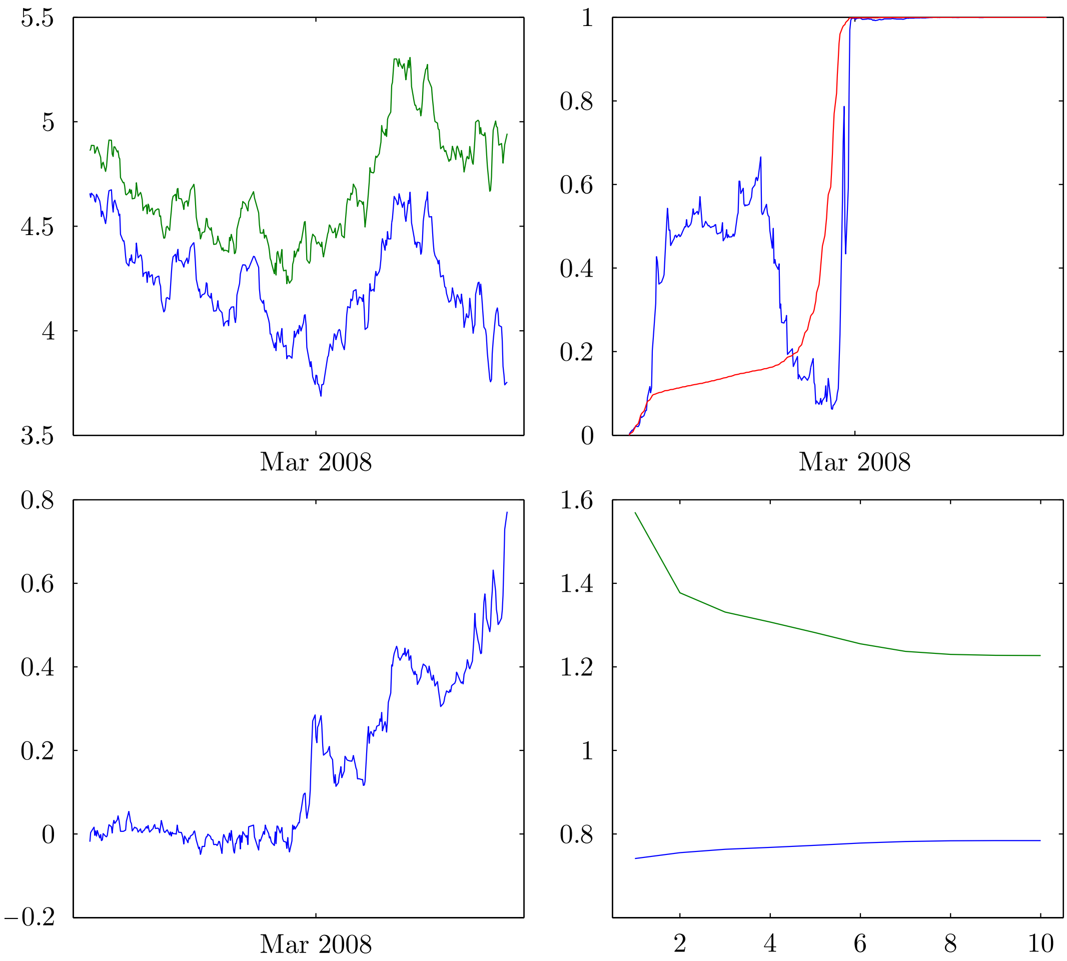


Figure 5.12: Results of learning for the Euro-area bond yields. I show (NW) the yield series x_t, y_t; (SW) residuals ϵ_t calculated with the final estimates for α, β; (NE) filtered (blue) and smoothed (red) posterior for random walk $i_t = 1$; and (SE) the convergence of parameters α, β—note the relative change in the estimate for intercept α (green).

of cointegration, which I expect to hold until the yield on Greek debt spiralled as Greece's debt burden took toll. The results are shown in figure 5.12. These data provide a good example of series which, over the time window shown in the figure, do not show cointegration according to the classical test (the Dickey-Fuller test shows p-value 0.927241, and does not reject the null hypothesis of random walk), but the partial cointegration model does show a segment of cointegration—the classical test passes in the detected region of cointegration, and fails for the remainder.

5.6 Conclusion

This chapter has made three significant contributions to cointegration analysis, a key topic for time-series analysts in a variety of fields—most notably finance. First, I showed that by placing the problem of cointegration in a probabilistic framework and taking a Bayesian approach to the latent variable ϕ, maximum likelihood in a generative model can deliver more accurate estimates of the cointegration coefficients. Second, I showed how we can augment the Bayesian analysis with a random walk model to deliver a combined cointegration detection algorithm that is demonstrably more robust than the classical Engle-Granger two-stage approach. Finally, I showed by application of the reset model framework of chapter 3 that a model for intermittent cointegration with no restriction on the number or location of cointegration regimes is possible, and delivers intuitively meaningful results.

PART III

Conclusions and Extensions

CHAPTER 6

Summary Conclusions

This thesis has contributed novel modelling techniques for sequential data. The contributions, set out in the three chapters comprising Part II, are unified under the title of inference in Bayesian time-series models, and split into three approximately-homogeneous contribution chapters. Chapter 3 discussed probabilistic inference in a particular class of change-point model described as a 'reset model', in which a latent Markov state is maintained with resets possible at multiple unknown points. In chapter 4, the linear-dynamical system was extended to permit a Bayesian approach to variance modelling, and by incorporating insight from the treatment of reset models in chapter 3, a heteroskedastic linear-dynamical system was developed based on a pragmatic approximation. Finally, chapter 5 dealt with the topic of cointegration, which considers a linear relationship between two or more series, and contributed a novel estimation and detection scheme for possibly-cointegrated series inspired by a Bayesian modelling approach.

This chapter seeks to conclude the thesis by drawing together summary results and conclusions of the contributions set out in Part II.

6.1 Switch-Reset Models

Chapter 3 discussed probabilistic inference in reset models and switch-reset models, which have been used in various fields including bioinformatics (Boys and Henderson, 2004), finance (Davis et al., 2008) and signal processing (Fearnhead, 2005). The well-known β message passing algorithm derived from equation (2.2) is applicable and straightforward to derive in respect of reset models, but suffers some drawbacks. First, for certain classes of such model—notably the linear-dynamical system—numerical stability is a concern, as demonstrated in figure 3.2. Second, it is difficult to contrive a linear-time algorithm for smoothed inference, due to the abstract nature of the β components.

To address these issues I went on to derive a correction smoother, based on a redefinition of the reset model in terms of run-length. I then contributed an interpretation of smoothing in terms of messages relating to future resets, which I represent in the bracket smoother of section 3.2.4 with an additional reverse run-length index. The algorithms so defined overcome the numerical difficulties of the β approach and can be implemented with confidence using standard numerically-stable propagation routines in models such as the linear-dynamical system. Moreover, the derivation is didactically useful when considering approximations to the smoothed posterior, since the components are themselves distributions in the latent variable of interest. The resulting approximations based on dropping low weight components in the filtered and smoothed posteriors give a linear-time algorithm that exhibits excellent performance, superior to previous approaches based on α-β smoothing. Further applications include piecewise reset models (widely known simply as changepoint models), for which the inference algorithms are readily transferred.

A switch-reset model was also discussed, motivated by a desire for multiple generating processes, inspired by switching latent Markov models such as the switching linear-dynamical system. The reset nature of the model significantly reduces the complexity in comparison with other switching systems, and the linear-time routines are applicable. The reset models are highly practical and do not suffer from the numerical difficulties apparent in the more general switching models. Furthermore, with robust and effective linear-time smoothing and filtering algorithms, they are inexpensive to deploy.

The reset model framework provides inspiration and insight for other models described later in the thesis including the heteroskedastic linear-dynamical system of chapter 4 and intermittent cointegration model of chapter 5.

6.2 Heteroskedastic Linear-Dynamical System

Based on the observation that variance modelling in the switching linear-dynamical system (and notably the switch-reset variation of chapter 3) are important in the determination of posterior states, chapter 4 introduced a Bayesian approach to variance modelling in the linear-dynamical system.

It is straightforward to show a conjugate form for message passing inference routines in a linear-dynamical system augmented with a variable representing the precision of the sequence. I then went on to consider modelling the precision of the sequence as a process to permit resets (change-points) in the latent chain. Such a model can be developed easily by application of the bracket smoother applicable to any reset model. Unfortunately, exact inference cannot be written in a convenient form when the dynamics in the internal latent Markov state remain unbroken when the variance parameter changes. By application of a pragmatic approximation, I derived a tractable inference scheme, and by appealing to the approximation framework of chapter 3 showed that a linear-time algorithm is available.

The resulting model was shown to be effective on two real-world datasets, resulting in the detection of regions of similar variance in observed data even though differentiating characteristics of those regimes were not known *a priori*. The resulting heteroskedastic linear-dynamical system is a highly practical and scalable routine for detecting variance regimes in multi-dimensional data while allowing for a continuously-evolving latent variable state. Permitting resets in the precision regime has the desired result that the complexity in specifying a model in terms of parameters is significantly reduced compared with the standard switching linear-dynamical system while inference remains inexpensive.

6.3 Bayesian Conditional Cointegration

Cointegration is a prevalent topic in time-series analysis and has received considerable interest in the literature. Whilst the classical Engle-Granger approach is appealing to economists because of its conceptual simplicity, particularly when compared with the more complex method of Johansen, the approach does suffer some weaknesses as set out in section 5.1. In particular, the two steps of the Engle-Granger method make conflicting assumptions: under the null hypothesis of the unit root test in the second stage, the regression of the first stage was spurious. It is also known that the ordinary least-squares method used to estimate the relationship can deliver estimates with considerable bias when the sample size is small, even though the estimator is consistent. Furthermore, the lack of exogeneity of the independent variable on the residuals means the test statistic does not follow a well-known distribution.

I presented a novel scheme for inference in a cointegration model based on Bayesian modelling, which I believe provides a significant contribution to the long-standing problem of inconsistency in establishing cointegration, while retaining the conceptual simplicity of the Engle-Granger technique so appealing to the applied economist.

Three novel techniques for dealing with a linear cointegration relationship for time-series were presented based on the Bayesian scheme. First, I showed how a cointegration relationship can be estimated in a Bayesian framework by using Expectation Maximisation to deliver the maximum likelihood estimator in the cointegration model. Analysis demonstrated that the Bayesian approach can deliver more accurate estimates than ordinary least-squares for a smaller sample. The method in section 5.3 allows estimation of a linear relationship between variables, making no assumption about the underlying generating process by working with the conditional relationship $p(y_{1:T}|x_{1:T})$ not the joint $p(x_{1:T}, y_{1:T})$. Second, I showed how one may check for a unit root in the residuals process by another application of the Bayesian cointegration model, and comparing the data likelihood with a random walk model. When combined with the estimation scheme, the resulting decision algorithm has greater sensitivity than the Engle-Granger approach. In particular, I showed in section 5.4 that the Bayesian algorithm is less likely to deliver a spurious relationship than the classical OLS–unit root testing approach and I believe this to be a significant result. Third, I developed a switching model for the problem of intermittent cointegration, designed to detect a relationship and regimes in which that relationship is a random walk. Uniquely, this reset-model-inspired approach places no restriction on the number or position of the different regimes.

Taken together, these three approaches represent a significant contribution to the literature on estimation and detection methods for cointegrated variables.

CHAPTER 7

Further Work

The results of chapter 3 form an extensible, robust inference framework for latent Markov models which allow internal process resets, underpining the application of change-point models in chapters 4 and 5. In each of these contribution chapters, an individual insight relating to the topic of interest is first described before reset modelling is considered. In chapter 4, I first show a Bayesian approach to precision modelling in a linear-dynamical system, and for the topic of cointegration in chapter 5, novel estimation and detection schemes are contributed before application of the reset model framework.

It is interesting, therefore, to consider how one may further extend the studies in important models of interest throughout the thesis, and this chapter seeks to discuss some potential future directions.

7.1 Switch-Reset Models

Whilst the reset framework of chapter 3 is flexible by construction of the model and algorithms, there are some simple ways that the models can be enhanced. First of all, it is possible to augment the switch-reset model by removing the conditional independence assumption between the internal state s_t and the latent variable h_{t-1} so the selected prevailing state may depend also on the previous continuous state. A further extension of the model is possible by separating the deterministic relationship between occurrences of resets and changes in internal state s_t—this could be considered by introducing the binary reset variable c_t and defining a distribution over the occurrence of a reset $p(c_t|s_t, s_{t-1})$. Such a model could be considered a more general marriage between switching Markov models and reset models, though inference could be a difficult problem.

7.2 Heteroskedastic Linear-Dynamical System

Next, I consider extensions of the variance-resetting model.

7.2.1 Approximation schemes

The heteroskedastic linear-dynamical system of chapter 4 appeals to an approximation to the model posterior to render inference tractable for long, high-dimensional sequences. Conveniently, according to the approximation scheme described, the Gauss-Gamma messages share a common Gaussian component, meaning the inference task requires minimal storage space even for high-dimensional data. Alternative approximation schemes can also be considered however, and an obvious choice is to replace the troublesome dependency on λ_{t-1} in the filtering update in the case of a reset with a *scaled* λ_t. The scale factor may be selected, for example, by considering the Kullback-Leibler divergence,

$$
\begin{aligned}
&KL\Big(\mathcal{N}\big(h_{t-1}\big|f, \lambda_{t-1}^{-1}F\big)\Big\|\,\mathcal{N}\Big(h_{t-1}\Big|f, (\lambda_t\gamma)^{-1}\Big)\Big) \\
&\qquad\qquad = -\Big\langle \log\mathcal{N}\Big(h_{t-1}\Big|f, (\lambda_t\gamma)^{-1}\Big)\Big\rangle_{\mathcal{N}\left(h_{t-1}\middle|f,\lambda_{t-1}^{-1}F\right)p(\lambda_{t-1})p(\lambda_t)} + \text{constant}
\end{aligned}
$$

and we seek to minimise the divergence by choice of scale factor γ,

$$
\begin{aligned}
\widehat{\gamma} &= \arg\max_{\gamma} \Big\langle \log\mathcal{N}\Big(h_{t-1}\Big|f, (\lambda_t\gamma)^{-1}\Big)\Big\rangle_{\mathcal{N}\left(h_{t-1}\middle|f,\lambda_{t-1}^{-1}F\right)p(\lambda_{t-1})p(\lambda_t)} \\
&= \arg\max_{\gamma} \left[\log\sqrt{\gamma} - \frac{\gamma}{2}\Big\langle \lambda_t\left(h_{t-1}-f\right)^2\Big\rangle_{\mathcal{N}\left(h_{t-1}\middle|f,\lambda_{t-1}^{-1}F\right)p(\lambda_{t-1})p(\lambda_t)}\right] \\
&= \arg\max_{\gamma} \left[\log\gamma - \gamma\left\langle\lambda_t\right\rangle_{p(\lambda_t)}\left\langle\lambda_{t-1}^{-1}\right\rangle_{p(\lambda_{t-1})}F\right] = \left(\left\langle\lambda_t\right\rangle\left\langle\lambda_{t-1}^{-1}\right\rangle F\right)^{-1}
\end{aligned}
$$

and we would therefore optimally approximate

$$
\mathcal{N}\big(h_{t-1}\big|f, \lambda_{t-1}^{-1}F\big) \approx \mathcal{N}\big(h_{t-1}\big|f, \lambda_t^{-1}\left\langle\lambda_t\right\rangle\left\langle\lambda_{t-1}^{-1}\right\rangle F\big)\,.
$$

In such an approximation scheme, however, the Gauss-Gamma messages no longer share a common Gaussian component and so inference is less efficient. Myriad alternative approaches could also be considered.

7.2.2 Switching systems

There are a number of ways that one may consider switching extensions of the heteroskedastic linear-dynamical system of chapter 4. First, it is possible to allow for a multimodal prior for λ as a mixture of Gamma distributions by developing a switching model, in which the λ at a reset point may be drawn from one of two or more Gamma distributions with differing *a priori* parameters selected by the prevailing state $s_t \in S$, governed by $p(s_t|s_{t-1})$. Inference in such a model should remain tractable. Second, the example with bee tracking data given in section 4.3.1 shows a good example where a combined heteroskedastic-switching linear-dynamical system may be useful, to model different internal states s_t designed to correspond to different real-world activity; in this case, we may select transition matrices designed to correspond to differentiate the turn left and right phases. In such a model, precision resets would coincide with changes in the internal state s_t.

7.3 Bayesian Conditional Cointegration

A great number of extensions are possible for the cointegration models of chapter 5.

7.3.1 Stationarity Condition

One may broaden the stationarity condition placed on the residuals process ϵ_t, just as the augmented Dickey-Fuller test extends to original basic Dickey-Fuller test. Throughout the chapter I assumed that the process can be modelled as a first-order autoregressive process, and sought to 'test' this process for stationarity. An obvious extension allows for an order-p autoregression,

$$\epsilon_t = \sum_{\tau=1}^{p} \phi_\tau \epsilon_{t-\tau} + \eta_t = \boldsymbol{\phi}^\top \hat{\boldsymbol{\epsilon}}_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}\big(0, \sigma^2\big)$$

which must be a stationary process for cointegration to hold. Hamilton (1994) derived a number of results relating to stability in such processes, which in general can be shown to be stationary when all the roots of the characteristic equation

$$\lambda^p - \phi_1 \lambda^{p-1} - \phi_2 \lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = 0$$

lie in the unit circle. In the case that $p = 1$, this corresponds to the condition used in chapter 5, $|\phi_1| < 1$, which provided a tractable prior for ϕ_1. In the case that $p = 2$, the condition can be written as

$$|\phi_2| < 1, \quad |\phi_1| < 1 - \phi_2$$

from which one may write a prior distribution, although inference is difficult in closed form. The problem may be solved by specifying a suitable prior and applying the well-known information filter, see for example Cappé et al. (2005).

7.3.2 Vector Cointegration

Whilst chapter 5 deals with simple cointegration between two series, we may generalise the notion of cointegration to a vector time series $\mathbf{y}_t$ by writing

$$\begin{aligned} &\boldsymbol{\alpha}^\top \mathbf{y}_t = \epsilon_t \\ &\qquad \epsilon_t = \phi \epsilon_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}\big(0, \sigma^2\big), \quad |\phi| < 1 \end{aligned}$$

in which the cointegrating vector $\boldsymbol{\alpha}$ governs the linear combination responsible for the stationary process ϵ_t.

Using this vector generalisation, based on any estimate of the cointegrating vector inference in the Bayesian cointegration model follows exactly as set out in section 5.3. However, the EM update for $\boldsymbol{\alpha}$ is inevitably different. In this case the energy terms are given as

$$\begin{aligned} &\sum_{t=2}^{T} \langle \log p(\epsilon_t | \epsilon_{t-1}, \phi) \rangle_{q(\phi|\epsilon_{1:T})} \\ &\qquad = -\frac{1}{2\sigma^2} \sum_{t=2}^{T} \left\langle \big(\boldsymbol{\alpha}^\top \mathbf{y}_t - \phi \boldsymbol{\alpha}^\top \mathbf{y}_{t-1}\big)^2 \right\rangle_{q(\phi|\epsilon_{1:T})} - \frac{T-1}{2} \log 2\pi\sigma^2 \end{aligned}$$

and differentiating the terms in $\boldsymbol{\alpha}$ yields

$$2\left[\mathbf{y}_t \mathbf{y}_t^\top - \langle \phi \rangle \left(\mathbf{y}_t \mathbf{y}_{t-1}^\top + \mathbf{y}_{t-1} \mathbf{y}_t^\top\right) + \left\langle \phi^2 \right\rangle \mathbf{y}_{t-1} \mathbf{y}_{t-1}^\top\right] \boldsymbol{\alpha}$$

and seeking a stationary point by setting this to the zero vector $\mathbf{0}$ yields a system of solutions. As set out by Hamilton (1994), it is therefore necessary to impose a restriction such as the first element of $\boldsymbol{\alpha}$ to unity (corresponding to the case set out earlier).

To do this, we may split out the series and write

$$\begin{aligned} & y_t - \boldsymbol{\alpha}^\top \mathbf{x}_t = \epsilon_t \\ & \qquad \epsilon_t = \phi \epsilon_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}\big(0, \sigma^2\big), \quad |\phi| < 1 \end{aligned}$$

and by again differentiating the energy, we find the solution

$$\begin{aligned} \boldsymbol{\alpha} = & \left[\sum_t \mathbf{x}_t \mathbf{x}_t^\top - |\phi| \left(\mathbf{x}_t \mathbf{x}_{t-1}^\top + \mathbf{x}_{t-1} \mathbf{x}_t^\top \right) + \left|\phi^2\right| \mathbf{x}_{t-1} \mathbf{x}_{t-1}^\top \right]^{-1} \\ & \qquad \left[\sum_t y_t \mathbf{x}_t - |\phi| \left(y_t \mathbf{x}_{t-1} + y_{t-1} \mathbf{x}_t \right) + \left|\phi^2\right| y_{t-1} \mathbf{x}_{t-1} \right] \end{aligned}$$

and by defining

$$\mathbf{X}_T = \begin{bmatrix} \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_T^\top \end{bmatrix}, \mathbf{X}_{T-1} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_{T-1}^\top \end{bmatrix}, \mathbf{y}_T = \begin{bmatrix} y_2 \\ \vdots \\ y_T \end{bmatrix}, \mathbf{y}_{T-1} = \begin{bmatrix} y_1 \\ \vdots \\ y_{T-1} \end{bmatrix}$$

we can write the solution as

$$\begin{aligned} \boldsymbol{\alpha} = & \left[\mathbf{X}_T^\top \mathbf{X}_T - |\phi| \left(\mathbf{X}_T^\top \mathbf{X}_{T-1} + \mathbf{X}_{T-1}^\top \mathbf{X}_T \right) + \left|\phi^2\right| \mathbf{X}_{T-1}^\top \mathbf{X}_{T-1} \right]^{-1} \\ & \qquad \left[\mathbf{X}_T^\top \mathbf{y}_T - |\phi| \left(\mathbf{X}_{T-1}^\top \mathbf{y}_T + \mathbf{X}_T^\top \mathbf{y}_{T-1} \right) + \left|\phi^2\right| \mathbf{X}_{T-1}^\top \mathbf{y}_{T-1} \right] \end{aligned}$$

from which the analogy to the OLS solution is clear. Whereas OLS requires mean-zero errors, by contrast the inference-update estimation algorithm of this thesis allows serial correlation provided that the errors form a stationary process.

7.3.3 Cointegration regimes

Whilst the Bayesian approach to intermittent cointegration of chapter 5 is both flexible and simple, there are limitations. In particular, the switching model requires that the regions of cointegration are governed by a time-invariant linear relationship. For example, in the case that deviations from the relationship in the random-walk segments are permanent, the model presented in section 5.5 would struggle to properly estimate a relationship since piecewise-constant values of the constant term α would be required. Such considerations provide opportunities for future research.

7.3.4 Further Bayesian Analysis

In the case of the cointegration, the parameters of fundamental interest are the linear coefficients α and β. The classical approach estimates these with ordinary least-squares, and goes on to test for stationary residuals. A natural Bayesian approach would therefore be to model these parameters as variables;

the approach in this thesis is somewhat different. Rather, the model seeks an estimation of the linear coefficients, but without the inconsistency of requiring the residuals process to have constant, zero mean. Whilst it is tempting to consider how to extend the model to find a posterior for the coefficients, I have left this more computationally-advanced problem for further work. First, one may provide prior distributions for the parameters of the cointegration relationship and seek to calculate the posteriors, although this is likely to require numerical methods for integration. Secondly for the purpose of detecting cointegration, one could approach the problem by marginalising the cointegration parameters from the model.

One may consider approximation schemes to approximate the posterior of the model by for example factorising the posterior (Jordan et al. (1999) provide a useful introduction), or alternatively take a sampling approach (see for example Chen et al. (2000)). The benefit of the recursive maximum likelihood approach taken in this article is that the algorithm is easily implemented since the inference is both conceptually and computationally simple.

An easy way to augment the model would be to place a prior over the variance σ^2. This may be done in closed analytical form since the inverse Gamma distribution is a conjugate prior for the variance in a Gaussian distribution, as shown in chapter 4.

7.3.5 An Online Algorithm

In big data applications, for example high-frequency finance, it is necessary to consider how one may implement an online algorithm for Bayesian cointegration. An interesting idea is to use online EM which seeks to update the estimates as new data are observed, see for example Cappé and Moulines (2009) or Neal and Hinton (1998).

Appendices

APPENDIX A

Probability Density Functions

This appendix gives the definitions and some properties for the standard probability distributions used for inference.

A.1 Dirac Delta Distribution

The first standard distribution is an implied density for a trivial distribution over a degenerate continuous variable x which places all mass at a single point a. This distribution, known as the Dirac delta distribution has a probability density function denoted by

$$\delta(x - a)$$

and whilst no analytical form exists for the density, the distribution is fully described the property that

$$\int_x \delta(x - a)\, f(x) \equiv f(a)\,.$$

A.2 Uniform Distribution

The uniform distribution is often used as an "uninformative prior", and assigns equal mass to all possible values in the distribution. When the variable is discrete, the probability is trivially set by dividing the total mass by the number of possible discrete values S,

$$p(s) = \frac{1}{S}$$

but when the variable is continuous it can be more difficult to specify a density. Fortunately when the range of possible values is limited to some interval the density is given as

$$\mathcal{U}(x|\,(a, b)) \equiv \frac{1}{b - a}\,[x \in (a, b)]$$

which assigns equal mass within the interval and zero mass elsewhere. When a uniform distribution is required over all real values $x \in \mathbb{R}$ the distribution is *improper* since it is impossible to normalise the density.

A.3 Normal Distribution

The multivariate Gaussian distribution for variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is given by the density

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}} \exp -\tfrac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)$$

I also write for the density function $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\,.$$

A.4 Gamma Distribution

For the gamma, we use the reciprocal notation to ease the algebra.

$$\mathit{Gam}(x|a,b) \equiv \frac{b^a x^{a-1}}{\Gamma\left(a\right)} \exp -bx$$

for which the parameters can be found from required mean μ and standard deviation σ by $a = \left(\frac{\mu}{\sigma}\right)^2$ and $b = \frac{\mu}{\sigma^2}$. Alternatively, we can find the scale parameter b from a desired mode m and shape a with $b = \frac{a-1}{m}$.

A.5 Student's t distribution

For the student's t distribution, I use the three-parameter density[1] given for variable $\mathbf{x}$ ($\dim \mathbf{x} = d$) by

$$\mathit{Student}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\det \nu\pi\boldsymbol{\Sigma}}} \left[1 + \frac{1}{\nu}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{-\frac{\nu+d}{2}}$$

[1] This characterisation of the student-t density is sometimes known as the *location-scale* form.

APPENDIX B

Bayes' Theorem for Gaussians

The family of Gaussian-distributed variables is, generally speaking, well-behaved under Bayesian manipulation of linear combinations. This appendix sets out the derivations of several utility results, most of which are well-known results for inference with Gaussian variables. The aim is to present the results in a coherent, clear, and pragmatic manner.

Fact B.1 *Marginal* $p(x) = \int_y p(x|y)$

If

- $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^\top & \boldsymbol{\Sigma}_{yy} \end{bmatrix} \right)$

then

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$

Proof. The derivation is rather long-winded, and requires calculation of the Schur complement as well as completing the square of the Gaussian p.d.f. to integrate out the variable. For a full work-through, see Bishop (2007, Section 2.3.2). ■

Fact B.2 *Joint* $p(x, y) = p(y|x)\,p(x)$

If

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}\big(\mathbf{M}\mathbf{x} + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\big)$

then

- $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_x + \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\mathbf{M}^\top \\ \mathbf{M}\boldsymbol{\Sigma}_x^\top & \mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\top + \boldsymbol{\Sigma}_y \end{bmatrix} \right)$

Proof. We can write $\mathbf{y} = \mathbf{M}\mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(\mu_y, \mathbf{\Sigma}_y)$. Then the covariance can be written[1,2] $\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \rangle = \left\langle \Delta\mathbf{x}\left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)^\top \right\rangle = \langle \Delta\mathbf{x}\Delta\mathbf{x}^\top \rangle \mathbf{M}^\top + \langle \Delta\mathbf{x}\Delta\epsilon^\top \rangle$. Since $\langle \Delta\mathbf{x}\Delta\epsilon^\top \rangle = 0$ we therefore have $\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \rangle = \mathbf{\Sigma}_x\mathbf{M}^\top$. Similarly, $\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \rangle = \left\langle \left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)\left(\mathbf{M}\Delta\mathbf{x} + \Delta\epsilon\right)^\top \right\rangle = \mathbf{M}\langle \Delta\mathbf{x}\Delta\mathbf{x}^\top \rangle \mathbf{M}^\top + \langle \Delta\epsilon\Delta\epsilon^\top \rangle = \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y$. The result follows. ■

Corollary B.3 *Marginal* $p(y) = \int_x p(y|x)\,p(x)$ *(linear transform of a Gaussian)*

If

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \mathbf{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}\big(\mathbf{M}\mathbf{x} + \boldsymbol{\mu}_y, \mathbf{\Sigma}_y\big)$

then

- $p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x}) = \mathcal{N}\big(\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\big)$.

Proof. Immediate from Fact B.1 and Fact B.2. ■

Fact B.4 *Conditioning* $p(x|y) \propto p(x,y)$

If

- $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy}\\ \mathbf{\Sigma}_{xy}^\top & \mathbf{\Sigma}_{yy}\end{bmatrix}\right)$

then

- $\mathbf{x}|\mathbf{y} \sim \mathcal{N}\big(\boldsymbol{\mu}_x + \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}_y\right), \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{xy}^\top\big)$

Proof. Again the derivation is long-winded, and appeals to the Schur complement. See Bishop (2007, Section 2.3.1). ■

Corollary B.5 *Conditioning* $p(x|y) \propto p(y|x)\,p(x)$ *(inverse linear transform, dynamics reversal)*

If

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \mathbf{\Sigma}_x)$ and
- $\mathbf{y}|\mathbf{x} \sim \mathcal{N}\big(\mathbf{M}\mathbf{x} + \boldsymbol{\mu}_y, \mathbf{\Sigma}_y\big)$

then

- $\mathbf{x}|\mathbf{y} \sim \mathcal{N}\big(\mathbf{R}\left(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right) + \boldsymbol{\mu}_x, \mathbf{\Sigma}_x - \mathbf{R}\mathbf{M}\mathbf{\Sigma}_x^\top\big)$ where
- $\mathbf{R} = \mathbf{\Sigma}_x\mathbf{M}^\top\left(\mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\right)^{-1}$

Equivalently, we have $\mathbf{x} = \mathbf{R}\mathbf{y} + \epsilon$, $\epsilon \sim \mathcal{N}\big(\boldsymbol{\mu}_x - \mathbf{R}\left(\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}_y\right), \mathbf{\Sigma}_x - \mathbf{R}\mathbf{M}\mathbf{\Sigma}_x^\top\big)$.

Proof. From Fact B.2, we have

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x\\ \mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_x & \mathbf{\Sigma}_x\mathbf{M}^\top\\ \mathbf{M}\mathbf{\Sigma}_x^\top & \mathbf{M}\mathbf{\Sigma}_x\mathbf{M}^\top + \mathbf{\Sigma}_y\end{bmatrix}\right) \tag{*}$$

[1] *Displacement* of a variable $\mathbf{x}$ is given by $\Delta\mathbf{x} = \mathbf{x} - \langle\mathbf{x}\rangle$.

[2] $\mathbf{y} = \mathbf{M}\mathbf{x} + \epsilon \Longrightarrow \Delta\mathbf{y} = \mathbf{y} - \langle\mathbf{y}\rangle = \mathbf{M}\mathbf{x} + \epsilon - \langle\mathbf{M}\mathbf{x} + \epsilon\rangle = \mathbf{M}\mathbf{x} + \epsilon - \mathbf{M}\langle\mathbf{x}\rangle - \langle\epsilon\rangle = \mathbf{M}\Delta\mathbf{x} + \Delta\epsilon$

and by Fact B.4,

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}\big(\boldsymbol{\mu}_x + \mathbf{R}\left(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right), \boldsymbol{\Sigma}_x - \mathbf{R}\mathbf{M}\boldsymbol{\Sigma}_x^\top\big)$$

from which the result follows.

Alternative derivation (originally by and based on Barber (2012, Chapter 8)) follows equation (*) by aiming to write $\mathbf{x} = \mathbf{R}\mathbf{y} + \epsilon$ for Gaussian ϵ with $\left\langle \Delta\epsilon\Delta\mathbf{y}^\top \right\rangle = 0$. Consider the covariance

$$\begin{aligned}
\left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \right\rangle &= \left\langle (\mathbf{R}\Delta\mathbf{y} + \Delta\epsilon)\,\Delta\mathbf{y}^\top \right\rangle \\
&= \mathbf{R}\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \right\rangle + \cancel{\left\langle \Delta\epsilon\Delta\mathbf{y}^\top \right\rangle} \\
\mathbf{R} &= \left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \right\rangle \left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \right\rangle^{-1}
\end{aligned}$$

from equation equation (*) we have $\left\langle \Delta\mathbf{x}\Delta\mathbf{y}^\top \right\rangle = \boldsymbol{\Sigma}_x\mathbf{M}^\top$ and $\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \right\rangle = \mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\top + \boldsymbol{\Sigma}_y$. The desired mean and covariance are therefore obtained from

$$\begin{aligned}
\langle \epsilon \rangle &= \langle \mathbf{x} \rangle - \mathbf{R}\langle \mathbf{y} \rangle = \boldsymbol{\mu}_x - \mathbf{R}\left(\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}_y\right) \\
\left\langle \Delta\epsilon\Delta\epsilon^\top \right\rangle &= \left\langle \Delta\mathbf{x}\Delta\mathbf{x}^\top \right\rangle - \mathbf{R}\left\langle \Delta\mathbf{y}\Delta\mathbf{y}^\top \right\rangle \mathbf{R}^\top = \boldsymbol{\Sigma}_x - \mathbf{R}\mathbf{M}\boldsymbol{\Sigma}_x^\top
\end{aligned}$$

■

Corollary B.6 *Joint Conditioning* $p(x)\,p(y|x) = p(y)\,p(x|y)$

We may write

$$\begin{aligned}
&\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)\,\mathcal{N}\big(\mathbf{y}\big|\mathbf{M}\mathbf{x} + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\big) \\
&\quad = \mathcal{N}\big(\mathbf{y}\big|\mathbf{M}\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\top + \boldsymbol{\Sigma}_y\big)\,\mathcal{N}\big(\mathbf{x}\big|\mathbf{R}\left(\mathbf{y} - \mathbf{M}\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\right) + \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x - \mathbf{R}\mathbf{M}\boldsymbol{\Sigma}_x^\top\big)
\end{aligned}$$

where $\mathbf{R} = \boldsymbol{\Sigma}_x\mathbf{M}^\top\left(\mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\top + \boldsymbol{\Sigma}_y\right)^{-1}$.

Proof. The result follows from Corollary B.5 and Corollary B.3. ■

APPENDIX C

Truncated Gaussian Moments

This appendix sets out some analytical results relevant to the cointegration model set out in chapter 5. In particular, I derive reults relating to the statistics of the posterior of ϕ according to the sequential inference scheme of section 5.3.1, useful for the intermittent cointegration model derived in section 5.5.

Each truncated Gaussian component in the posterior for ϕ is written in the form $p(\phi)\,\mathcal{N}(\phi|f, F)$.

C.1 Normalisation

The normalisation constant for each Gaussian component is used for calculating the likelihood and normalising the inference messages. For each Gaussian component, we require

$$\int_\phi p(\phi)\,\mathcal{N}(\phi|f, F) = \int_{-1}^{1} \frac{1}{2}\mathcal{N}(\phi|f, F)\ \mathrm{d}\phi$$

which is calculated easily since

$$\int_a^b \mathcal{N}(x|\mu, \sigma^2)\ \mathrm{d}x = \left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]_a^b$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

C.2 Moments

For the update in each Expectation-Maximisation iteration, the first and second non-central moments $\langle\phi\rangle$ and $\langle\phi^2\rangle$ for each posterior component are required. These can also be calculated exactly since

$$\begin{aligned}\int_a^b x\mathcal{N}(x|\mu, \sigma^2)\ \mathrm{d}x &= \int_{a-\mu}^{b-\mu} y\mathcal{N}(y|0, \sigma^2)\ \mathrm{d}y + \mu\int_{a-\mu}^{b-\mu} \mathcal{N}(y|0, \sigma^2)\ \mathrm{d}y \\ &= -\sigma^2\left[\mathcal{N}(x|\mu, \sigma^2)\right]_a^b + \mu\left[\Phi\left(\frac{x-\mu}{\sigma}\right)\right]_a^b\end{aligned}$$

and

$$\begin{aligned}\int_a^b x^2 \mathcal{N}\big(x\big|\mu,\sigma^2\big)\ \mathrm{d}x &= \int_{a-\mu}^{b-\mu} y^2 \mathcal{N}\big(y\big|0,\sigma^2\big)\ \mathrm{d}y + 2\mu \int_{a-\mu}^{b-\mu} y \mathcal{N}\big(y\big|0,\sigma^2\big)\ \mathrm{d}y \\ &\quad + \mu^2 \int_{a-\mu}^{b-\mu} \mathcal{N}\big(y\big|0,\sigma^2\big)\ \mathrm{d}y \\ &= \int_{a-\mu}^{b-\mu} \frac{y}{\sqrt{2\pi\sigma^2}} \frac{\mathrm{d}}{\mathrm{d}y} \left(-\sigma^2 \exp - \frac{y^2}{2\sigma^2}\right)\ \mathrm{d}y \\ &\quad - 2\mu\sigma^2 \left[\mathcal{N}\big(x\big|\mu,\sigma^2\big)\right]_a^b + \mu^2 \left[\Phi\bigg(\frac{x-\mu}{\sigma}\bigg)\right]_a^b \\ &= -\sigma^2 \left[(x+\mu)\,\mathcal{N}\big(x\big|\mu,\sigma^2\big)\right]_a^b + \big(\mu^2+\sigma^2\big) \left[\Phi\bigg(\frac{x-\mu}{\sigma}\bigg)\right]_a^b .\end{aligned}$$

REFERENCES

R. P. Adams and D. J. C. MacKay. Bayesian Online Changepoint Detection. Technical report, University of Cambridge, 2007.

C. Alexander. *Practical Financial Econometrics*, volume II of *Market Risk Analysis*. Wiley, 2008.

D. Alspach and H. Sorenson. Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

H. Attias. A Variational Bayesian Framework for Graphical Models. *Advances in Neural Information Processing Systems*, 12(1-2):209–215, 2000.

I. E. Auger and C. E. Lawrence. Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.

A. Banerjee, J. J. Dolado, D. F. Hendry, and G. W. Smith. Exploring Equilibrium Relationships in Econometrics Through Static Models: Some Monte Carlo Evidence. *Oxford Bulletin of Economics and Statistics*, 48(3):253–277, 1986.

D. Barber. Expectation Correction for Smoothed Inference in Switching Linear Dynamical Systems. *The Journal of Machine Learning Research*, 7:2515–2540, 2006.

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

D. Barber and A. T. Cemgil. Graphical Models for Time Series. *IEEE Signal Processing Magazine*, 27 (6):18–28, 2010.

D. Barber and S. Chiappa. Unified Inference for Variational Bayesian Linear Gaussian State-Space Models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19, pages 81–88, 2006.

D. Barry and J. A. Hartigan. Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260–279, 1992.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

T. Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31 (3):307–327, 1986.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

R. J. Boys and D. A. Henderson. A Bayesian Approach to DNA Sequence Segmentation. *Biometrics*, 60 (3):573–581, 2004.

C. Bracegirdle and D. Barber. Switch-Reset Models : Exact and Approximate Inference. In *Proceedings of The Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15. JMLR, 2011.

C. Bracegirdle and D. Barber. Bayesian Conditional Cointegration. In *29th International Conference on Machine Learning (ICML)*, 2012.

O. Cappé and E. Moulines. On-line ExpectationMaximization Algorithm for Latent Data Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

C. K. Carter and R. Kohn. Markov Chain Monte Carlo in Conditionally Gaussian State Space Models. *Biometrika*, 83(3):589–601, 1996.

C. M. Carvalho and H. F. Lopes. Simulation-Based Sequential Analysis of Markov Switching Stochastic Volatility Models. *Computational Statistics & Data Analysis*, 51(9):4526–4542, 2007.

M. J. Cassidy and W. D. Penny. Bayesian Nonstationary Autoregressive Models for Biomedical Signal Analysis. *IEEE Transactions on Biomedical Engineering*, 49(10):1142–1152, 2002.

P. Cheeseman. An Inquiry Into Computer Understanding. *Computational Intelligence*, 4(2):58–66, 1988.

M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer, 2000.

J. E. H. Davidson, D. F. Hendry, F. Srba, and S. Yeo. Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom. *The Economic Journal*, 88(352):661–692, 1978.

R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Break Detection for a Class of Nonlinear Time Series Models. *Journal of Time Series Analysis*, 29(5):834–867, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

D. A. Dickey and W. A. Fuller. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 74(366):427–431, 1979.

D. A. Dickey, D. W. Jansen, and D. L. Thornton. A Primer on Cointegration with an Application to Money and Income. *Federal Reserve Bank of St. Louis Review*, (Mar):58–78, 1991.

O. Dikmen and A. T. Cemgil. Inference and Parameter Estimation in Gamma Chains. Technical Report CUED/F-INFENG/TR.596, University of Cambridge, 2008.

A. Doucet, S. Godsill, and C. Andrieu. On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing*, 10(3):197–208, 2000.

R. F. Engle. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50(4):987–1007, 1982.

R. F. Engle and C. W. J. Granger. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276, 1987.

P. Fearnhead. Exact Bayesian Curve Fitting and Signal Segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166, 2005.

P. Fearnhead and P. Clifford. On-Line Inference for Hidden Markov Models via Particle Filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

P. Fearnhead and Z. Liu. Online Inference for Multiple Changepoint Problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

K. Fergusson and E. Platen. On the Distributional Characterization of Daily Log-Returns of a World Stock Index. *Applied Mathematical Finance*, 13(01):19–38, 2006.

E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 457–464, 2008.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

R. Garnett, M. A. Osborne, and S. J. Roberts. Sequential Bayesian Prediction in the Presence of Changepoints. In *Proceedings of the 26th International Conference on Machine Learning*, pages 345–352, 2009.

W. R. Gilks and C. Berzuini. Following a Moving Target—Monte Carlo Inference for Dynamic Bayesian Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113, 1993.

C. W. J. Granger. Developments in the Study of Cointegrated Economic Variables. *Oxford Bulletin of Economics and Statistics*, 48(3):213–228, 1986.

C. W. J. Granger and P. Newbold. Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2): 111–120, 1974.

A. W. Gregory and B. E. Hansen. Residual-Based Tests for Cointegration in Models with Regime Shifts. *Journal of Econometrics*, 70(1):99–126, 1996.

J. D. Hamilton. *Time Series Analysis*. Cambridge University Press, 1994.

R. Harris and R. Sollis. *Applied Time Series Modelling and Forecasting*. Wiley, 2003.

A. Hatemi-J. Tests for Cointegration with Two Unknown Regime Shifts with an Application to Financial Market Integration. *Empirical Economics*, 35(3):497–505, 2008.

R. G. Jarrett. A Note on the Intervals Between Coal-Mining Disasters. *Biometrika*, 66(1):191–193, 1979.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

W. H. Jefferys and J. O. Berger. Sharpening Ockhams Razor on a Bayesian Strop. Technical Report 91-44C, 1991.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 1961.

S. Johansen. Statistical Analysis of Cointegration Vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254, 1988.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal Detection of Changepoints with a Linear Computational Cost. *Arxiv preprint arXiv:1101.1438*, 2011.

C. J. Kim. Dynamic Linear Models with Markov-Switching. *Journal of Econometrics*, 60(1-2):1–22, 1994.

C. J. Kim and C. R. Nelson. *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press, 1999.

J. Y. Kim. Inference on Segmented Cointegration. *Econometric Theory*, 19(4):620–639, 2003.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

T. P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369. Morgan Kaufmann, 2001.

T. P. Minka. Divergence Measures and Message Passing. Technical Report MSR-TR-2005-173, Microsoft Research, Cambridge, UK, 2005.

K. P. Murphy. Hidden Semi-Markov Models (HSMMs). Technical report, 2002.

R. M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

R. M. Neal and G. E. Hinton. A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In *Learning in Graphical Models*, pages 355–368. Kluwer, 1998.

J. J. K. Ó Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

S. M. Oh, J. M. Rehg, and F. Dellaert. Parameterized Duration Mmodeling for Switching Linear Dynamic Systems. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1694–1700, 2006.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115, 1954.

E. S. Page. A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika*, 42(3-4): 523–527, 1955.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

H. E. Rauch, F. Tung, and C. T. Striebel. Maximum Likelihood Estimates of Linear Dynamic Systems. *Journal of the American Institute of Aeronautics and Astronautics*, 3(8):1445–1450, 1965.

S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian Processes for Timeseries Modelling. *Philosophical Transactions of the Royal Society (Part A)*, 2012. To appear.

S. E. Said and D. A. Dickey. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71(3):599–607, 1984.

A. J. Scott and M. Knott. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3):507–512, 1974.

B. Seong, S. K. Ahn, and P. A. Zadrozny. Cointegration Analysis with Mixed-Frequency Data. CESifo Working Paper Series 1939, CESifo Group Munich, 2007.

M. E. Tipping and N. D. Lawrence. Variational Inference for Student-t Models: Robust Bayesian Interpolation and Generalised Component Analysis. *Neurocomputing*, 69(1):123–141, 2005.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

R. E. Turner and M. Sahani. Two Problems With Variational Expectation Maximisation for Time-Series Models. In *Bayesian Time-Series Models*, pages 104–124. Cambridge University Press, 2011.

M. Verhaegen and P. Van Dooren. Numerical Aspects of Different Kalman Filter Implementations. *IEEE Transactions on Automatic Control*, 31(10):907–917, 2002.

G. Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*. Wiley, 2004.

P. K. Watson and S. S. Teelucksingh. *A Practical Introduction to Econometric Methods: Classical and Modern*. University of West Indies Press, 2002.

J. M. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western, 2009.

X. Xuan and K. Murphy. Modeling Changing Dependency Structure in Multivariate Time Series. In *24th International Conference on Machine Learning (ICML)*, pages 1055–1062. ACM, 2007.