The University of Adelaide

Inference of Markovian-regime-switching models with application to South Australian electricity prices

by

Angus Lewis

A thesis submitted in partial fulfilment for the
degree of Master of Philosophy

in the

Faculty of Engineering, Computer and Mathematical Science
School of Mathematical Sciences

December 2018

THE UNIVERSITY
of ADELAIDE

Declaration of Authorship

I, Angus Lewis, declare that this thesis titled, 'Inference of Markovian-regime-switching models with application to South Australian electricity prices' and the work presented in it are my own.

- I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.
- I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.
- I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: _____

Date: _____

Abstract

Markovian-Regime-Switching (MRS) models are commonly used for modelling economic time series, including electricity prices. In this application it is common to include *independent regimes* as these can more accurately capture the dynamics of electricity prices compared to traditional MRS models. The advantage of independent regime MRS specifications is that they allow us to seperate dynamics between regimes. Despite their popularity, parameter inference for MRS models with independent regimes is underdeveloped. Until this thesis, there was no computationally feasible method to evaluate the likelihood of, or find maximum likelihood estimate for, MRS models with independent regimes. Moreover, there are no good discussions of Bayesian methods for such models applied to electricity prices. In this thesis we develop both maximum likelihood and Bayesian inference methodologies for MRS models with independent regimes, and use simulations to investigate their behaviours. We use our methods to investigate the South Australian wholesale electricity market, and find evidence of a significant jump in price volatility which coincides with the closure of South Australia's only coal generation facility, and therefore a significant change in market structure. Our work also suggests that Bayesian methods can be advantageous compared to maximum likelihood, since Bayesian methods can avoid issues with inferring parameters of shifted distributions, which are commonly used in this context.

Acknowledgements

First and foremost, I would like to acknowledge my supervisors, Giang Nguyen and Nigel Bean. In all regards they are brilliant at what they do. They were always willing to make time for me, even regarding the not-so-exciting task of editing this thesis. I am very much looking forward to working with them over the next few years.

I also want to thank other students and academics in the Maths faculty here at Adelaide. I recall having useful conversations with Gary Glonek, Andrew Black, Joshua Ross, Jono Tuke, and there would be others I have missed too, so thank you.

I want to thank my close family for their support, especially my mum, Kylie, my partner, Alice, and my brother, Jack. I particularly want to thank Alice for putting up with my long hours and minimal pay.

Thanks also to my friends and peers, Vanessa and Caitlin, who provided much needed sources of empathy.

Finally, a thanks to ACEMS, who have generously provided much appreciated resources to aid my development.

Contents

List of Figures

List of Tables

Symbols

$\mathbb{E}$	expectation operator
$\mathbb{P}$	probability measure
$\mathbb{P}^{\boldsymbol{\theta}}$	probability measure with parameter $\boldsymbol{\theta}$
$f^{\boldsymbol{\theta}}$	probability density with parameter $\boldsymbol{\theta}$
$f(\boldsymbol{x}\|\boldsymbol{\theta})$	the posterior distribution
F	cumulative distribution function
$\ell(\theta)$	the log-likelihood
$L(\theta)$	the likelihood
$\boldsymbol{\theta}_n$	the value of $\boldsymbol{\theta}$ at the nth iteration of the EM algorithm
$\boldsymbol{\theta}^{(n)}$	the value of $\boldsymbol{\theta}$ at the nth iteration in an MCMC chain
$\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$	parameter vector
$\alpha, \phi, \sigma, \mu, q, p_{ij}$	various parameters
$\widehat{\theta}$	hat ($\widehat{\ }$) denotes the maximum likelihood estimate (MLE)
$\boldsymbol{\Theta}$	parameter space
$\mathbb{N} = \{0, 1, 2, ...\}$	the natural numbers
$\mathbb{N}_+ = \{1, 2, ...\}$	the positive integers
$\mathbb{I}$	the indicator function
$p_{ij} = \mathbb{P}(R_{t+1} = j \| R_t = i)$	a transition probability
$\boldsymbol{R} = (R_0, ..., R_T)$	a (hidden) regime sequence
R_n	the nth element of $\boldsymbol{R}$
$\boldsymbol{H}_n = (\boldsymbol{N}_n, R_n)$	the nth element of the augmented regime sequence

$\boldsymbol{N}_n = (N_{t,0}, N_{t,1}, ..., N_{t,k})$	a vector of counters
$N_{t,i}$	the ith element of $\boldsymbol{N}_t$
$\boldsymbol{x} = \boldsymbol{x}_{0:T} = (x_0, ..., x_T)$	the complete sequence of observed values (prices)
$\boldsymbol{x}_{r:s} = (x_r, x_{r+1}, ..., x_s)$	observed values between times r and s, $r \leq s$
$\{X_t\}_{t\in\mathbb{N}}$ or $\{X_t\}$	a sequence
$\{t_i \vert t \in \mathbb{N}\}$	a set
$\{\varepsilon_t\}$	a sequence of N(0,1) random variables
$\mathcal{S} = \{1, 2, ..., M\}$	the state space of R_t
$\mathcal{S}_{AR} = \{1, 2, ..., k\}$	the states with AR(1) components
M	the total number of regimes in a model
k	the total number of AR(1) regimes in a model
$\mathcal{S}_{AR}^c = S \setminus S_{AR} = \{k+1, ..., M\}$	the compliment of S_{AR}
$\mathcal{S}^{(t)}$	all states that $\boldsymbol{N}_t$ can be in with positive probability at time t
Δ_i	at time t, this is a state denoting there is no time ℓ in $0, 1, ..., t-1$ where $R_\ell = i$
$\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)$ $:= f^{\boldsymbol{\theta}}_{\boldsymbol{H}_t, X_t \vert \boldsymbol{X}_{0:t-1}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, R_t), x_t \vert \boldsymbol{x}_{0:t-1})$	the joint density of an observation x_t and augmented hidden regime $\boldsymbol{H}_t$, given all previous observations, $\boldsymbol{x}_{0:t-1}$
$c^{(t)} := f^{\boldsymbol{\theta}}_{X_t \vert \boldsymbol{X}_{0:t-1}}(x_t \vert \boldsymbol{x}_{0:t-1})$	the conditional density of x_t given all previous observations, $\boldsymbol{x}_{0:t-1}$
$\widehat{\alpha}_{\boldsymbol{N}_t}^{(t)}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) \vert \boldsymbol{x}_{0:t})$	forward/filtered probabilities/inferences
$a_{\boldsymbol{N}_t}^{(t)}(i) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i) \vert \boldsymbol{x}_{0:t-1})$	prediction probabilities
$\gamma_{\boldsymbol{N}_t}^{(t)}(i) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i) \vert \boldsymbol{x}_{0:T})$	smoothed probabilities/inferences
$\propto$	proportional to
η_{ij}	the number of transitions from state i to state j in $\boldsymbol{R}$
$\mathcal{O}(T)$	indicates an algorithm has order T complexity
PPC	posterior predictive check
MRS	Markovian-Regime-Switching model

Dedicated to my mum, Kylie.

Chapter 1

Introduction

1.1 Introductory background

Electricity is a unique commodity as it cannot currently be stored efficiently and requires immediate delivery to consumers. This, coupled with the facts that electricity demand is inelastic, highly variable, and dependent on weather conditions and business activities, causes electricity spot prices to exhibit extreme behaviour in deregulated markets [35]. The electricity spot price is the price that commercial generators, large consumers and electricity retailers buy and sell electricity at, and not the price faced by the general public. Large price spikes, periods of high volatility and regulations that restrict electricity retailers passing risk on to consumers mean that market participants face significant risk. To hedge this risk, derivative contracts are used, and to price derivatives a model of spot prices is needed. Since electricity markets display spike characteristics not commonly found in other markets, models developed for other markets do not adequately capture the price dynamics of electricity, and modelling them is an active area of research.

The issues mentioned above are particularly relevant in South Australia, where relative isolation and generation mix lead to high and volatile electricity spot prices; not to mention issues regarding power system stability and blackouts. The power system in South Australia, and more generally in Australia, has also become a popular issue for politicians. The Australian Government is currently debating elements of the National Energy Guarantee, a plan which aims to lower electricity prices and increase reliability to stimulate economic growth, among other things.

A commonly used model for electricity prices is the Markovian-regime-switching (MRS) model whereby multiple stochastic processes are interweaved by a Markov chain. The general idea is that there exist multiple regimes underlying the price process, and depending on which regime the system is in, different characteristics are displayed. For

example, for electricity prices we could suppose that there is a *normal* or *base regime* where prices are relatively low and comparatively non-volatile, and a *spike regime* where prices are high and volatile. For MRS models, we assume that the regime the system is in follows a discrete-time Markov chain which is not directly observable. MRS models can be seen as extensions of hidden Markov models, since MRS models also allow for dependence between observations, given the hidden regime process. In electricity price modelling applications, this dependence is typically specified as an autoregressive process of order 1 which relates random variables through the equation $X_t = \alpha + \phi X_{t-1} + \sigma\varepsilon_t$, where α, ϕ and σ are parameters, and $\{\varepsilon_t\}$ is a sequence of independent, identically distributed $\mathrm{N}(0, 1)$ random variables. More broadly, MRS models find application in biology [2], weather modelling [72, 99], speech recognition [70] and more, and we hope our contributions here can extend to these fields also.

In the electricity price modelling literature, it has become popular to specify MRS models with *independent* regimes. To define an MRS model, for any time t, let us denote the (hidden) regime of the system as R_t, and the observed price as X_t. Independent regime MRS models are MRS models where, given $R_t = i$, X_t depends on previous prices from Regime i only. The advantage of MRS models with independent regimes is that there is no transitional behaviour after a change in regime. When applied to electricity price modelling, this means these models can capture the rapid return to base levels after a price spike. However, likelihood evaluation for MRS models with independent regimes is complicated by this dependence structure – the dependence between prices is governed by the hidden regime process and therefore is random. Parameter estimation for MRS models with independent regimes is still underdeveloped, and it is the goal of this thesis to address this.

The current method of inference for MRS models with independent regimes is an approximation to the EM algorithm, which we show can be unreliable. We then develop and implement two solutions to this inference problem. We first develop a novel, computationally feasible, and exact likelihood-based framework, and then a data-augmented Bayesian framework.

1.2 The South Australian electricity market

The South Australian (SA) electricity market is a particularly interesting case study due to its relative isolation, extreme weather and high penetration of renewables – almost 50% of the generation in SA in 2016 was from renewables [7] – all of which can contribute to high and volatile prices [7]. The SA market is part of a broader network of connected markets called the National Electricity Market (NEM) which was established

in December 1998. The NEM is comprised of five interconnected states of Australia which also act as price regions: Queensland, New South Wales (including the Australian Capital Territory), South Australia, Victoria, and Tasmania. Each state has its own generation capacity and can also import/export electricity and ancillary services via interconnectors between states. One aspect of the SA market, and more generally the NEM, that makes Australia's energy network interesting is the relative sparsity of the network. The NEM stretches from Port Lincoln in the west of SA, across Bass Straight to Hobart in the south of Tasmania, and up to Port Douglas in far north Queensland. Compared to other electricity grids around the world the NEM is relatively sparse since it services only 9 million customers [4]. Nonetheless, the NEM is crucial to Australia's economy supplying about 2000 terawatt hours of electricity to consumers each year. There are currently over 300 registered participants (large generators, energy retailers and large consumers) in the NEM who traded $16.6 billion through the NEM in the financial year 2016-2017.

The majority of electricity generation in the NEM is from coal, accounting for 77% of annual generation in the financial year 2016-2017; gas accounted for 9% of total generation, hydro power 8%, followed by wind, 5%, and a small amount of solar, 0.3% (not including behind-the-meter residential rooftop solar) and 0.7% other sources [5]. Small scale behind-the-meter solar accounts for approximately 2.5% of total electricity generation in the whole NEM, but is not traded in the wholesale market. South Australia is in contrast to this: there is a larger contribution from wind and no contribution from coal. In SA, gas produced 50.5%, wind 39.2%, rooftop solar 9.2% and diesel and other non-scheduled generators 1.1% of the 11,077 GWh of electricity produced in SA in 2016-2017 [7]. Note that these numbers are only for energy production in SA and not consumption. In 2016-2017 SA had significant net energy imports – 164 GWh was exported while 2,869 GWh was imported from other states where there is a high percentage of coal generation. There was no coal generation for this period since SA's last coal generator was decommissioned in May 2016.

The NEM is managed by the Australian Energy Market Operator (AEMO). AEMO has many roles including organising and dispatching energy markets (which are the markets that we are interested in), maintaining system stability, operating ancillary services markets, roles in long-term forecasting and system planning, and also operating gas markets in Australia. The five energy markets in the NEM (SA, Victoria, Queensland, Tasmania and New South Wales) are dispatched every five minutes of every day. In this process, scheduled generators submit initial bids to AEMO at a 30-minute resolution, stating the amount of electricity they are willing to supply and at what price, as a stack of ten quantity-price pairs. Generators may bid prices anywhere between the market floor, $-\$1,000$ per MWh, and the market cap, $\$14,000$ per MWh. Some large consumers

also submit demand bids. *Initial bids* must be submitted before 12:30pm on the day before dispatch (the trading day is defined as 4:00am one day to 3:59am of the next day), but rebids are allowed up to 5 minutes before dispatch. Generators are allowed to rebid the amounts of energy they are willing to supply across the ten prices only, and are not allowed to change the bid prices. Companies participating in this market are informed of the *pre-dispatch* prices, which they use along with other updated information such as weather forecasts, to inform their rebids for the 5-minute intervals. If a rebid is made within a 30-minute interval, or less than 15-minutes before dispatch, it must be accompanied by an explanation which can be reviewed by the Australian Energy Regulator (AER).

AEMO also forecasts energy production from non-scheduled generators (such as wind and solar) and consumer demand for each 5-minute interval. The system is dispatched by AEMO who match supply with forecast demand using a large optimisation program, with an objective to minimise costs subject to demand and system constraints. The price set by the optimisation program is the *dispatch price* which is the lowest bid price that causes generators to fulfil demand, and all generators are paid this price. The *spot price* is the average of 6 dispatch prices in a half hour interval resulting in 48 realisations of the spot price per day. The spot price is the price that we are interested in modelling, as it is the price at which transactions actually take place and on which contracts are valued.

Due to the nature of the bid and dispatch processes, as well as other factors such as the number and composition of generators in the market, the NEM is vulnerable to strategic bidding of generators [31]. For example, Hurn *et al.* [54] suggest that a base-load generator could influence the price by rebidding generation from lower prices to higher prices forcing the dispatch price upwards for some 5-minute interval. The inflated dispatch price affects the spot price for the entire 30-minute interval, and the generator can then offer a large portion of its generation at lower prices for the other intervals in the trading period, ensuring that it is dispatched and knowing it will receive an inflated price.

Of course price spikes occur naturally in this market also. A series of large price spikes in the SA market occurred in winter, 2016, and these were largely attributed to poor wind generation forecasts – AEMO estimated there was going to be much more wind energy available than there was, leaving the market unprepared and peaking generators had to be turned on to meet demand, causing the price to spike. Other natural causes of spikes are generator breakdowns or transmission failures, which unexpectedly remove supply from the market.

Significant price drops also occur in this market, hence the price floor of $-\$1,000$. For example, price drops can occur when there is an unexpected glut of wind generation, which has a low marginal cost and is given priority over other generators in SA. Negative prices occur when there is supply surplus. It can be impossible for some generators to turn off at short notice, and it can take hours for them to restart, so it can occasionally be cheaper for generators to momentarily pay to produce energy than to shut down.

Electricity prices, in general, exhibit trends which can largely be explained by demand. For example, it is known that demands, and therefore prices, are higher on business days than on non-business days on average, and that prices are often lower overnight when fewer people are awake, and highest in the afternoon and evening. In South Australia this is no different. However, there is one emerging trend in SA that may become significant in the future: AEMO predict that by the year 2025-2026 the minimum demand for electricity will become negative at times due to an oversupply of solar generation [7].

The strength of the SA electricity grid has come into question over the past few years due to recurrent blackouts during extreme weather, including a 'system black' event which occurred at 4:20pm on September 28, 2016, when all electricity supply to the state was lost. The first customers had power restored by 7pm the same day, while the whole system was not fully restored until the 11^{th} of October. The system black event was caused when strong winds severed two critical transmission lines resulting in six voltage drops within two minutes. This was followed by nine wind farms reducing power production as a precautionary control – generation from one wind farm was lost completely, while eight others continued to produce at a reduced level. This totalled a generation reduction of 456 MW in less than seven seconds. To compensate, energy imports across the Haywood inteconnector were increased until they reached a level that tripped the interconnector and SA was left separated from the rest of the NEM. At this point, generators in SA were unable to maintain the frequency of the grid and supply to the entire state was lost. Modelling performed by AEMO suggests that the system black event could have been avoided had the wind generators continued to operate normally and not reduced capacity.

The system black event caused AEMO to suspend market operation from the trading period beginning at 4:00pm, on the 28^{th} of September until 10:30pm on the 11^{th} of October. During this period, prices were set by AEMO and not the market. The prices for this period were calculated as the average price in SA in the 'same trading interval' over the last four weeks. For this calculation the 'same trading interval' means different things for weekdays and weekends. For a given 30-minute trading interval on a weekday, the price was calculated as the average price at the same time only on weekdays over the last four weeks. For a given 30-minute trading interval on a weekend, the price

was calculated as the average price at the same time on weekend-days only. During the market suspension, market participants were instructed to continue to submit price bids in the usual way, and AEMO used these to dispatch generators in some sort of economic merit order.

SA's vulnerability is caused by many factors including its isolation, weather, sparsity of customers and reliance on asynchronous generators. Sparsity can leave a system vulnerable since it can be uneconomic to build precautionary backups in parts of the network that do not supply many customers. Synchronous generators provide frequency regulation which is necessary for stable operation – most of SA's synchronous generators are located in the Adelaide region, while the next closest synchronous generators that are likely to be active are in the Latrobe Valley, 800km away in Victoria. In addition, SA only has one alternating current interconnector to the rest of the NEM through which frequency control can be received. Hence, as a result of being relatively weakly connected to the rest of the NEM and having synchronous generators localised to the Adelaide region, SA only has a marginal benefit from the strength in numbers of synchronous generators [7]. However, new technologies such as battery storage and control systems for wind generators are helping to manage this. Lastly, SA's weather can cause issues in a different way. Hot weather in SA's summer can overheat infrastructure and also increases demand (mainly from people using air conditioning) which can cause blackouts and load shedding, where parts of the network are cut off to maintain system stability. To help maintain system stability, SA has recently installed the world's largest lithium ion battery farm connected to the Hornsdale wind farm – the battery has equivalent generation capacity of 100 megawatts and storage capacity of 125 megawatt hours.

These are exciting times in the energy industry as we debate economic, environmental and system reliability trade-offs and explore innovative solutions to these problems. In SA there are plans for another battery farm and investment in solar thermal storage solutions. There is debate about incentivising private investment in small-scale batteries and solar panels, and control systems to manage distributed storage solutions. There are more wind generators planned, debate about another coal-fired plant and plans for a solar thermal storage solution. And, of direct relevance to this thesis, there are plans to change from a 30-minute pricing scheme, to a 5-minute pricing scheme in 2021 [6].

1.3 Structure of the thesis

Chapter 2 consists of three main sections. In Section 2.1 we provide definitions of different classes of MRS models; Type I, a *dependent regime* model, and Types II and III, which are two different specifications of *independent regime models*. There are slight

subtleties in these definitions, and we hope to make them clear in Section 2.1. Section 2.2 gives a lengthy overview of technical concepts related to this thesis. While we do not use *every* concept from Section 2.2 directly, we believe that a thorough understanding of these concepts will assist the reader in fully appreciating our work. Section 2.3 gives a broad overview of different types of electricity price models, before thoroughly exploring the development of MRS models for electricity prices. Section 2.3 also provides a literature review of some current methods of inference for MRS models. First, the *forward* algorithm of Hamilton [45], which is used to evaluate the likelihood of MRS models of Type I, is presented. Then, the *backward* algorithm of Kim [67], together with more work of Hamilton [45] to implement the EM algorithm for Type I models, is presented. We also briefly mention the approximation of the EM algorithm by Janczura and Weron [60], which we title the 'EM-like' algorithm and, until this thesis, was the current method of choice for inference of MRS models with independent regimes. We save a more thorough discussion of the EM-like algorithm for Chapter 3. Finally, we conclude this chapter with a literature review of detrending techniques for electricity prices.

In Chapter 3, we first discuss the EM-like algorithm of Janczura and Weron [60], then present some theoretical issues with the algorithm and examples where the EM-like methodology can fail. In Section 3.2, we develop a novel forward algorithm to evaluate the likelihood of MRS models with independent regimes. Then, in Section 3.3, we develop a backward algorithm which gives a computationally feasible way to calculate *smoothed inferences* for MRS with independent regimes. Using our backward algorithm we derive an Expectation-Maximisation (EM) algorithm to find the maximum likelihood estimates of independent regime MRS models. The construction of our algorithms is similar to the construction of analogous algorithms for hidden Markov models, traditional (dependent-regime) MRS models, and hidden semi-Markov models. The general idea is to augment the hidden regime process with a 'last-visit' counter which keeps track of the last time the system was in a given regime. Then, for each time t, the augmented hidden regime process contains all the information needed to specify the distribution of the price X_t, given all past prices. Furthermore, the hidden process remains Markovian and so techniques from the existing literature can be applied. We conclude this section with a discussion where we compare our EM algorithm to 'black-box' optimisation procedures using our forward algorithm, empirically investigate bias and consistency of the MLE, discuss some difficulties relating to *shifted* distributions which are commonly used in electricity price modelling, and conclude by mentioning some potential extensions of our methods and future work.

Chapter 4 presents a Bayesian approach to parameter inference using a data-augmented Markov chain Monte Carlo algorithm. Bayesian inference is a paradigm where model

parameters are thought to be random variables, and the goal is to infer an entire distribution of these parameters, called the *posterior distribution*. Data-augmented MCMC is a powerful technique which enables efficient sampling of parameters from the posterior distribution by extending the sample space to provide further information. We take time to detail our MCMC implementation in Section 4.2, highlight the motivation for each element of the algorithm, and describe its intricacies. One element of our MCMC implementation that we particularly enjoyed was implementing *adaptive steps*, to automatically tune our algorithm, making our algorithm much more practical. In the application of our Bayesian methodology we rely heavily on *posterior predictive checks* (PPCs) to assess model fit. Our implementation of the PPC methodology is qualitative, and similar to traditional diagnostic plots for simple regression models. For this reason we investigate the power of our methods to distinguish between different models in Section 4.4.

Our last major chapter is an application to the South Australian electricity market. We first detail our trend estimation technique. Extreme observations such as spikes in electricity prices can bias our estimate of trend components. For this reason we implement an iterative filtering method, where we iterate between estimating the trend components of the model, classifying prices as extreme and removing them from the data. In Section 5.2, we present the candidate models which we then fit to the South Australian electricity price dataset. In Sections 5.3 and 5.4 we apply our Bayesian and maximum likelihood methodologies respectively. We also present some of our analysis of the South Australian dataset in Appendix B to avoid repetition. Of note, we find that there is a significant jump in volatility around April 2016, which corresponds to the closure of South Australia's only coal generation facility. We also find no need for a regime to capture significant price drops for the South Australian market. In Section 5.4 we resort to 'common-sense' model checking since we cannot use the Akaike Information Criterion, or related information-theoretic model comparisons, due to the entwinement of our deterministic trend estimation methods with our estimation of the parameters of our MRS models. We conclude this section with a discussion of our analyses, and comment on some lessons learned.

Finally, in Chapter 6 we conclude the thesis, commenting on our significant contributions, possible future work, and the (many) lessons learned from this work.

Chapter 2

Background

2.1 Introduction to MRS models

An MRS model is built from two pieces, an unobservable regime sequence and an observable sequence. As the name suggests, MRS models assume the unobservable regime sequence is a Markov chain. Let $\{R_t\}_{t\in\mathbb{N}}$ be a Markov chain on a finite state space $\mathcal{S} = \{1, 2, ..., M\}$, with transition matrix P, and $\{X_t\}_{t\in\mathbb{N}}$ be the observation sequence, in this case electricity prices. Then $R_t = i$ represents the event that the process X_t is in Regime i at time $t \geq 0$.

In this thesis we consider MRS models with regimes that either have independent and identically distributed (i.i.d.) prices, or *autoregressive of order 1* (AR(1)) prices. Typically, we define i.i.d. distributions with two or three parameters, a location parameter μ_j, and a scale (or variance) parameter σ_j^2, and a shifting parameter q_i may also be included, where j is the index denoting the regime. We parameterise AR(1) regimes with a location parameter α_j, correlation ϕ_j and conditional variance σ_j. That is, suppose $\{Y_t\}_{t\in\mathbb{N}}$ is an AR(1) process for Regime j defined by

$$Y_t = \alpha_j + \phi_j Y_{t-1} + \sigma_j \varepsilon_t,$$

where $\{\varepsilon_t\}_{t\in\mathbb{N}}$ is a sequence of i.i.d. N(0,1) random variables. The parameter ϕ_j is the correlation between Y_t and Y_{t-1}; moreover, the correlation between Y_t and Y_{t-m} is ϕ_j^m. The parameter α_j shifts the long-run mean level of the AR(1) process to $\dfrac{\alpha_j}{1-\phi_j}$. The parameter σ_j^2 is the conditional variance of Y_t given Y_{t-1}, i.e. $var(Y_t|Y_{t-1}) = \sigma_j^2$.

The simplest MRS model is the hidden Markov model (HMM), where observations X_t take values in a discrete set, and X_t is independent of $X_{t-1}, \ldots, X_0$ and $X_{t+1}, X_{t+2}, \ldots$

given the regime at time t, R_t. In general, MRS models are specified in terms of distributions that allow dependence on past observations, given the current regime. That is, the model defines distributions,

$$X_t|R_t, X_{t-1}, X_{t-2}, ..., X_0 \sim F^{R_t},$$

for some distribution F^{R_t}. Dependent-regime MRS models were first developed by Hamilton [44] in the context of modelling financial markets. Hamilton defined MRS models where each regime followed autoregressive dynamics and developed estimation techniques for these models. For example, a simple MRS model that fits into the class introduced by Hamilton is the following. Let $X_t = \alpha_{R_t} + \phi X_{t-1} + \sigma\varepsilon_t$, where ε_t is a sequence of i.i.d. N(0,1) random variables, so that

$$X_t|R_t, X_{t-1}, X_{t-2}, ..., X_0 \sim N(\alpha_{R_t} + \phi X_{t-1}, \sigma^2).$$

Here we have assumed that X_t follows AR(1) dynamics, and that the constant term α_{R_t} is the only term dependent on the hidden regime. More generally, the traditional MRS model specifies that

$$X_t|R_t, X_{t-1}, X_{t-2}, ..., X_0$$

follows some time-series model with finite dependence on past observations for each $R_t \in \mathcal{S}$, and this dependence structure does not depend on $R_0, .., R_{t-1}$. That is, the dependence structure does not take into account which regime the past observations

$$X_{t-1}, X_{t-2}, ..., X_0$$

belong to. We label these dependent-regime models as *MRS models of Type I.*

Example 2.1 (An MRS model with dependent regimes (Type I))**.** *Let* $\mathcal{S} = \{1, 2\}$*, and*

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

and specify

$$X_t|R_t = 1, X_{t-1}, X_{t-2}, ..., X_0 \sim N(0.6X_{t-1}, 1),$$
$$X_t|R_t = 2, X_{t-1}, X_{t-2}, ..., X_0 \sim N(1 + 0.9X_{t-1}, 1).$$

So X_t follows AR(1) dynamics in both Regime 1 and Regime 2. This is an MRS model of Type I since X_t depends on X_{t-1} regardless of which regime the lagged observation, X_{t-1}, came from. In Figure 2.1 we plot a simulation of this model and colour the observations from each regime to illustrate the characteristics of these models.

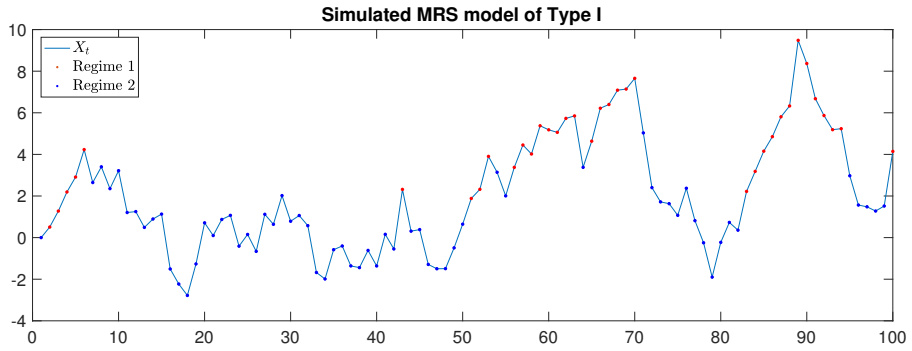


FIGURE 2.1: A simulation of the Type I MRS model in Example 2.1 where we have coloured the observations from each regime. Red points are from Regime 1 and blue points are from Regime 2.

In the electricity price modelling literature it has become popular to specify MRS models with *independent* regimes. These are models where, given $R_t = i$, X_t depends only on lagged values from Regime i. We further classify these models in two groups depending on what happens within each regime between times when they are observed. If the processes within the regimes evolve regardless of whether they are observed or not, we call them *MRS models of Type II* and describe them as *MRS models with independent regimes that evolve at all time points*. An example of this type of model is in Example 2.2.

Example 2.2 (An MRS model of Type II)**.** *Let* $\mathcal{S} = \{1, 2\}$*, and*

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

and define the following AR(1) processes

$$\begin{aligned} B_t &= 0.6B_{t-1} + \varepsilon_t^B, \\ S_t &= 1 + 0.9S_{t-1} + \varepsilon_t^S, \end{aligned}$$

where ε_t^B *and* ε_t^S *are i.i.d. sequences of N(0,1) random variables. Then, construct the MRS model* X_t *as follows*

$$X_t = \begin{cases} B_t, & \text{if } R_t = 1, \\ S_t, & \text{if } R_t = 2. \end{cases}$$

A simulation of this process is plotted in Figure 2.2.

The advantage of MRS models with independent regimes is that the behaviour of each regime is distinct, and we do not need to have any transitional behaviour after a change in regime. When applied to electricity price modelling, this means the model can capture

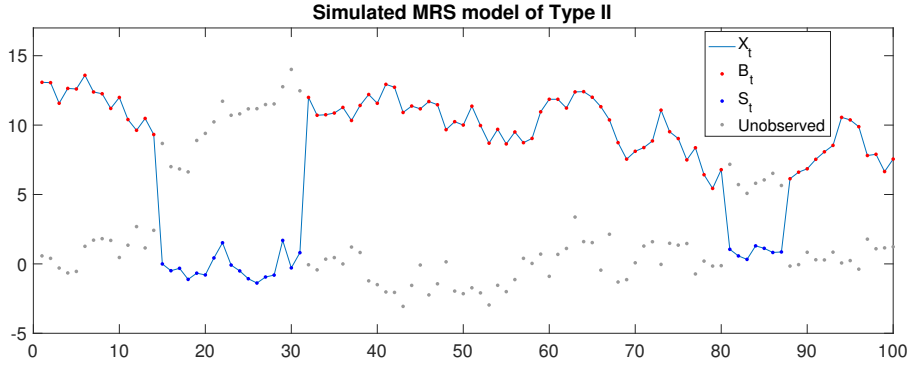


FIGURE 2.2: A simulation of the Type II MRS model in Example 2.2 where observations are coloured according to which regime generated them. The blue points are observed values of the process B_t, the red points are observed values of the process S_t, and the unobserved values within each processes are represented by the grey dots. Notice that there are no 'transition' periods after a change of regime, rather there is a distinct jump in the process at transition times.

a rapid return to base levels after a price spike occurs, which is a phenomenon that is commonly observed [3].

We also introduce a new MRS model to the electricity pricing literature, which we label an *MRS model of Type III* and describe these models as *MRS models with independent regimes which evolve only when observed*. This model is similar to the MRS model of Type II except the processes within each regime stop between times when they are observed, which slightly simplifies the analysis since there are no unobserved values. An example of an MRS model of Type III is illustrated in the following.

Example 2.3. *Again, let* $\mathcal{S} = \{1, 2\}$*, and*

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

and define the following AR(1) processes

$$B_{\tau_B(t)} = 0.6B_{\tau_B(t-1)} + \varepsilon^B_{\tau_B(t)},$$
$$S_{\tau_S(t)} = 1 + 0.9S_{\tau_S(t-1)} + \varepsilon^S_{\tau_S(t)},$$

where $\varepsilon^B_{\tau_B(t)}$ *and* $\varepsilon^S_{\tau_S(t)}$ *are i.i.d. sequences of N(0,1) random variables,* $\tau_B(t) = \sum_{i=0}^{t} \mathbb{I}(R_i = 1)$ *and* $\tau_S(t) = \sum_{i=0}^{t} \mathbb{I}(R_i = 2)$*. Then, construct the MRS model* $\{X_t\}$ *as follows*

$$X_t = \begin{cases} B_{\tau_B(t)}, & \text{if } R_t = 1, \\ S_{\tau_S(t)}, & \text{if } R_t = 2. \end{cases}$$

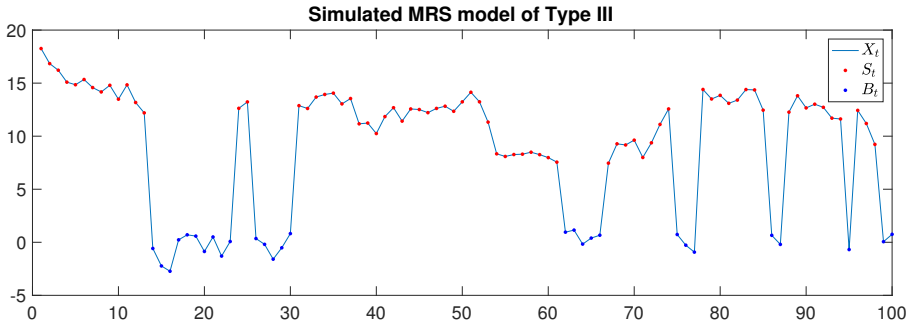


FIGURE 2.3: A simulation of the MRS model of Type III from Example 2.3 where observations are coloured according to which regime generated them. Red points are from the process B_t and blue points are from the process S_t. Similar to the other independent regime MRS model specification, there are no 'transition' periods after a change of regime since regimes are independent. Notice that there are now no unobserved values of within-regime processes, and hence correlations are equally strong regardless of the gap in that regime.

A realisation of this process is plotted in Figure 2.3.

To make precise what we mean by dependent and independent regime models, define the sets $\mathcal{A}_i := \{t \in \mathbb{N} : R_t = i\}$, $i \in \mathcal{S}$. We say that a model has *independent regimes* if the sets $\{X_t : t \in \mathcal{A}_i\}$, $i \in \mathcal{S}$, are independent (that is, if $A = \{X_t : t \in \mathcal{A}_i\}$ and $B = \{X_t : t \in \mathcal{A}_j\}$, for $i \neq j$, then A and B are independent) events. Otherwise, it is a *dependent regime* model

2.2 Technical concepts

In this section we review technical concepts needed for this thesis.

2.2.1 Markov chains

Here, we focus only on discrete-time Markov chains on a finite *state space*, for clarity of exposition. Later in this thesis we also deal with discrete-time Markov chains on general spaces, for example AR(1) processes are Markov chains on $\mathbb{R}$, as is our implementation of Markov chain Monte Carlo algorithms. However, since the concepts behind general-state-space Markov chains are more technical, we detail them in Appendix A rather than here.

A discrete-time Markov chain is a sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ that have the *Markov property*. We denote by $\mathcal{S}$ the *state space*, which is the set of possible values X_t can take. The Markov property says that the probability of moving into a state $i \in \mathcal{S}$

at time $t+1$, given the entire history of the process $X_t = i_t, X_{t-1} = i_{t-1}, ..., X_0 = i_0$ for $i_0, ..., i_t \in \mathcal{S}$, depends only on the current position of the chain, $X_t = i_t$;

$$\mathbb{P}\left(X_{t+1} = i | X_t = i_t, X_{t-1} = i_{t-1}, ..., X_0 = i_0\right) = \mathbb{P}\left(X_{t+1} = i | X_t = i_t\right),$$

for all $i_0, ..., i_t, i \in \mathcal{S}$, and all $t \in \mathbb{N}$, assuming both conditional probabilities are well defined, i.e.

$$\mathbb{P}(X_t = i_t, X_{t-1} = i_{t-1}, ..., X_0 = i_0) > 0.$$

A Markov chain is called *time homogeneous* if $\mathbb{P}(X_{t+1} = i | X_t = j) = \mathbb{P}(X_1 = i | X_0 = j)$ for all $t \in \mathbb{N}$. In this thesis, we assume that Markov chains are time homogeneous unless otherwise stated.

Without loss of generality, the state space $\mathcal{S}$ is assumed to be $\{1, 2, ..., M\}$, where $M < \infty$. The probabilities $p_{ij} := \mathbb{P}\left(X_{t+1} = j | X_t = i\right)$, $i, j \in \mathcal{S}$, known as *transition probabilities*, are represented collectively as the *transition matrix*

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \\ p_{M1} & p_{M2} & \cdots & p_{MM} \end{bmatrix}.$$

The *n-step transition probabilities* are defined as $\mathbb{P}(X_{t+n} = j | X_t = i)$ and represented collectively as $P^{(n)}$. We can show that the n-step transition probabilities are given by

$$P^{(n)} = P^n.$$

A Markov chain is said to be *irreducible* if the process can get from any state to any other state with positive probability; that is, if $\mathbb{P}(X_n = j | X_0 = i) > 0$ for some $n \in \mathbb{N}$ and each $i, j \in \mathcal{S}$.

2.2.2 Maximum likelihood

Maximum likelihood estimation is a popular technique to estimate the parameters of a probabilistic model from observed data. The estimates produced by this method are called *maximum likelihood estimates* (MLEs) and have numerous nice properties [21, 78]. Maximum likelihood supposes that observed data, $\boldsymbol{x}_{0:t} = (x_0, ..., x_t)$, were generated from some distribution $f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t})$ with unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ that belongs to a parameterised family of distributions, $\{f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}) | \boldsymbol{\theta} \in \Theta\}$, where Θ is the parameter space. The *likelihood function* is defined as $L(\boldsymbol{\theta}) := f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t})$, that is, the density function $f^{\boldsymbol{\theta}}$ is evaluated at the observed data and treated as a function of

the parameters $\boldsymbol{\theta}$. Maximum likelihood estimation finds the parameters that maximise the likelihood, which are commonly denoted by $\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$. In practice, it is often easier to work with the log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{x}_{0:t}) = \log f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t})$; note that $\widehat{\boldsymbol{\theta}}$ maximises the likelihood if and only if it also maximises the log-likelihood.

Desirable properties of the maximum likelihood estimators are:

- Consistency ([78], Section 2): Under certain regularity conditions, if the data are generated by $f^{\boldsymbol{\theta}^*}$, where $\boldsymbol{\theta}^*$ are the true parameters, then the MLEs *converge in probability* to the true parameters as the number of observations grows. Under slightly stronger conditions, the MLEs *converge almost surely* to the true parameters as the number of observations grows.

- Functional invariance ([21], Section 7.2): If $h(\boldsymbol{\theta})$ is some transformation of the parameter vector, $\boldsymbol{\theta}$, then the MLE for $h(\boldsymbol{\theta})$ is $\widehat{h}(\boldsymbol{\theta}) = h(\widehat{\boldsymbol{\theta}})$.

- Asymptotically normally distributed ([78], Section 3): Define the *Fisher information* as

$$I(\boldsymbol{\theta}^*) := \mathbb{E}_{f^{\boldsymbol{\theta}^*}}\left[\frac{d^2}{d\boldsymbol{\theta}^2}\log f^{\boldsymbol{\theta}}(\boldsymbol{X})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right], \tag{2.1}$$

  where $\boldsymbol{X}$ are random variables with distribution $f^{\boldsymbol{\theta}^*}$, and $\dfrac{d}{d\boldsymbol{\theta}}$ is the vector derivative. Then, under certain regularity conditions,

$$\sqrt{t}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \to N_p\left(0, I(\boldsymbol{\theta}^*)^{-1}\right) \text{ in distribution,}$$

  where N_p is the p-dimensional normal distribution. This means that, for large sample sizes, we can expect that the MLEs approximately follow a Normal distribution with known variance, which has practical applications in model selection and hypothesis testing.

- Asymptotic efficiency ([78], Section 5): Under certain regularity conditions then

$$var\left(\widehat{\boldsymbol{\theta}}\right) \to I(\boldsymbol{\theta}^*)^{-1}.$$

  This means the MLE is an estimator that asymptotically has the smallest variance of all unbiased estimators. That is, the MLE satisfies the *Cramer-Rao lower bound* with equality as the sample size goes to infinity.

2.2.3 The EM algorithm

General Expectation Maximisation theory was initially developed by Dempster *et al.* [29] in the late 70s. However, it turned out that work on maximum likelihood estimation of HMMs by Baum in the late 1960s [9–11] was a specific application of the EM algorithm. The EM algorithm is a maximisation technique typically used to find the MLE for problems involving missing or latent data. In missing data problems, the likelihood is typically hard to evaluate and therefore the MLE is difficult to find directly. The EM algorithm provides a way to find the MLE numerically, often without ever evaluating the likelihood itself.

Suppose data, $\boldsymbol{x}_{0:t} = (x_0, ..., x_t)$, is observed from some model with density $f^{\boldsymbol{\theta}^*}$ and suppose this density can only be written as a marginal density,

$$f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}) = \int f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}, \boldsymbol{Y}) d\boldsymbol{Y}, \tag{2.2}$$

where $\boldsymbol{Y}$ is the missing, or latent, data in the problem, and the integral is over the support of $\boldsymbol{Y}$. We call $f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}, \boldsymbol{Y})$ the *complete data likelihood* and $f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t})$ the *incomplete data likelihood*. The EM algorithm circumvents having to work with the integral in Equation (2.2). The EM algorithm iterates between 'expectation' (E) and 'maximisation' (M) steps to produce a sequence $\{\boldsymbol{\theta}_n\}_{n\in\mathbb{N}}$ that converges to the maximiser of the likelihood, $\boldsymbol{\theta}^*$ (under certain regularity conditions). The EM algorithm can also be extended to find the maximum *a posteriori* estimate (MAPE) in a Bayesian setting.

At the n^{th} iteration of the EM algorithm, suppose the current parameters are $\boldsymbol{\theta}_n$. In the E-step of the algorithm, the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ is constructed as follows

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) := \mathbb{E}_{f^{\boldsymbol{\theta}_n}(\boldsymbol{Y}|\boldsymbol{x}_{0:t})} \left[\log f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}, \boldsymbol{Y}) | \boldsymbol{x}_{0:t} \right].$$

This expectation can be much easier to calculate than Equation (2.2), which is usually the case for exponential families.

In the M-step of the algorithm the maximisers $\boldsymbol{\theta}_{n+1}$ are found,

$$\boldsymbol{\theta}_{n+1} := \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n),$$

which are then used at the $(n+1)^{\text{st}}$ iteration to construct $Q(\cdot, \boldsymbol{\theta}_{n+1})$.

The EM algorithm is a subclass of MM (majorisation-minimisation or minorisation-maximisation) algorithms, and works because the functions $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ *minorise* the likelihood. That it, Q has the property

$$\log f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t}) - \log f^{\boldsymbol{\theta}_n}(\boldsymbol{x}_{0:t}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) - Q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n).$$

So increasing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ past $Q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n)$ must cause $\log f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:t})$ to increase past $\log f^{\boldsymbol{\theta}_n}(\boldsymbol{x}_{0:t})$. Thus, if $Q(\boldsymbol{\theta}_{n+1}, \boldsymbol{\theta}_n) > Q(\boldsymbol{\theta}_n, \boldsymbol{\theta}_n)$ then $f^{\boldsymbol{\theta}_{n+1}}(\boldsymbol{x}_{0:t}) > f^{\boldsymbol{\theta}_n}(\boldsymbol{x}_{0:t})$, and so the sequence $\{\boldsymbol{\theta}_n\}_{n\in\mathbb{N}}$ must increase the log-likelihood.

Unlike other optimisation methods, such as *steepest decent*, the EM algorithm does not rely on evaluating or approximating derivatives of the log-likelihood, and this can be advantageous. However, in general, there is no guarantee that the EM algorithm will converge to the true maximiser, in particular, it is known that when the likelihood is multimodal, the EM algorithm can get stuck at local maximisers or saddle points [105]. So, in practice, the algorithm should be initialised at a range of initial values to increase the chances of finding the true MLE.

The EM algorithm is a first-order algorithm, in that the convergence of the sequence $\{\boldsymbol{\theta}_n\}_{n\in\mathbb{N}}$ to the true maximiser is linear. Consider the mapping M defined by $\boldsymbol{\theta}_{n+1} := M(\boldsymbol{\theta}_n)$ given by the EM algorithm, with a fixed point $\boldsymbol{\theta}^*$, then

$$\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}^* = \frac{d}{d\boldsymbol{\theta}} M(\boldsymbol{\theta}^*)\,(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*) + \mathcal{O}(||\boldsymbol{\theta}_n - \boldsymbol{\theta}^*||^2),$$

where $\frac{d}{d\boldsymbol{\theta}} M(\boldsymbol{\theta}^*)$ is the Jacobian matrix, and $||\cdot||$ is the Euclidean distance [73]. In practice the EM algorithm is terminated after some finite number of iterations, when the step size becomes small, for example, when $\max |\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n| < \varepsilon$, and/or $|f^{\boldsymbol{\theta}_{n+1}}(\boldsymbol{x}) - f^{\boldsymbol{\theta}_n}(\boldsymbol{x})| < \varepsilon$. This is a necessary but not sufficient condition for the EM algorithm to have found the true MLE. As such, using this termination criterion the EM algorithm may terminate in places where the likelihood is relatively flat, but not a local maximum.

2.2.4 Bayesian inference

Bayesian inference is a separate parameter estimation paradigm to maximum likelihood inference. In a Bayesian setting, the parameters $\boldsymbol{\theta}$ are treated as unknown random variables, which is different from the maximum likelihood setting where the parameters are treated as unknown constants. The goal of Bayesian inference is to infer the distribution of the unknown parameters, given observed data; this is known as the *posterior distribution*.

Prior distributions The first step in a Bayesian inference problem is to define the *prior distribution*, which is the distribution the observer thinks the parameters follow *before* any observations are made. It is not always clear how to choose a prior distribution, and this is one of the points of contention for some statisticians about Bayesian methods. However, in a sense it is not much different to choosing a model for the data in the first place. There have been many attempts to derive *uninformative* or *objective* prior distributions, but these depend on the definition of *uninformative* or *objective*. A simple objective prior distribution is the uniform distribution,

$$\pi(\boldsymbol{\theta}) \propto c,$$

for $\boldsymbol{\theta} \in \Theta$ with Θ bounded, and c a constant. In general we use the notation $\pi(\cdot)$ for prior distributions. One interpretation of the uniform prior distribution is that no information about the parameters is known before the data is observed. In many circumstances, it is still valid to specify a uniform prior when Θ is unbounded and arrive at a posterior distribution that is well-defined, but care needs to be taken as $\pi(\cdot)$ is technically no longer a distribution since $\int_\Theta \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \infty$. A drawback of uniform objective prior distributions is that they are not invariant to transformations. For example, specifying $\pi(\theta) \propto c$ for some scalar parameter θ, is not equivalent to specifying $\pi(\theta^2) \propto c$. Another attempt at an objective prior is *Jeffreys' prior* [63], which is defined as

$$\pi(\boldsymbol{\theta}) \propto \sqrt{det(I(\boldsymbol{\theta}))}.$$

Jeffreys' prior is invariant to transformation, but comes with problems of its own, such as it is sometimes not a well-defined distribution since $\int_\Theta \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \infty$.

Another important type of prior distributions is the *conjugate prior distribution*, which is a distribution such that the posterior and prior distributions belong to the same family. Conjugate prior distributions are particularly 'nice' since the posterior distribution is available in closed form and very little computations are needed. Developed in the age before computers, conjugate prior distributions are practical because heavy computations are not needed, but with modern computing power this is no longer a restriction.

The likelihood Bayesian inference also relies on the likelihood. In Bayesian inference the likelihood is treated slightly differently than in maximum likelihood inference: it is interpreted as the *conditional distribution* of the data given the parameters, $f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta})$, where $\boldsymbol{x}_{0:t} = (x_0, ..., x_t)$ is observed data.

Posterior distribution The *posterior distribution* is accessed via Bayes' Theorem which states

$$f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t}) = \frac{f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{x}_{0:t})},$$

where $f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})$ is the posterior distribution, $f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta})$ is the likelihood and

$$f(\boldsymbol{x}_{0:t}) = \int_{\Theta} f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

is a normalising constant (with respect to $\boldsymbol{\theta}$). Often the constant term, $f(\boldsymbol{x}_{0:t})$, is not available in closed form, or is not computable, and we only have the proportional relationship

$$f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t}) \propto f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

As a result, numerical methods are needed to approximate the posterior distribution, such as Markov chain Monte Carlo (Section 2.2.5).

Properties Like maximum likelihood inference, Bayesian inference also has many nice properties. Under certain regularity conditions the following can all be shown [40]:

- The posterior distribution is consistent: as sample size grows, the posterior distribution converges in distribution to a point mass at the true parameter. Formally, let $\{f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}_{0:t})\}_{t\in\mathbb{N}}$ be a sequence of posterior distributions, and let $\boldsymbol{\theta}^*$ be the true parameter that generated the data $\boldsymbol{x}_{0:t}$. The posterior distribution is called consistent if, for every neighbourhood $\mathcal{N}$ of $\boldsymbol{\theta}^*$,

  $$\int_{\mathcal{N}} f_{\boldsymbol{\theta}}(\boldsymbol{u}|\boldsymbol{x}_{0:t})d\boldsymbol{u} \to 1,$$

  as $t \to \infty$, with probability 1 (i.e. under the assumption the data is generated from the distribution $f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta}^*)$). It might seem odd that there exists a 'true parameter value' in a Bayesian setting since $\boldsymbol{\theta}$ is a random variable, however, consider the example of observing a single time series $\boldsymbol{x}_{0:t} = (x_0, ..., x_t)$, then the assumption that a single parameter vector generated this series is natural.

- Assuming again that there exists a true parameter value $\boldsymbol{\theta}^*$, then, as the sample size grows the posterior distribution is asymptotically normally distributed. Formally,

  $$\lim_{t\to\infty} \int_{\Theta_{\boldsymbol{u}}} \left| f_{\boldsymbol{U}_t}(\boldsymbol{u}|\boldsymbol{X}_{0:t}) - \frac{1}{\sqrt{2\pi}} det(I(\boldsymbol{\theta}^*))^{-1/2} e^{-\frac{1}{2}\boldsymbol{u}' I(\boldsymbol{\theta}^*)\boldsymbol{u}} \right| d\boldsymbol{u} = 0,$$

  with probability 1 when the data is generated by the distribution $f(\boldsymbol{x}_{0:t}|\boldsymbol{\theta}^*)$, where $f_{\boldsymbol{U}_t}(\boldsymbol{u}|\boldsymbol{X}_{0:t})$ is the posterior distribution of $\boldsymbol{U}_t = \sqrt{t}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ given $\boldsymbol{X}_{0:t}$, $I(\boldsymbol{\theta}^*)$ is the Fisher information as in (2.1) and $\Theta_{\boldsymbol{u}}$ is the parameter space of the transformation

$\boldsymbol{U}_t$. This observation is useful for approximations to the posterior distribution such as Laplace's approximation, which provides a way to approximate posterior distributions with a Normal distribution without heavy computations.

- A consequence of the consistency of posterior distributions is robustness to the choice of prior distribution. Suppose $\pi_1(\cdot)$ and $\pi_2(\cdot)$ are prior distributions that are positive and continuous at $\boldsymbol{\theta}^*$. Furthermore, suppose the posterior distributions constructed with these prior distributions, $f_1(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})$ and $f_2(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})$ respectively, are both consistent at $\boldsymbol{\theta}^*$, then

$$\sup_{A\in\Theta}\left|\int_A f_1(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})d\boldsymbol{\theta} - \int_A f_2(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})d\boldsymbol{\theta}\right| \to 0,$$

as $t \to \infty$. This says regardless of the choice of prior distributions, as long as the posterior distributions are consistent, then we end up with the same posterior distribution asymptotically.

Representing the posterior Another aspect of Bayesian inference is how to best represent posterior inferences. The natural way is to present the entire posterior distribution, but sometimes it is more informative to present simpler summaries of the posterior distribution. For example, when the posterior is high-dimensional the entire posterior cannot be visualised so it is hard to get any intuitive sense for what the posterior distribution looks like. Other common ways to report posterior inferences are to use marginal distributions, i.e. report $f(\theta_i|\boldsymbol{x}_{0:t})$ (or $f(\theta_i,\theta_j|\boldsymbol{x}_{0:t})$ when there is some important dependence between parameters θ_i and θ_j), or to calculate posterior summaries from the posterior distribution. For example, point estimates of the parameters might be useful and one could report the posterior mean, median or *maximum a posteriori estimator* (MAPE, which is defined as $\arg\max_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})$), and the spread of the posterior distribution can be summarised using the posterior variance.

Bayesian point estimates are sometimes justified by showing that they minimise some *loss function*, $L(\boldsymbol{\theta}^*, \boldsymbol{\theta}(\boldsymbol{x}_{0:t}))$. Let $\boldsymbol{\theta}(\boldsymbol{x}_{0:t})$ be an estimator of the parameter $\boldsymbol{\theta}^*$, and suppose we are interested in minimising the expected loss as measured by

$$\mathbb{E}\left[L(\boldsymbol{\theta}^*, \boldsymbol{\theta}(\boldsymbol{x}_{0:t}))\right] = \mathbb{E}\left[(\boldsymbol{\theta}^* - \boldsymbol{\theta}(\boldsymbol{x}_{0:t}))^2\right],$$

where the expectation is taken with respect to the joint distribution $f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})$. We can show that the estimator that minimises this expected loss is the posterior mean

$$\boldsymbol{\theta}(\boldsymbol{x}_{0:t}) = \int_\Theta \boldsymbol{\theta} f(\boldsymbol{\theta}|\boldsymbol{x}_{0:t})d\boldsymbol{\theta}.$$

Similarly, when the loss function of interest is

$$L(\boldsymbol{\theta}^*, \boldsymbol{\theta}(\boldsymbol{x}_{0:t})) = |\boldsymbol{\theta}^* - \boldsymbol{\theta}(\boldsymbol{x}_{0:t})|,$$

then the posterior median minimises the expected loss.

2.2.5 Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo is a broad class of algorithms that involve simulating Markov chains, typically for the purpose of sampling from probability distributions. The general idea is to construct a Markov chain with limiting distribution f, then, by simulating a long realisation of this Markov chain, we can assume the chain is close to stationary and the collection of samples towards the end of the chain are approximately distributed according to f (although the samples may *not* be independent). The theory behind MCMC techniques is vast and a popular area of research, which shows its importance. MCMC algorithms provide an alternative to other sampling techniques, such as *inverse sampling*, which is not always tractable, or *rejection sampling*, which can take prohibitively long to compute a sufficient number of samples. This section is mostly based on Robert and Casella [88].

Metropolis Algorithm The simplest MCMC algorithm is the *Metropolis algorithm* [74] which has the following structure. Suppose we want to construct a Markov chain with $\frac{f(x)}{c}$ as its limiting distribution, where c is a normalising constant, and $f(x)$ is a probability density on a state space $\mathcal{S}$. We explicitly write the normalising constant, c, to emphasise the fact that these algorithms do not depend on it. Let $q(x, \cdot)$ be a *symmetric* probability density for each $x \in \mathcal{S}$, that is $q(x, y) = q(y, x)\ \forall y \in \mathcal{S}$. Here, q is known as the *proposal distribution*. Given the current state of the chain is $X_n = x$, the Metropolis algorithm simulates transitions of a Markov chain using the following:

1. Simulate y from the distribution $q(x, \cdot)$.

2. Calculate the acceptance ratio α and acceptance probability a,

$$\alpha(x, y) = \frac{\frac{f(y)}{c} q(y, x)}{\frac{f(x)}{c} q(x, y)} = \frac{f(y)}{f(x)},$$
$$a(x, y) = \min(1, \alpha(x, y)).$$

3. Set $X_{n+1} = y$ with probability $a(x, y)$, otherwise set $X_{n+1} = x$.

In Step 2 of the Metropolis algorithm, the density $q(x, y)$ cancels from the top and $q(y, x)$ from the bottom of the ratio since q is symmetric. Also notice that the ratio $\alpha(x, y)$ does not depend on the normalising constant c. The significance of this is that the density f does not need to be normalised to implement this algorithm, which makes MCMC algorithms particularly useful in Bayesian settings where the constant, c, may be intractable. That f is the limiting distribution of this Markov chain follows immediately after showing that the transition kernel of the Metropolis chain,

$$K(x, A) = m(x)\mathbb{I}(x \in A) + \int_A q(x, y)a(x, y)dy,$$

is reversible, where $m(x) = 1 - \int_{\mathcal{S}} q(x, y)a(x, y)dy$, and $\mathbb{I}(x \in A)$ is the identity operator.

Metropolis-Hasting algorithm In 1970 Hastings [46] extended the Metropolis algorithm to use arbitrary proposals $q(x, y)$ (i.e. q no longer has to be symmetric). This extension means the acceptance probabilities in Step 2 remain,

$$a(x, y) = \min\left(1, \frac{f(y)q(y, x)}{f(x)q(x, y)}\right),$$

without the benefit of the cancellation of the $q(x, y)$ like in the Metropolis algorithm. The arguments as to why f is the stationary distribution of this chain remain the same. This algorithm is referred to as the *Metropolis-Hasting (MH) algorithm*. It turns out that the choice of proposal distribution is relatively arbitrary, in that the chain will have the correct stationary distribution regardless of which proposal distribution we choose (up to some not very restrictive conditions). However, since we do not have the luxury of simulating the chain for an infinite number of transitions, this choice is crucial so that the chain is close to stationary in a computationally feasible number of steps.

Gibbs sampler Another popular MCMC algorithm is the *Gibbs sampler* [39]. The Gibbs sampler is applicable when we wish to sample a random vector, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, from a distribution $f(\boldsymbol{\theta})$, and the conditional distributions,

$$f_{\theta_i}(\cdot|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p),$$

for each $i = 1, ..., p$, can be derived. Suppose the current state of the chain is $\boldsymbol{\theta}_n = (\theta_{n,1}, ..., \theta_{n,p})$, the Gibbs sampler updates the chain using the following steps:

1. Sample $\theta_{n+1,1}$ from $f_{\theta_1}(\cdot|\theta_{n,1}, ..., \theta_{n,p})$.
2. Sample $\theta_{n+1,2}$ from $f_{\theta_2}(\cdot|\theta_{n+1,1}, \theta_{n,3}, ..., \theta_{n,p})$.

3. Sample $\theta_{n+1,3}$ from $f_{\theta_3}(\cdot|\theta_{n+1,1}, \theta_{n+1,2}, \theta_{n,4}, ..., \theta_{n,p})$.
   $\vdots$

$p-1$. Sample $\theta_{n+1,p-1}$ from $f_{\theta_{p-1}}(\cdot|\theta_{n+1,1}, ...\theta_{n+1,p-2}, \theta_{n,p})$.

p. Sample $\theta_{n+1,p}$ from $f_{\theta_p}(\cdot|\theta_{n+1,1}, ..., \theta_{n+1,p-1})$.

It is interesting to note that the Gibbs sampler is a special case of the Metropolis-Hastings algorithm. To see this, suppose that $q(x, y) = f_{\theta_i}(y|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$ is the proposal distribution of an MH algorithm. Note that this proposal can be written as

$$q(x, y) = f_{\theta_i}(y|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p) = \frac{f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)}{f(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)}.$$

Then, the acceptance ratio is

$$\begin{aligned}
\alpha &= \frac{f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)q(y, x)}{f(\theta_1, ..., \theta_{i-1}, x, \theta_{i+1}, ..., \theta_p)q(x, y)} \\
&= \frac{f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)\dfrac{f(\theta_1, ..., \theta_{i-1}, x, \theta_{i+1}, ..., \theta_p)}{f(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)}}{f(\theta_1, ..., \theta_{i-1}, x, \theta_{i+1}, ..., \theta_p)\dfrac{f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)}{f(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)}} \\
&= \frac{f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)f(\theta_1, ..., \theta_{i-1}, x, \theta_{i+1}, ..., \theta_p)}{f(\theta_1, ..., \theta_{i-1}, x, \theta_{i+1}, ..., \theta_p)f(\theta_1, ..., \theta_{i-1}, y, \theta_{i+1}, ..., \theta_p)} \\
&= 1.
\end{aligned}$$

Thus, an MH algorithm with this proposal always accepts updates and therefore does the same thing as the Gibbs sampler.

The way the Gibbs sampler is presented here might suggest that we must do the updates in a specific order, but this is not the case. The theory of the Gibbs sampler holds if we execute the update steps in any order, or if we update the parameters in blocks, or even if we randomly choose a parameter to update at each step, so long as we use the appropriate conditional probability.

Metropolis-within-Gibbs An extension to the Gibbs sampler is the *Metropolis-within-Gibbs* (also known as *block-Metropolis-Hastings*) algorithm. In this extension to the Gibbs sampler, the update steps are replaced by Metropolis-Hastings updates. That is, rather than sample directly from $f_{\theta_i}(\cdot|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$, we instead sample using the following:

1. Simulate θ_i^* from $q_i(\cdot|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_p)$.

2. Calculate

$$a = \min\left(1, \frac{f(\theta_1, ..., \theta_{i-1}, \theta_i^*, \theta_{i+1}, ...\theta_p) q_i(\theta_i | \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_p)}{f(\theta_{n,1}, ..., \theta_{i-1}, \theta_i, \theta_{i+1}, ...\theta_p) q_i(\theta_i^* | \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ...\theta_p)}\right).$$

3. Set $\theta_i = \theta_i^*$ with probability a, otherwise set $\theta_i = \theta_i$,

where q_i is some proposal distribution. The advantage of the Metropolis-within-Gibbs sampler is that it can be used even when the conditional distributions, f_{θ_i}, are not known. The Metropolis-within-Gibbs algorithm can also be seen as an extension of the MH algorithm. The original MH algorithm is a multivariate update-all-elements-at-a-time algorithm, whereas this extension breaks up the update step into univariate moves. Other modifications of the MH algorithm also exist, for example, it can be beneficial to update parameters in blocks of parameters that are strongly dependent.

Choosing an algorithm As mentioned before, the choice of proposal distribution for MH-style updates is essentially arbitrary but it does affect the rate of convergence of the chain to its stationary distribution. Similarly, whether a block-wise algorithm or an update-all-elements-at-once algorithm is used, also affects the rate of convergence of the chain. One advantage of using a block-updating algorithm is that suitable proposal distributions are more obvious to find. For example, suppose the search for a suitable proposal distribution is restricted to normal distributions centred around the current position of the chain. For a p-dimensional parameter vector, if an update-all-elements-at-once algorthm is chosen, then this leaves the covariance matrix of the proposal distribution to specify, which has $\frac{1}{2}n(n+1)$ elements. If a one-parameter-at-a-time algorithm is used instead (such as a Metropolis-within-Gibbs algorithm) then only p variances need to be specified to define the proposal distributions. Of course, there are also good reasons not to use a block-updating algorithm, for example it is known that block-wise algorithms exhibit high correlation between steps of the algorithm and, as a result, the algorithm might take a prohibitively long time to sample the entire state space. Literature investigating optimal proposal distributions for MCMC algorithms is available [89] but the models for which this theory hold is limited, although numerical examples show that the theory can be insightful for more general problems [90].

Adaptive methods One way to choose proposal distributions is to use an adaptive algorithm. Adaptive MCMC algorithms typically work by restricting the proposal distribution to a certain family of distributions (for example, normal distributions centred around the current state of the chain) and the algorithm automatically adjusts the

variance (or covariance matrix) of the proposal distribution(s) to target a prespecified acceptance ratio that is theoretically optimal (such as that in [89]). Such an adaptive algorithm was developed by Haario *et al.* [43], and other examples of adaptive algorithms are in Roberts and Rosenthal [90].

Choice of proposal distribution A sufficient condition on the proposal distribution of the Metropolis-Hastings algorithm that guarantees f is the stationary distribution is the following [88].

Theorem 2.1. *Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain produced by a Metropolis-Hastings algorithm. For every proposal distribution $q(x_t|x_{t-1})$ whose support includes the support of $f(\cdot)$, the transition kernel of the chain is reversible, and f is the stationary distribution of the chain.*

Another sufficient condition on the proposal of the MH algorithm is the following [91].

Lemma 2.2. *Assume f is bounded and positive on every compact set of its support. If there exists $\varepsilon > 0$ and $\delta > 0$ such that*

$$q(x_t|x_{t-1}) > \varepsilon \quad if \quad |x_t - x_{t-1}| < \delta,$$

then the Metropolis-Hastings chain is f-irreducible and aperiodic. Moreover f is the stationary distribution of the chain.

Burn-in One issue with MCMC algorithms is that convergence only holds asymptotically, in that the chain reaches stationarity only after an infinite number of samples. As one obviously cannot simulate an infinitely long Markov chain, we must decide when the chain is close to stationary. One common technique to assess stationarity of MCMC chains is to look at trace plots of the chain, which plot the value of the chain against iteration. From trace plots we look to see when the chain 'settles down' and shows behaviour we would expect from a stationary chain. We can then conclude whether the assumption of stationarity is reasonable, or more samples are needed. More rigorous methods are described in [20]. The samples of the MCMC chain are useful only if they are (close to) stationary, thus the portion of the chain that is deemed non-stationary is discarded as *burn in*.

Data-augmented MCMC As mentioned in Section 2.3.3, the likelihood function is not always easy to evaluate for MRS models, particularly MRS models of Type II with independent regimes that evolve at all times, and Type III with independent regime

that evolve only when observed. MCMC algorithms rely on being able to simulate long realisations of the MCMC chain. So, for MCMC algorithms to be an effective solution, the likelihood function needs to be computed efficiently, but sometimes this is not possible. In some cases this can be overcome by using a technique called *Data Augmented Markov Chain Monte Carlo* (DA-MCMC), which we shall use for MRS models.

To set the scene for DA-MCMC, suppose we are fitting a model to data, $\boldsymbol{x}$, in a Bayesian setting and that numerical methods are needed to compute the posterior distribution. Furthermore, suppose that the likelihood function $f(\boldsymbol{x}|\boldsymbol{\theta})$ is not computationally feasible, but can be written as a marginal distribution

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{R}\in\mathcal{R}} f(\boldsymbol{x}, \boldsymbol{R}|\boldsymbol{\theta}),$$

where $\boldsymbol{R} = (R_0, ..., R_T)$ is the hidden sequence of regimes and $\mathcal{R}$ is the set of all possible regime sequences. Note that $\boldsymbol{R}$ could be a vector of continuous random variables too, in which case the sum would be replaced by the appropriate integrals. In a standard MCMC setting, samples from the posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{x})$, are obtained via the proportionality relationship

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

but since $f(\boldsymbol{x}|\boldsymbol{\theta})$ is computationally intractable this cannot be implemented. However if the joint distributions, $f(\boldsymbol{x}, \boldsymbol{R}|\boldsymbol{\theta})$, are easy to evaluate, DA-MCMC provides a way for us to proceed. In a DA-MCMC algorithm the joint posterior distribution, $f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x})$, is inferred, and MCMC is used to sample from this posterior distribution via the proportionality relationship

$$f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x}) \propto f(\boldsymbol{x}, \boldsymbol{R}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

To obtain the (marginal) posterior distribution of interest, $f(\boldsymbol{\theta}|\boldsymbol{x})$, we need to integrate the joint posterior distribution over $\boldsymbol{R}$, that is,

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{\mathcal{R}} f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{\theta}).$$

Conveniently, this is equivalent to ignoring the elements $\boldsymbol{R}$ in the DA-MCMC chain, and estimating the posterior distribution using the remaining dimensions of the chain, with no extra computations needed.

2.2.6 Model checking and selection

Once a model is fitted to data, we would like to assess how well it fits the data (model checking) and to be able to compare models (model selection). Commonly used methods to compare two models are the *likelihood ratio test* [79] or the *Akaike Information Criterion* (AIC) [1]. AIC can sometimes be more useful since it accounts for model complexity and does not rely on the models being nested. Derived as the solution to the problem of choosing a model that has the best out-of-sample predictive error, the AIC statistic is $\text{AIC} = 2p - 2\ell(\widehat{\boldsymbol{\theta}})$, where p is the number of parameters in the model. An implicit assumption of likelihood-based measures of goodness-of-fit is that each model is capturing exactly the same data [84].

To assess model fit, goodness-of-fit statistics can be used. A simple way to check distributional assumptions for MRS models is to compare the stationary distribution of the observed data with the stationary distribution of candidate models using, for example, the Kolmogorov-Smirnov test, which compares empirical and theoretical cumulative distribution functions. This checks stationarity assumptions only and does not warn us if within-regime distributional assumptions are valid. A goodness-of-fit procedure developed by Janczura and Weron [62] enables one to check distributional assumptions within each regime for MRS models. However, as it is presented, their method relies on the EM-like algorithm being correct, which is not the case (see Section 3.1). It may be that the theory of [62] still holds but more work is needed to show this.

A Bayesian approach to model checking is *posterior predictive checks* [37]. The idea is to sample parameters (and latent variables) from the posterior distribution, use these samples to calculate statistics of the observed data, and compare these to statistics calculated under the assumption that the model is true. Repeating this for many samples can warn us if there are any obvious ways in which our model fails.

2.2.7 Wavelet and Fourier filtering

In this context, filtering refers to estimating deterministic components of electricity prices. Filters take electricity price series as inputs, and output a smoothed series. Wavelet filtering is a popular technique used in the electricity price modelling literature to model long-term deterministic components, and simulation studies have shown that wavelet filters are good at capturing the complex deterministic patterns in electricity prices for model estimation purposes [57, 82, 83]. However, due to the fact that wavelet functions are localised in time, there are issues with out-of-sample forecasting of trend components when using wavelets.

To summarise the following discussion: electricity price series can be smoothed by projecting the data onto dilations and translations of father wavelets that are only able to represent the data up to some level of detail. This projection takes the form of a linear combination of dilations and translations of the father wavelet (Equation (2.4)). Due to nice properties of mother and father wavelets, and the families of functions defined from them, there exist simple recursive formulas to calculate the coefficients $\lambda_{m,n}$ in Equation (2.4).

Now, in more details, wavelet filtering is achieved by progressively projecting time series data onto progressively coarser bases. The bases onto which the data is projected are defined by wavelet functions. There are two types of wavelet functions, *mother* wavelets, denoted ψ and *father* wavelets, denoted ϕ. It is easiest to first define the mother wavelet and the family of functions it defines, since this then allows us to present families defined from the father wavelets in an insightful way.

The following is closely based on Valens [97] and the interested reader should consult this (and references therein) for more details.

The mother wavelet is a function, $\psi(x) : \mathbb{R} \to \mathbb{R}$, which is non-zero only on a compact interval of its domain and has the properties

$$\int_{\mathbb{R}} \psi(x)dx = 0, \quad \int_{\mathbb{R}} \psi^2(x)dx = 1, \quad \int_{\mathbb{R}} \frac{|s_\infty(x)|^2}{x}dx < \infty,$$

where s_∞ is the Fourier transform of ψ. We use this to define a family of wavelet functions,

$$\psi_{m,n}(x) = \frac{1}{\sqrt{2^m}}\psi\left(2^{-m}x - n\right)$$

for $m, n \in \mathbb{Z}$. Each function $\psi_{m,n}$ has the same properties as ψ; they integrate to 0 and are non-zero on a compact interval of their domain. Moreover, they also have the property

$$\int_{-\infty}^{\infty} \psi_{m,n}(x)\psi_{m',n'}(x)dx = 0$$

if $m \neq m'$ or $n \neq n'$, so $\{\psi_{m,n} : m, n \in \mathbb{Z}\}$ are orthogonal. Any square-integrable function, f, can be represented as

$$f(x) = \sum_{m,n\in\mathbb{Z}} a_{m,n}\psi_{m,n}(x),$$

for some coefficients $a_{m,n}$, $m, n \in \mathbb{Z}$; this is the projection of f on to the wavelet family defined using ψ. In the context of modelling data, this means that the data can be modelled exactly (without error) by wavelet functions.

For a finite number of discrete, equally- spaced time series observations, a finite number of terms are needed in the projection onto the wavelet basis. Assume that the length of the observation sequence is $N = 2^M$. Then the data are completely captured by

$$f(x) = \sum_{m=0}^{M} \sum_{n=1}^{2^{M-m}} a_{m,n}\psi_{m,n}(x), \tag{2.3}$$

in the sense that f passes through every data point.

Now, the father wavelet is also non-zero only on a compact set of its domain, and families of wavelets are defined from it in the same way; by scaling and dilation,

$$\phi_{m,n}(x) = \frac{1}{\sqrt{2^m}}\phi\left(2^{-m}x - n\right).$$

Father wavelets must have the property that, for a given m,

$$span\{\psi_{m,n}|n \in \mathbb{Z}\} + span\{\phi_{m,n}|n \in \mathbb{Z}\} = span\{\psi_{m-1,n}|n \in \mathbb{Z}\}.$$

As a consequence, since our data can be represented as the sum in Equation (2.3), then the data can also be represented as

$$f(x) = \sum_{n=1}^{N} \lambda_{0,n}\phi(x - n),$$

for some coefficients $\lambda_{0,n}$. We also have the property

$$f(x) = \sum_{n=1}^{N} \lambda_{0,n}\phi(x - n) = \sum_{n=1}^{N/2} \lambda_{1,n}\phi(2^{-1}x - n) + \sum_{n=1}^{N/2} a_{1,n}\psi_{1,n}(x).$$

This decomposition comes with an insightful interpretation: the first sum represents the 'trend' while the second represents the 'detail' in the data. Applying this decomposition recursively J times, we can represent the data as

$$f(x) = \sum_{n=1}^{\frac{N}{2^J}} \lambda_{J,n}\phi(2^{-J}x - n) + \sum_{j=1}^{J} \sum_{n=1}^{\frac{N}{2^j}} a_{j,n}\psi_{j,n}(x).$$

The first sum captures the trend at scale J, and the second sum captures all the details up to and including scale J. This decomposition has nice properties that give simple recursive formulas for the coefficients $a_{m,n}$ and $\lambda_{m,n}$, which are used in practice. The trend in the data is estimated as the first sum

$$\widehat{f}(x) = \sum_{n=1}^{2^{-J}N} \lambda_{J,n}\phi(2^{-J}x - n), \tag{2.4}$$

and is sometimes referred to as an S_J approximation, or decomposition, of the data.

There are many different functions ψ that define wavelet families, and their corresponding father wavelet, each of which has their own desirable properties. Popular classes of functions ψ are Coiflets and Daubechies wavelets, which are typical in electricity price modelling. The wavelets in the Daubechies family are indexed by the number of vanishing moments they possess, which can be thought of as a measure of the complexity of signals they can replicate.

One issue with implementing wavelet filtering is that the dataset must be of a length that is a power of 2, which is rarely the case. This can be remedied by, for example, making the data circular, or appending repeated mean or median values to each end of the dataset to increase it to an appropriate length.

Wavelet filtering is more flexible than filters based on periodic functions (such as Fourier filtering), since the wavelet bases are localised in frequency and time, whereas the periodic functions are only localised in frequency. The result is that periodic filters are not as robust to outliers and cannot represent aspects of time series that are not periodic.

2.3 Literature review

2.3.1 Modelling detrended electricity prices

Typically electricity prices are modelled as a sum of two processes, a deterministic trend component, T_t, and a stochastic component, X_t, so the price at time t is given by $P_t = T_t + X_t$ (or $P_t = T_t X_t$). Our main interest is in the stochastic component X_t, but we also need to deal with the trend component, the literature of which we review in Section 2.3.4.

In this section we briefly review some of the many stochastic models used for electricity prices, before specifically focusing on MRS models in Section 2.3.2. We categorise models, using a similar structure to Weron [102], into the following categories: reduced form, statistical-forecast, fundamental, spike only, agent based and computer intelligence. Of course, there is some overlap between these categories. Each type of model has its advantages and disadvantages and serves its own purpose.

Reduced form models The goal of reduced form models is to capture the probabilistic properties of electricity prices for use in derivative valuation and risk management.

These models should accurately replicate electricity-price behaviour. Two popular models in this category are the *MRS models*, which are reviewed extensively in Section 2.3.2, and *jump diffusion processes*.

Diffusion processes are continuous-time continuous-state stochastic processes that have continuous sample paths with probability 1. The simplest diffusion process is standard Brownian motion, denoted $\{B_t\}_{0\leq t}$. Standard Brownian motion is characterised by normally distributed increments,

$$B_t - B_s \sim N(0, t-s) \quad for \quad 0 \leq s \leq t,$$

independent increments ($B_{t_1} - B_{s_1}$ is independent of $B_{t_2} - B_{s_2}$ provided that the intervals $[s_1, t_1]$ and $[s_2, t_2]$ do not overlap), continuous sample paths with probability 1, and $B_0 = 0$. Traditional financial markets are often modelled by geometric Brownian motion; at time t, the price P_t is related to a standard Brownian motion through

$$P_t = P_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma B_t},$$

where P_0 is an initial price and μ and σ are parameters of the geometric Brownian motion. There is an abundance of literature modelling financial markets with geometric Brownian motion, including the celebrated Black-Scholes model. However, due to the characteristics of electricity prices – spikes, drops, negative prices and mean reversion – these models are unsuitable for electricity markets [14, 35].

A popular model for electricity prices is the stochastic jump-diffusion model [14, 35, 52], where prices, or more commonly log-prices, are modelled as the sum of an Ornstein-Uhlenbeck (O-U) process and a jump process. An O-U process is related to standard Brownian motion through the equation,

$$P_t = P_0 e^{-\theta t} + \mu\left(1 - e^{-\theta t}\right) + \sigma \int_0^t e^{-\theta(t-s)} dB_t,$$

where P_0 is an initial condition, μ, θ and σ are parameters and the integral is a *stochastic integral* with respect to standard Brownian motion. More general specifications of O-U processes exist where the stochastic integral is with respect to more general *Lévy processes* rather than Brownian motion.

The O-U process is a mean-reverting process – if we ignore the stochastic fluctuations, the process naturally tends to the value μ. O-U processes have the property that, given the value of the process at time s, the distribution of the process at time t, $t > s$, follows

a normal distribution,

$$X_t|X_s \sim \mathrm{N}\left(\mu + (X_s - \mu)e^{-\theta t}, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta t}\right)\right).$$

If we consider observing an O-U process at discrete times, then the resultant discrete-time process $\{X_t\}$ is a Gaussian AR(1) process. This fact is often used to simplify models with O-U process for electricity prices, since prices are only observed at discrete equally-spaced times.

The behaviour of jump-diffusion models for electricity prices is to follow O-U dynamics between spikes, and jump discontinuously at random times. Jump times are often specified as a Poisson process and jump sizes have been specified as log-normal [16, 35], normal [17], exponential [42] or truncated exponential random variables [38]. More general processes have also been studied, for example, models with more than one jump process [42], allowing for upward and downward jumps, incorporating periodicity into the rate of the jump arrival process, or allowing the arrival rate of jumps to depend of price level [38].

For models consisting of a single O-U process with additive jump components, the parameter θ of the O-U process must capture the mean reversion between, and immediately after, spikes, even though it is known that prices return to the mean level at a much faster rate immediately following spikes than between spikes [49]. To overcome this, Benth *et al.* [15] model the spot price using a sum of generalised O-U processes driven by non-Gaussian processes, with each O-U process having its own mean reversion term. Similarly, Gonzalez *et al.* [42] model prices as a sum of a Gaussian O-U process and non-Gaussian O-U processes which capture jumps, and they allow each jump process to have its own mean reversion parameter. They fit their model to the Amsterdam Power Exchange United Kingdom (APXUK) and European Energy Exchange (EEX) markets using Bayesian methods, and show that their model fits well using posterior predictive checks (Section 2.2.6).

All of these models are continuous-time models, however, electricity price evolution is a discrete-time process as prices are only realised when the market is dispatched. Continuous-time processes, particularly continuous-time jump processes, can be complicated to fit to discretely observed data since, for example, we do not observe the process at jump times that occur between observations. For this reason, either approximate methods such as approximate likelihood [38] or data-augmented Bayesian methods [42] are used. MRS models of Type I can be seen as a discrete-time analogue to some jump-diffusion models and are often easier to fit to data.

Statistical-forecast Models Statistical models are typically built to make forecasts (usually point forecasts) for electricity markets or can also be used as fundamental models to determine the effects of fundamental price drivers. There is a vast array of statistical techniques and models in the literature ranging from simple, so called *similar day* methods, which we elaborate on below, to complex non-linear ARMA-GARCH models incorporating exogenous factors.

Similar day methods are sometimes used as benchmark models for more extravagant methods [94]. A similar day method works, for example, by comparing the attributes of the forecast day to attributes of previous days, then the price forecast is made as the average of the prices on all previous similar days. Attributes might include day of the week, whether the day is a public holiday, the forecast weather, time of year or available generation.

Autoregressive (AR) models are popular statistical models for financial time series. When applied to electricity markets, AR models often include exogenous factors and are sometimes labelled ARX models, e.g. the model defined by

$$P_t = \phi_1 P_{t-24} + \phi_2 P_{t-48} + \phi_3 P_{t-168} + \phi_4 mp_t + \psi_1 z_t + d_1 D_{\text{Mon}} + d_2 D_{\text{Sat}} + d_3 D_{\text{Sun}} + \varepsilon_t,$$

where P_t is the logarithm of the current price, P_{t-24}, P_{t-48} and P_{t-168} are the logarithms of the prices at the same hour yesterday, two days ago and last week, mp_t is the minimum of yesterday's log-prices, z_t is the logarithm of the load at time t, D_{Mon}, D_{Sat} and D_{Sun} are indicators for Monday, Saturday and Sunday, respectively, ϕ_i, $i = 1, 2, 3, 4$, ψ_1, d_j, $j = 1, 2, 3$ are parameters which are estimated from the data, and $\{\varepsilon_t\}$ is a sequence of $\mathrm{N}(0, 1)$ random variables. This model was applied in Weron and Misiorek [104] to the Nord-Pool and Pennsylvania-New Jersey-Maryland (PJM) markets.

Typically, simple AR models are not adequate to capture characteristics observed in electricity markets [56] and more complex models are considered. For example, Chen and Bunn [22] employ advanced regression techniques to fit mixture regression models to electricity prices and use these to obtain point forecasts. They compare their models to MRS models of Type I with exogenous variables and conclude that while MRS models fit the data best, they may overfit the data and perform poorly out-of-sample.

Swider and Weber [56] fit ARMA, ARMA-GARCH, Gaussian mixture and MRS models of Type I to the EEX market and compare models using a range of metrics: likelihood values, Bayesian Information Criterion, R^2 value, and mean error. They conclude that standard Gaussian ARMA process are not adequate to model electricity prices.

Panagiotelis and Smith [85] use a vector autoregressive model with skew-t-distributed errors to model New South Wales electricity prices and fit their model in a Bayesian

setting. To evaluate their model, they obtain forecasts over a 72-hour horizon and use a continuous ranked probability score (CRPS) to assess the quality of their forecasts, where the CRPS is defined as $\int_{-\infty}^{\infty} (F(h) - \mathbb{I}(h > y^{obs}))^2 dh$, where F is the model's predictive cumulative distribution function.

Recently, Pape *et al.* [86] apply sophisticated regression and time series techniques to forecast the distribution of electricity prices at an hourly resolution. They use multiple regression on log-prices with an offset to estimate a function for mean prices using ordinary least squares. They then transform the residuals estimated using this mean function, and model these residuals with an ARMA(1,1)-GARCH(1,1) process to capture the distributional characteristics of electricity prices. They conclude that their model can capture the complex nature of electricity prices and produce accurate point and density forecasts.

Nowotarski *et al.* [82] review and implement a range of statistical models to forecast electricity prices and investigate the forecasting performance of combining individual forecasts models. They conclude that combining forecasts of statistical models can be advantageous for forecasting performance.

Fundamental models These models capture the fundamental drivers of electricity prices, such as load, weather and fuel prices, and quantifies their effect on prices. For example, Kanamura and Ōhashi [65] model the electricity supply function of the PJM market using a piece-wise polynomial function, where the range of the two pieces of the polynomial function define a base regime and a spike regime. They use the fact that, at the market-clearing price, supply and demand are equal to obtain a relationship between price and demand. Then they model electricity demand using an AR(1) process and relate prices to demand. Applying their methodology to real data, they show that the frequency of jumps between prices in the base regime and spike regime are non-constant over time.

Another example is Karakatsani and Bunn [66], where the effect of fundamentals and prices are both stochastic. They model prices P_{jt}, where j indicates the time of day and t represents the day, using

$$\begin{aligned} P_{jt} &= \boldsymbol{\beta}'_{jt}\boldsymbol{x}_{jt} + \varepsilon_{jt} \quad \text{measurement equation,} \\ \boldsymbol{\beta}_{jt} &= \boldsymbol{\beta}_{j(t-1)} + \boldsymbol{v}_{jt} \quad \text{transition equation,} \end{aligned}$$

where $\boldsymbol{\beta}_{jt}$ are time-varying regression coefficients, $\boldsymbol{x}_{jt}$ are regressors, $\varepsilon_{jt} \sim$ i.i.d. $\mathrm{N}(0, \sigma^2_{\varepsilon_j})$ and $\boldsymbol{v}_{jt} = (v_{jt1}, ..., v_{jtk}) \sim$ i.i.d. $\mathrm{N}_k(0, \Sigma_j)$ with $\Sigma_j = diag(\sigma^2_{v_{jk}})$. They conclude that capturing the time-varying nature of prices can improve model forecasts. Some applications

of MRS models also fall into this category, such as Norén [80], Knapik [68], and Mount [77].

Spike-only models The term 'spike-only models' is used to describe models that are concerned with modelling the arrival of price spikes, and are not necessarily concerned with the actual value of prices. That is, the arrival of spike events is modelled as a point process. Researchers in this area focus on modelling the rate of arrival of spikes, typically using time-varying functions which include exogenous information. Examples are Clements *et al.* [26] who use multivariate self-exciting marked point processes to model spikes in the National Energy Market (NEM) and capture the dependence between connected markets. Herrera and González [48] use self-exciting marked point processes to model markets in the NEM and conclude that the arrival of spikes depends on the time between spikes. They also show that their model can improve value at risk forecasts. Eichler *et al.* [32] use a logit model and autoregressive conditional hazard model to capture price spikes in the NEM and conclude that these models can improve spike forecasts.

Becker *et al.* [12] model prices in the NEM using Hawkes point processes and Poisson autoregressive models, and use these to determine exogenous variables that affect the arrival rate of spikes, such as load, temperature and the number of spikes in the previous day. They compare their models and show they exhibit different behaviour when applied to the same dataset, in particular, spike predictions from the Hawkes model are more variable than the Poisson models in periods with lots of spikes. They also evaluate forecasting performance of their models and show that the Hawkes model misses less spikes but at the cost of more false alarms.

Christensen *et al.* [24] is one of the earlier papers to propose the autoregressive conditional hazard model for electricity spike events and apply this to the NEM. Christensen *et al.* [25] propose the Poisson autoregressive model to capture persistence in spike events and show, using data from the NEM, that this persistence is a significant aspect of the model. The reader may have noticed that all of these papers model markets in the NEM. This is probably because the NEM has some of the most spiky markets in the world, in both frequency of spikes (e.g. QLD) and the size of spikes (e.g. SA) [49].

Agent-based models Agent-based models attack the problem by modelling individual market participants on both sides of the market, and then simulating market operation to determine prices. They are typically useful for investigating market design and policy changes in electricity markets, and not so relevant for derivative valuation or price forecasting [98, 102]. Agent based models are not particularly relevant to our

modelling approach and we do not investigate them further here; for a review of agent based models, see Ventosa *et al.* [98].

Computer intelligence The last category is computer intelligence models, which utilise recent developments in computing power to fit very complex non-linear models to data, usually for point forecasts. Examples of computer intelligence models are *neural networks* [23], which are complex non-linear functions which take data as inputs and produce price forecasts as outputs. The forecast combining method of Nowotarski *et al.* [82] can also be seen as a computer intelligence model. Computer intelligence models are not particularly relevant to this thesis as they lack probabilistic interpretation, and cannot be used for derivative pricing.

2.3.2 Development of MRS models for electricity prices

Some good review papers of MRS models for electricity prices are Janczura and Weron [59, 102]. The earliest applications of MRS models for electricity prices are Ethier and Mount [34] and Deng [30]. Ethier and Mount [34] use a MRS model of Type I (with dependent regimes) with two AR(1) regimes to model log daily on-peak electricity prices in American and Australian markets. They conclude the regime-switching nature of prices is better modelled by a Markov chain than a simple independent process, and there is significant evidence of different means and variances within the two regimes. Deng [30] uses a slightly different two-state MRS model of Type I (with dependent regimes). In [30], both regimes follow AR(1) dynamics with the same parameters, except one regime is also shifted by an exponentially distributed random variable.

In the models of Ethier and Mount [34] and Deng [30], prices must decay back to base levels following a spike; however, this is not consistent with observations from the market, where there is a more immediate return to base levels. For this reason, Huisman and Mahieu [53] introduce a three-regime MRS model of Type I (with dependent regimes) which separates the behaviour of base and spike prices. They use one regime to capture base prices, one regime to capture spikes, and one regime to return prices to base levels following a spike: we label these regimes b, s and r respectively. The authors specify a transition matrix of the form

$$P = \begin{array}{c} \\ r \\ b \\ s \end{array} \begin{array}{c} \begin{array}{ccc} r & b & s \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & p & 1-p \\ 1 & 0 & 0 \end{bmatrix} \end{array},$$

where p is the only parameter to be estimated. The dynamics of the hidden regime process is as follows. Suppose, for simplicity, the process starts in the base regime, then the process stays in the base regime for a geometrically distributed amount of time with parameter p, after which the process transitions to the spike regime for one time step, and then to the spike-reverting regime for one time step, and then back to the base regime where it stays for a geometrically distributed amount of time. There are clearly shortcomings of this model, namely the model does not allow for consecutive spikes while this is clearly a feature in the market [25].

Higgs and Worthington [49] apply the model of Huisman and Mahieu [53] to Australian electricity markets. In their paper they compare the performance of Huisman and Mahieu's model with a simple i.i.d. model, and an AR(1) model. They find that Huisman and Mahieu's model is best in terms of predicting prices, both in and out-of-sample, that the probability of switching to the spike regime varies across markets (5% in NSW up to 10% in VIC), the size of spikes vary between markets, with SA having the largest average spikes, and that interconnectors appear to have lowered prices in QLD and SA. They acknowledge the limitations of Huisman and Mahieu's model and flag this as an area for future research, along with extending MRS models to include exogenous factors and a multivariate analysis of the NEM markets.

In 2003, MRS models for electricity prices started to evolve in two directions simultaneously. One set of literature develops MRS models of Type II (with independent regimes), while another develops MRS models of Type I (with dependent regimes), including exogenous factors in their analyses. The former aims to overcome the shortfalls of Huisman and Mahieu's model and this is where MRS models of Type II (with independent regimes) are born. The first of the independent regime models is presented in Huisman and de Jong [52] where a two-regime model is proposed. Their paper models base prices by an AR(1) regime and captures spikes by a Gaussian distribution. They apply their model to the Dutch APX market, compare it to Huisman and Mahieu's dependent regime model [53] and a simple AR(1) model. It is unclear how they fit the model to data in this paper. They state that the Kalman filter is used to get a soft classification of states in the model, and this soft classification is for weighting the likelihood function. However, it is not obvious how they apply Kalman filter methodology to MRS models with independent regimes.

Since then, there has been a plethora of literature applying MRS models with independent regimes to electricity prices, each paper contributing novel aspects to the area. Of note, Weron *et al.* [103] propose a two-regime model with log-normally distributed spikes and Bierbrauer *et al.* [17, 18] propose two-regime models with Pareto spikes and exponential spikes, respectively. All three papers model log-prices and conclude that

log-normal spikes are best. Weron [101] challenges the idea of modelling log-prices and concludes that modelling raw prices can be advantageous. Janczura and Weron [58] introduce *heteroskedastic variance* structures to within-regime dynamics by modelling base prices with an autoregressive constant elasticity of variance (CEV) model

$$B_t = \alpha + \phi B_{t-1} + \sigma |B_{t-1}|^{\gamma} \varepsilon_t,$$

where α, ϕ, σ and γ are parameters, and $\{\varepsilon_t\}$ is a sequence of i.i.d. N(0,1) random variables. Janczura and Weron [58] also introduce *shifted* regimes into the literature. A shifted regime is a regime that can only capture prices above (or below) a specified level. For example, a shifted log-normal distribution, with shifting parameter q, can only capture prices above the level q. Janczura and Weron [58] motivate shifted regimes by citing that shifting is necessary for the calibration procedure to correctly separate spikes (and drops) from 'normal' price behaviour. More recently, Regland and Lindström [87] introduce Gamma distributed spikes to the MRS modelling literature.

Until 2010 it had gone unnoted that estimation for independent regime models was underdeveloped. Up until that time, papers did not properly detail how they fit their model to data and it is unclear if the methods they use are theoretically valid. In 2012 two papers addressing this issue appeared. One paper by Janczura and Weron [60] develops an approximate maximum likelihood method which we explore in more detail in Sections 2.3.3 and 3.1. The other, by Regland and Lindström [87], compares the algorithm of Janczura and Weron [60] to another maximum likelihood method and a Bayesian method utilising an MCMC algorithm. However, they neither detail their new maximum likelihood method, nor mention any philosophical difference between maximum likelihood and Bayesian methods. It is the aim of this thesis to develop new algorithms for exact maximum likelihood and Bayesian inference for these models.

In parallel, the literature examining exogenous factors affecting electricity markets using MRS models with dependent regimes was evolving. The first of these papers is Mount *et al.* [77] who use a two-regime model of Type I, where both regimes are AR(1) processes and include exogenous predictors. They also include exogenous factors in the switching probabilities of the hidden Markov chain via a logistic transform of a linear combination of exogenous variables. They use data from the PJM market in the United States, and show that reserve margin and load can be used to predict mean prices and regime switches. Huisman [51] notes that reserve margin is not often available to the market and uses temperature as a proxy. His analysis is similar to Mount *et al.* [77] and concludes that temperature is a significant predictor of spikes, but notes that temperature does not provide as much information as the reserve margin. Becker *et al.* [13]

model Queensland's electricity prices using a two-regime model where within-regime dynamics are modelled as independent, scaled, beta random variables that dependend on exogenous factors. Becker *et al.* [13] agree with Huisman [51] concluding that weather can have a significant impact on prices. Karakatsani and Bunn [66] also make use of MRS models with dependent regimes when analysing the impact of fundamentals on electricity prices. More recently, the PhD thesis of Knapik [68] uses MRS models with dependent regimes to examine the effects of load, water reservoir level, temperature and wind on prices and switching probabilities.

These two avenues of literature were then recombined by Janczura and Weron [59] when they produced a review article of MRS models for electricity prices and extended this literature by fitting a MRS model with independent regimes and time-varying parameters to electricity prices. They fit their models using their approximate algorithm and then, post hoc, use kernel smoothing methods to estimate transition probabilities with seasonal fluctuations. Another link between the two streams of literature came via the Masters thesis of Norén [80], which looks at MRS models with independent regimes and transition probabilities that depend on exogenous variables. Norén employs the approximation of the EM algorithm developed by Janczura and Weron to estimate model parameters.

2.3.3 Estimation of MRS models for electricity prices

Due to the specification of MRS models, the distribution of each observation depends the hidden regime sequence, and therefore the likelihood is written as a marginal distribution. Let $\boldsymbol{x} = (x_0, ..., x_T)$ be a sequence of observed prices, and $\boldsymbol{R} = (R_0, ..., R_T)$ be a sequence of unobserved regimes where each R_t lives on the state space $\mathcal{S} = \{1, ..., M\}$. Then, the likelihood is the marginal distribution

$$L(\boldsymbol{\theta}) := f^{\boldsymbol{\theta}}_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{\boldsymbol{R}\in\mathcal{R}} f^{\boldsymbol{\theta}}_{\boldsymbol{X},\boldsymbol{R}}(\boldsymbol{x}, \boldsymbol{R}) = \sum_{\boldsymbol{R}\in\mathcal{R}} f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}|\boldsymbol{R}) f^{\boldsymbol{\theta}}_{\boldsymbol{R}}(\boldsymbol{R}), \tag{2.5}$$

where $\mathcal{R}$ is the space of all possible regime sequences of length T[1]. The number of sequences in $\mathcal{R}$ is M^T. For most realistic datasets, it is computationally infeasible to enumerate all M^T sequences and the sum in (2.5) cannot be calculated as it is presented. The same problem arises for hidden Markov models (HMMs), which can be seen as simplifications of MRS models where the observations only take discrete values and are independently distributed, given the hidden regime. In the context of HMMs, the sum (2.5) is made computationally feasible by the forward algorithm [11], while maximisation

[1]Note that in this section we use the notation $f^{\boldsymbol{\theta}}_{X|Y}(x|y)$ for the conditional density of X given Y evaluated at the point x, y, and with parameters $\boldsymbol{\theta}$. This notation is necessary to avoid ambiguity.

of the likelihood is commonly performed via the Baum-Welch algorithm [11], which is a specific case of the EM algorithm [29] and uses the backward algorithm [11].

Hamilton [44, 45] extends the methods for HMMs to MRS models with dependent regimes by extending the forward algorithm, developing a new algorithm to replace the backward algorithm and applying the EM algorithm. Kim [67] refines the work of Hamilton, developing a more efficient implementation of Hamilton's algorithm which mimics the backward algorithm for HMMs. In this section we review the algorithms of Hamilton and Kim which are used for inference of MRS model of Type I (dependent regime models). As we shall see, these methods are extended by Janczura and Weron [60] to develop an approximate algorithm for MRS models of Type II (with independent regimes), and are related to our own methods (Chapter 3).

Likelihood evaluation for dependent-regime models: The forward algorithm

Define $\boldsymbol{x}_{r:s} = (x_r, x_{r+1}, ..., x_s)$ for $r \leq s$, then, observe that the definition of conditional densities can be used to write the likelihood as

$$L(\boldsymbol{\theta}) := f^{\boldsymbol{\theta}}_{\boldsymbol{X}}(\boldsymbol{x}) = f^{\boldsymbol{\theta}}_{X_0}(x_0) \prod_{t=1}^{T} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{x}_{0:t-1}).$$

The forward algorithm calculates $f^{\boldsymbol{\theta}}_{X_0}(x_0)$ and $f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{x}_{0:t-1})$ for $t = 1, 2, ..., T$, from which it is straightforward to calculate the likelihood or log-likelihood.

Recall, $\{R_t\}_{t\in\mathbb{N}}$ is a Markov chain that lives on the state space $\mathcal{S} = \{1, ..., M\}$ and has transition probabilities p_{ij}, $i, j \in \mathcal{S}$. The forward algorithm is initialised with probabilities $\mathbb{P}^{\boldsymbol{\theta}}(R_0 = i) = \pi_i$; π_i is commonly taken to be the stationary probabilities of the hidden Markov chain $\{R_t\}_{t\in\mathbb{N}}$. Using the law of total probability and the definition of conditional probability, the first term, $f^{\boldsymbol{\theta}}_{X_0}(x_0)$, is calculated as

$$\begin{aligned} f^{\boldsymbol{\theta}}_{X_0}(x_0) &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_0,R_0}(x_0, i) \\ &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|i)\mathbb{P}^{\boldsymbol{\theta}}(R_0 = i) \\ &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|i)\pi_i. \end{aligned}$$

The density $f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|i)$ is known from the model specification. Similar arguments can be used to construct a recursive procedure to calculate $f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{x}_{0:t-1})$ for

$t = 1, ..., T$:

$$\begin{aligned} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{x}_{0:t-1}) &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_t,R_t|\boldsymbol{X}_{0:t-1}}(x_t, i|\boldsymbol{x}_{0:t-1}) \\ &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i, \boldsymbol{x}_{0:t-1})\mathbb{P}^{\boldsymbol{\theta}}(R_t = i|\boldsymbol{x}_{0:t-1}) \\ &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i, \boldsymbol{x}_{0:t-1}) \\ &\quad \times \sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(R_t = i|R_{t-1} = j, \boldsymbol{x}_{0:t-1})\mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = j|\boldsymbol{x}_{0:t-1}) \\ &= \sum_{i\in\mathcal{S}} f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i, \boldsymbol{x}_{0:t-1}) \sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = j|\boldsymbol{x}_{0:t-1})p_{ji}, \end{aligned}$$

where $f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i, \boldsymbol{x}_{0:t-1})$ is also known from the model specification. The probabilities $\mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = j|\boldsymbol{x}_{0:t-1})$, can be calculated using Bayes' Theorem,

$$\mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = j|\boldsymbol{x}_{0:t-1}) = \frac{f^{\boldsymbol{\theta}}_{X_{t-1}|R_{t-1},\boldsymbol{X}_{0:t-2}}(x_{t-1}|j, \boldsymbol{x}_{0:t-2})\mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = j|\boldsymbol{x}_{0:t-2})}{f^{\boldsymbol{\theta}}_{X_{t-1}|\boldsymbol{X}_{0:t-2}}(x_{t-1}|\boldsymbol{x}_{0:t-2})}, \tag{2.6}$$

and are known as the *forward probabilities*. Required in Equation (2.6) are the *prediction probabilities*,

$$\mathbb{P}^{\boldsymbol{\theta}}(R_{t-1} = i|\boldsymbol{x}_{0:t-2}) = \sum_{j\in\mathcal{S}} p_{ji}\mathbb{P}^{\boldsymbol{\theta}}(R_{t-2} = j|\boldsymbol{x}_{0:t-2}). \tag{2.7}$$

In some applications, the forward and prediction probabilities may be quantities of interest [57], or used for model checking [62] and also in the backward algorithm.

Maximum likelihood for MRS models with dependent regimes using the EM algorithm

Hamilton's forward algorithm is a computationally feasible way to evaluate the log-likelihood, from which it is possible to use 'black-box' optimisation methods to find the MLEs. However, it is common to use the EM algorithm [29, 76] instead, particularly when the 'E-step' and 'M-steps' of the algorithm are available in closed form. The EM algorithm for MRS models with dependent regimes [45] can be seen as an extension of the Baum-Welch algorithm for hidden Markov models [9–11]. The EM algorithm, as described in Section 2.2.3, proceeds by iterating between constructing functions $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$, and then maximising Q with respect to $\boldsymbol{\theta} \in \Theta$, where Θ is the parameter space, which results in a sequence $\{\boldsymbol{\theta}_n\}_{n\in\mathbb{N}}$ that converges to local maximisers of the likelihood. The

function Q is constructed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) = \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}, \boldsymbol{R})|\boldsymbol{x}; \boldsymbol{\theta}_n\right].$$

The EM algorithm for MRS models with dependent regimes proceeds as follows [45]. Define η_{ij} as the number of transitions from state i to state j in the sequence $\boldsymbol{R} = (R_0, ..., R_T)$ and let $\mathbb{I}(\cdot)$ be the indicator function. The joint log-density of $\boldsymbol{x}$ and $\boldsymbol{R}$ can be written as

$$\begin{aligned}
&\log f^{\boldsymbol{\theta}}_{\boldsymbol{X},\boldsymbol{R}}(\boldsymbol{x}, \boldsymbol{R}) \\
&= \log f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}|\boldsymbol{R}) + \log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{R}) \\
&= \log f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|R_0) + \sum_{t=1}^{T} \log f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|R_t, \boldsymbol{x}_{0:t-1}) + \log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{R}) \\
&= \sum_{j\in\mathcal{S}} \log\{f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j)\}^{\mathbb{I}(R_0=j)} + \sum_{t=1}^{T}\sum_{j\in\mathcal{S}} \log\{f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|j, \boldsymbol{x}_{0:t-1})\}^{\mathbb{I}(R_t=j)} \\
&\quad + \sum_{i,j\in\mathcal{S}} \log p_{ij}^{\eta_{ij}} + \sum_{i\in\mathcal{S}} \log \pi_j^{\mathbb{I}(R_0=j)} \\
&= \sum_{j\in\mathcal{S}} \mathbb{I}(R_0 = j) \log\{f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j)\} + \sum_{t=1}^{T}\sum_{j\in\mathcal{S}} \mathbb{I}(R_t = j) \log\{f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|j, \boldsymbol{x}_{0:t-1})\} \\
&\quad + \sum_{i,j\in\mathcal{S}} \eta_{ij} \log p_{ij} + \sum_{i\in\mathcal{S}} \mathbb{I}(R_0 = j) \log \pi_j.
\end{aligned}$$

Taking the expectation given parameters $\boldsymbol{\theta}_n$ and observed values $\boldsymbol{x}_{0:T}$ (i.e. with respect to $\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{R}|\boldsymbol{x}_{0:T})$) yields

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) = &\sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = j|\boldsymbol{x}_{0:T}) \log\{f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j)\} \\
&+ \sum_{t=1}^{T}\sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T}) \log\{f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|j, \boldsymbol{x}_{0:t-1})\} \\
&+ \sum_{i,j\in\mathcal{S}} \mathbb{E}[\eta_{ij}|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n] \log p_{ij} + \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = j|\boldsymbol{x}_{0:T}) \log \pi_j.
\end{aligned}$$

The densities $f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j)$ and $f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|j, \boldsymbol{x}_{0:t-1})$, $j \in \mathcal{S}$, $t = 1, ..., T$ are given by the model specification. For example, if Regime i is a Gaussian AR(1) regime then

$$f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i, \boldsymbol{x}_{0:t-1}) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x_t-\alpha_i-\phi_i x_{t-1})^2/(2\sigma_i^2)}. \tag{2.8}$$

The *smoothed probabilities*, also known as *smoothed inferences*, $\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})$, required to construct Q are obtained using a backward recursion after running the forward

algorithm with parameters $\boldsymbol{\theta}_n$, and storing the forward and prediction probabilities. This backward recursion was developed by Kim [67] and is derived as follows. First, note that $\mathbb{P}^{\boldsymbol{\theta}_n}(R_T = j|\boldsymbol{x}_{0:T})$ is already given by the last iteration of the forward algorithm. For $t = T-1, ..., 0$,

$$\begin{aligned}
\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T}) &= \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j, R_{t+1} = i|\boldsymbol{x}_{0:T}) \\
&= \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|R_{t+1} = i, \boldsymbol{x}_{0:T})\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:T}) \\
&= \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|R_{t+1} = i, \boldsymbol{x}_{0:t})\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:T}) \\
&= \sum_{i\in\mathcal{S}} \frac{\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|R_t = j, \boldsymbol{x}_{0:t})\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:t})}{\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:t})}\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:T}) \\
&= \sum_{i\in\mathcal{S}} p_{ji}^{(n)} \frac{\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:t})\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:T})}{\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:t})},
\end{aligned}$$

where $p_{ij}^{(n)}$ means the p_{ij} parameter under $\boldsymbol{\theta}_n$. Here the third equality holds since, given R_{t+1}, R_t is independent of $\boldsymbol{x}_{t+1:T}$ [55]. Intuitively we can think of this as $\boldsymbol{x}_{t+1:T}$ giving us no more information than R_{t+1} provides about R_t.

In the case that $R_t = j \in \mathcal{S}_{AR}$, the process is defined by $X_t = \alpha_j + \phi_j X_{t-1} + \sigma_j \varepsilon_t$, where α_j, ϕ_j and σ_j are parameters, and $\{\varepsilon_t\}$ is a sequence of i.i.d. N(0,1) random variables. If we assume

$$\log f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j) = g(x_0),$$

and g does not vary with parameters $\boldsymbol{\theta}$, then we can express the maximiser of Q at the $(n+1)^{\text{th}}$ iteration of the EM algorithm, $\boldsymbol{\theta}^{n+1}$, for $n \geq 1$, as the following system of equations [60].

$$\begin{aligned}
\phi_j^{(n+1)} &= \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})x_{t-1}B_{1,t}}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})x_{t-1}B_{2,t}}, \\
\alpha_j^{(n+1)} &= \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - \phi_j^{(n+1)}x_{t-1})}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})}, \\
(\sigma_j^2)^{(n+1)} &= \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - \alpha_j^{(n+1)} - \phi_j^{(n+1)}x_{t-1})^2}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})},
\end{aligned}$$

where

$$B_{1,t} = x_t - x_{t-1} - \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - x_{t-1})}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})},$$

$$B_{2,t} = \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})x_{t-1}}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})} - x_{t-1}.$$

In general, the switching probabilities are updated using the following [67]

$$p_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T}) \frac{p_{ij}^{(n)} \mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1} = i|\boldsymbol{x}_{0:t-1})}{\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:t-1})}}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1} = i|\boldsymbol{x}_{0:T})}, \tag{2.9}$$

which rely on the smoothed, forward and prediction probabilities. Recall that we may specify the π_j's as the stationary distribution of the Markov chain $\{R_t\}$, as parameters to be determined or some fixed distribution. However, note that the derivation of Equation (2.9) implicitly assumes the terms π_j are unrelated to the parameters p_{ij}, which is not the case if the stationary distribution of $\{R_t\}$ is used to specify the initial distribution of the chain. Nonetheless, we expect this issue to have a vanishing affect on the MLEs as the sample size grows, since the dependence of the likelihood on the initial distribution will be overwhelmed by other terms in the likelihood (provided, of course, that the Markov chain $\{R_t\}$ is ergodic).

Thus, we implement the EM algorithm by initialising it with a guess of the true parameters, then alternating between the forward and backward algorithms and calculating the maximisers of Q. The algorithm terminates when the step size is below a prespecified tolerance, i.e. $|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n|_\infty < \varepsilon$ where ε is some small tolerance.

Estimation of MRS models with independent regimes

The EM-like algorithm Currently the approximate method developed by Janczura and Weron [60] is used for inference for MRS models of Type II. We label this algorithm the 'EM-like' algorithm because of its resemblance to the EM algorithm; however, this is not the EM algorithm and no theory surrounding convergence is available. Simulations show that, for some cases, the EM-like algorithm can perform well [60]; however, examples can be constructed where the EM-like algorithm fails to recover the true parameters. We present and analyse the EM-like method in detail in Section 3.1. Related

to this, Regland and Lindström [87] examine the EM-like algorithm for MRS models with independent regimes, and conclude that it works well compared to MCMC and other maximum likelihood methods for their datasets. However, they detail neither their implementation of the EM algorithm, nor their method for likelihood evaluation, for which no computationally feasible method has been presented yet.

Monte Carlo Expectation Maximisation A tractable but approximate alternative to the EM and EM-like algorithms is Monte Carlo Expectation Maximisation (MCEM) [100], which has not been used in the electricity pricing literature to date. We do not investigate the MCEM algorithm for electricity price models further, since we develop exact likelihood methods in Sections 3.2-3.4, rather we simply mention MCEM for completeness. The MCEM algorithm has the same recipe as the EM algorithm, except the E-step is replaced by a Monte Carlo approximation. The idea is, if the hidden regime sequence can be sampled from the distribution $f^{\boldsymbol{\theta}_n}(\boldsymbol{R}|\boldsymbol{x})$, then the sum in the E-step can be approximated as

$$\begin{aligned}
&\mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X},\boldsymbol{R}}(\boldsymbol{x},\boldsymbol{R})|\boldsymbol{x};\boldsymbol{\theta}_n\right] \\
&= \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}|\boldsymbol{R})|\boldsymbol{x};\boldsymbol{\theta}_n\right] + \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{R}}(\boldsymbol{R})|\boldsymbol{x};\boldsymbol{\theta}_n\right] \\
&= \sum_{\boldsymbol{R}\in\mathcal{R}} \log\{f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}|\boldsymbol{R})\} f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}|\boldsymbol{x}) + \sum_{\boldsymbol{R}\in\mathcal{R}} \log\{f^{\boldsymbol{\theta}}_{\boldsymbol{R}}(\boldsymbol{R})\} f^{\boldsymbol{\theta_n}}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}|\boldsymbol{x}) \\
&\approx \sum_{j=1}^{J} \frac{1}{J} \log\{f^{\boldsymbol{\theta}}_{\boldsymbol{X}|\boldsymbol{R}}(\boldsymbol{x}|\boldsymbol{R}^*_j)\} f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}^*_j|\boldsymbol{x}) + \sum_{j=1}^{J} \frac{1}{J} \log\{f^{\boldsymbol{\theta}}_{\boldsymbol{R}}(\boldsymbol{R}^*_j)\} f^{\boldsymbol{\theta_n}}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}^*_j|\boldsymbol{x})
\end{aligned}$$

where $\boldsymbol{R}^*_j$ for $j = 1, ..., J$ are samples from the $f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}|\boldsymbol{x})$. Obviously J is chosen such that $J \ll 2^T$ to make the problem feasible. The strong law of large numbers ensures that, as $J \to \infty$ the approximation converges to the true value. Due to the Monte Carlo error introduced by this approximation, the monotonicity property of the EM algorithm is lost, however, it has been shown that the algorithm gets close to a maximiser with a high probability [19] for some cases.

Typical methods for sampling from the distribution, $f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}|\boldsymbol{x})$, are *rejection sampling* or *MCMC* algorithms. Sampling from $f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}}(\boldsymbol{R}|\boldsymbol{x})$ via MCMC algorithms is discussed in Section 4.2.

Bayesian methods An alternative parameter inference paradigm is *Bayesian inference*, where parameters are treated as random variables rather than unknown constants. The goal of Bayesian inference is to infer the distribution of the parameters given observed data, which is known as the *posterior distribution*. Section 2.2.4 provides a brief

introduction to Bayesian inference. For MRS models, the posterior distribution is not analytically available and numerical methods are needed to approximate it. Typically Markov Chain Monte Carlo algorithms are used (introduced in Section 2.2.5 and explored in Chapter 4). To our knowledge, there is just one paper that mentions Bayesian methods for MRS models with independent regimes in existing literature: Regland and Lindström [87] which briefly compares maximum likelihood and Bayesian methods. They simulate a single realisation of length 5000 of a 3-regime MRS model with independent regimes, recover point estimates of the parameters in a Bayesian framework using an MCMC algorithm, and compare these to likelihood-based estimates. They use their MCMC algorithm to generate a single MCMC chain of length 20,000, the first 20% of which is discarded as burn-in, and they use the rest of the samples to calculate posterior inferences. They do not detail which point estimates they report from the posterior distribution. They conclude that likelihood-based inferences compare favourably to Bayesian-based inferences, while more MCMC iterations are needed for a more accurate representation of the posterior distribution.

2.3.4 Detrending methods

As mentioned earlier, models of electricity prices are typically built out of two parts, a deterministic trend component, T_t and a stochastic component X_t. This section reviews literature focusing on the trend component T_t. The trend component is typically broken up into two parts, a periodic short-term seasonal component (STSC), s_t, which can capture weekly periodic behaviour, and a long-term component (LTC), ℓ_t, which can capture mean price movements over periods of months and years. In electricity price modelling literature, typical models for the LTC are:

- piecewise-constant functions and linear trends, e.g. [49] where the time series is projected onto piecewise-constant functions,
- superpositions of sinusoidal functions, e.g. [16, 33, 38, 42], where the time series is projected onto (typically one, two or three) sinusoidal functions of varying frequencies,
- wavelets, e.g. [59, 60, 101], where prices are recursively projected onto wavelet functions at progressively coarser scales,
- smoothing splines [71], where prices are projected onto basis functions, commonly piecewise cubics (smoothing splines are less common in the electricity pricing literature),
- utilising the forward price, [15, 80], where the trend price is the forward price.

The STSC is commonly modelled using piecewise-constant functions, as in De Jong [27].

There are numerous papers reviewing different models for the LTC and their effectiveness [57, 71, 82, 83]. Lisi and Nan [71] compare a range of models, including those mentioned above, to log-price data from the Pennsylvania-New Jersey-Maryland (PJM), Nord-Pool (NP) and Amsterdam Power Exchange (APX) markets. The comparison metrics they use focus on three aspects: (i) the residuals are stationary (that is, when the trend is removed from the data, the resultant series is stationary), (ii) the residuals are not periodic, (iii) out-of-sample predictions must be improved by the method.

To test (i) two models are fitted to the residuals, one with an extra seasonal component (modelled using a regression spline) and one without the extra seasonal component. The models are then compared to see if there is a significant difference between them and if there is no statistically significant difference then condition (i) is passed. Point (ii) is tested similarly: two models are fitted to the residuals, one with an extra term capturing weekly dependence; the fitted models are then compared, and if the extra term capturing weekly dependence is insignificant then the model passes test (ii). To test (iii) a regression model is fitted to the residuals (this regression model includes exogenous factors, demand and margin) and predictions made. The forecast performance is measured by evaluating the out-of-sample mean-squared error and mean absolute error between the prediction and the observed test data. These tests are all conducted across 24 different load periods for the APX and NP markets and over 48 load periods for the PJM market. They conclude that the best LTC model uses smoothing splines, while the best STSC model uses piecewise-constant functions and is estimated by trimmed means to avoid bias from extreme prices. Nowotarski *et al.* [81] briefly note that spline forecasting methods can perform poorly for longer range predictions.

Nowotarski *et al.* [82] compare 300 different models for the long-term trend component. The models they consider are based on either piecewise-constant functions, superpositions of sinusoids or wavelet methods. They fit all 300 models to data from New South Wales, European Energy Exchange (EEX), Nord-Pool (NP), New England Pool (NEP), New York Independent System Operator (NYISO) and Pennsylvania-New Jersey-Maryland (PJM) markets, and compare their predictive performance over a range of forecast time horizons. They conclude that wavelet-based methods are superior to sine-based methods for forecasting up to one year ahead as measured by three out-of-sample error measures (mean absolute error, mean squared error and mean absolute percentage error). They also find that an LTC model based on the Coiflets wavelet of order 4 is best. This specification also includes a linear decay to median prices, is calibrated to a three-year window, using the wavelet filter recursively 6 times, and is estimated after removing extreme values from the dataset and replacing them with the

mean deseasonalised price. They note there is not much difference between the methods using wavelets and wavelet-based methods outperform the rest. In their concluding remarks they also mention alternative models they have not considered. They agree with Stevenson *et al.* [95] in that using forward prices may not be wise, since they can be misleading, particularly in illiquid forwards markets.

Nowotarski and Weron [83] reach a similar conclusion regarding the superiority of wavelet models when they investigate the importance of modelling the long-term seasonal component of electricity prices for use in day-ahead forecasting. This study differs from [82] since the datasets used are different – hourly prices are modelled rather than daily average prices – and they also consider the Hodrick Prescott filter. They conclude that wavelet-base LTC methods are best as measured by weekly-weighted mean absolute error (the mean absolute error normalised by dividing by the mean weekly price), and that the Hodrick Prescott filter performs poorly.

Janczura *et al.* [57] investigate the effect of extreme prices on estimating the seasonal component from electricity price series. They explore a range of different definitions of spikes, ranging from fixed price thresholds, where prices over a certain level are classified as spikes; to using a recursive filter where prices are recursively removed from the data if they differ from the mean by 3 standard deviations or more, and the mean and standard deviation are recalculated each time a spike is removed; to classifying prices as spikes if they have over a 0.5 posterior probability of belonging to an extreme price regime of an MRS model fitted to the data. They remove extreme prices from the data using the different definitions and replace them with the mean of the detrended data, and then estimate the trend components on this altered dataset. The LTC model that they choose is based on the Daubechies wavelet family, and they estimate an S_6 approximation of the data. They compare the methods by simulating datasets using MRS models, applying these detrending methods and comparing the true trend to the estimated trend. They also re-fit a stochastic MRS model to the residuals and observe how close the parameters of the re-estimated MRS model are to the true parameters. They conclude that the classification of prices using MRS models is best in terms of recovering the parameters of the simulated data. In this paper, extreme prices were replaced by the mean of the deseasonalised data. However, there are many other methods proposed to replace extreme values, e.g. by a neighbouring point [38], by the mean of the two neighbouring points, or 'similar day' values [17] for example. Trueck *et al.* [96] explore some of these alternatives and suggest further work is needed to determine which is best.

Chapter 3

Likelihood methods for MRS models with independent regimes

For MRS models of Type I (with dependent regimes), the densities

$$f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i,\boldsymbol{x}_{0:t-1})$$

are simple and given directly by the model specification (e.g. Equation (2.8)), and the forward algorithm of Hamilton [44, 45] presented in Section 2.3.3 is an effective tool[1]. The same forward algorithm is theoretically applicable to MRS models with independent regimes (Type II and Type III models); however, it is not computationally feasible, even for relatively small datasets, since more information about the hidden sequence is needed to specify the conditional densities above. Specifically, the missing information required is the time of the last observations from each regime.

For MRS models of Type II there is the added complexity of unobserved values of the within-regime processes. Figure 3.1 illustrates this strange dependence structure for MRS models of Type II. For MRS models of Type III, the problem is slightly simpler since there are no unobserved values of the within-regime processes. Theoretically, the densities $f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i,\boldsymbol{x}_{0:t-1})$ can be determined but this is an $\mathcal{O}(M^t)$ operation, where M is the number of regimes, but this is not computationally feasible for practical values of M or T, where $T+1$ is the length of the observed price sequence, $\boldsymbol{x}_{0:T}$. This is the main challenge to overcome when evaluating the likelihood for MRS models with independent regimes.

[1] In this chapter we use the notation $f^{\boldsymbol{\theta}}_{X|Y}(x|y)$ for the conditional distribution of X given Y under the parameters $\boldsymbol{\theta}$, since it is more descriptive and necessary.

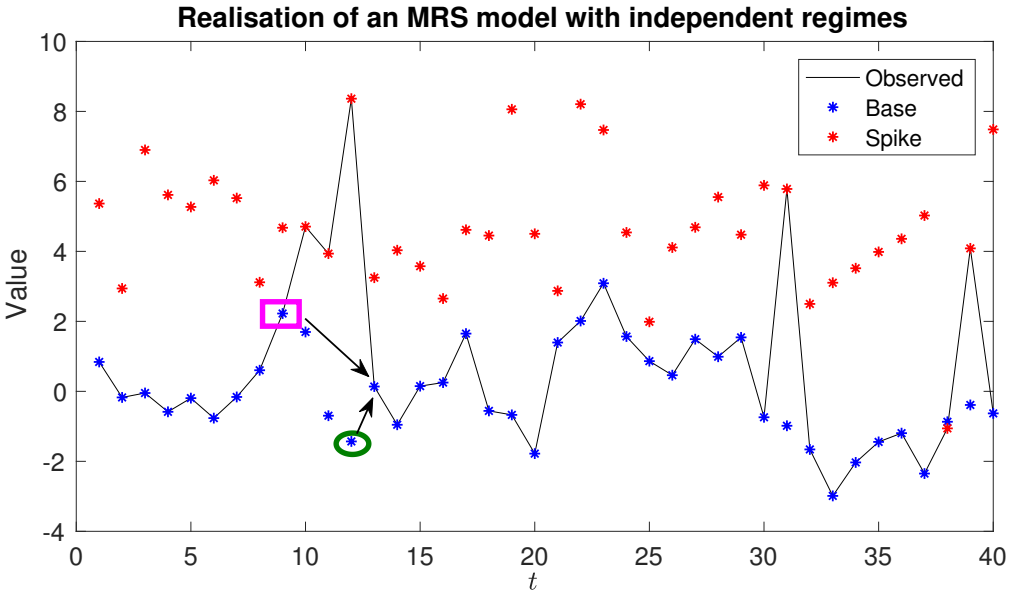


FIGURE 3.1: Simulation of an MRS model of Type II with two independent regimes; an AR(1) base regime and an i.i.d. spike regime. To write down the distribution of the observation at time $t = 13$, we need to know either the observation immediately before it, circled in green, or the time that the last observed price from that regime occurred, highlighted by the pink box. The value circled in green is unobserved and hence unknown. And since the regime sequence is unobserved we do not actually know the last time the process was in the AR(1) regime, i.e. we do not know which time point to put the pink box around.

This chapter details exact and approximate likelihood evaluation and maximisation techniques for MRS models with independent regimes, specifically focusing on models for electricity prices. That is, we restrict our attention to models with either AR(1) or i.i.d regimes. However, we believe our methods can be extended to more general models.

First, in Section 3.1, we examine the approximate algorithm of Janczura and Weron [60] who extend the works of Hamilton [44, 45] and Kim [67] and develop an ad-hoc algorithm for inference for MRS models of Type II. Their algorithm is motivated by, and similar to, the EM algorithm, so we title it the *EM-like algorithm*. However, the convergence results of the EM algorithm do not carry over to the EM-like algorithm, and there is currently no guarantee the algorithm will converge to the true parameters or MLEs. Simulations have shown that the EM-like algorithm can produce reasonable results [60], however it is easy to construct examples where the EM-like algorithm fails, as we do in Section 3.1.

We then present our own exact and computationally feasible algorithms for MRS models of Types II and III. In Section 3.2 we present a forward algorithm for likelihood evaluation, then in Section 3.3 we build on Section 3.2 and develop a backward algorithm to

calculate smoothed inferences which enables us to construct the EM algorithm presented in Section 3.4. We discuss our methods in Section 3.5.

3.1 The 'EM-like' algorithm

In the existing literature, there is currently no computationally feasible way to implement the EM algorithm for independent regime models [28, 60]. In particular, there is no efficient way to compute the E-step exactly (except when the within-regime processes are all i.i.d. in which case we have a hidden Markov model with continuous observation distributions). An efficient algorithm to approximately infer the parameters of MRS models of Type II, inspired by the EM algorithm of Hamilton [45], was developed Janczura and Weron [60] which we label the EM-like algorithm. This is *not* the EM algorithm and none of the EM theory holds.

For MRS models of Type II, the EM-like algorithm overcomes the problem of computational infeasibility by replacing lagged values for Regime i with sensible values, $\tilde{b}_{t-1,i}^{(n)}$ which are described as expectations [60],

$$\mathbb{E}[B_t^i|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n], \tag{3.1}$$

where $B_t^i := \mathbb{I}(R_t = i)x_t + \mathbb{I}(R_t \neq i)\left(\alpha_i + \phi_i B_{t-1}^i + \sigma_i \varepsilon_t^i\right)$.

Simply put, whenever an observation x_{t-1} appears in an expression related to Regime i in the EM algorithm in Section 2.3.3, it is replaced with $\tilde{b}_{t-1,i}^{(n)}$ at the n^{th} iteration. So, the E-step of the n^{th} iteration of the EM-like algorithm requires calculating the forward probabilities for $t = 1, ..., T$ using the recursion

$$\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_{t-1} = j|\boldsymbol{x}_{0:t-1}) = \frac{f^{\boldsymbol{\theta}_n}_{X_{t-1}|R_{t-1},B_{t-2}^{(j)}}\left(x_{t-1}|j,\tilde{b}_{t-2,j}^{(n)}\right)\sum_{i\in\mathcal{S}} p_{ij}^{(n)}\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_{t-2}=i|\boldsymbol{x}_{0:t-2})}{\sum_{j\in\mathcal{S}} f^{\boldsymbol{\theta}_n}_{X_{t-1}|R_{t-1},B_{t-2}^{(j)}}\left(x_{t-1}|j,\tilde{b}_{t-2,j}^{(n)}\right)\sum_{i\in\mathcal{S}} p_{ij}^{(n)}\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_{t-2}=i|\boldsymbol{x}_{0:t-2})}. \tag{3.2}$$

Assuming Regime i is an AR(1) regime, then the densities $f^{\boldsymbol{\theta}_n}$ are

$$\begin{aligned} &f^{\boldsymbol{\theta}_n}_{X_{t-1}|R_{t-1},B_{t-2}^{(j)}}\left(x_{t-1}|j,\tilde{b}_{t-2,j}^{(n)}\right) \\ &= \frac{1}{\sqrt{2\pi\left(\sigma_i^{(n)}\right)^2}}\exp\left\{-\frac{1}{2\left(\sigma_i^{(n)}\right)^2}\left(x_{t-1}-\alpha_i^{(n)}-\phi_i^{(n)}\tilde{b}_{t-2,j}^{(n)}\right)^2\right\}. \end{aligned}$$

This recursion is initialised with the parameter $\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_0 = i|x_0) = \rho_i^{(n)}$. We include a tilde (˜) over the probabilities of the EM-like algorithm to differentiate them from the smoothed and filtered probabilities calculated using an exact implementation of the EM algorithm. Also part of the E-step, the smoothed probabilities are calculated for $t = T-1, T-2, ..., 0$, using the backward recursion

$$\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T}) = \sum_{i\in\mathcal{S}} p_{ji}^{(n)} \frac{\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:t})\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_{t+1} = i|\boldsymbol{x}_{0:T})}{\sum\limits_{j\in\mathcal{S}} p_{ji}^{(n)}\tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:t})}.$$

The M-step of the EM-like algorithm for AR(1) regimes is,

$$\phi_j^{(n+1)} = \frac{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})\tilde{b}_{t-1,j}^{(n)} B_{1,t}}{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})\tilde{b}_{t-1,j}^{(n)} B_{2,t}},$$

$$\alpha_j^{(n+1)} = \frac{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - \phi_j^{(n+1)}\tilde{b}_{t-1,j}^{(n)})}{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})},$$

$$B_{1,t} = x_t - \tilde{b}_{t-1,j}^{(n)} - \frac{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - \tilde{b}_{t-1,j}^{(n)})}{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})},$$

$$B_{2,t} = \frac{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})\tilde{b}_{t-1,j}^{(n)}}{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})} - \tilde{b}_{t-1,j}^{(n)},$$

$$(\sigma_j^2)^{(n+1)} = \frac{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})(x_t - \alpha_j^{(n+1)} - \phi_j^{(n+1)}\tilde{b}_{t-1,j}^{(n)})^2}{\sum\limits_{t=1}^{T} \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T})},$$

and $\rho_i^{(n+1)} = \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_0 = i|\boldsymbol{x}_{0:T})$.

The $\tilde{b}_{t,i}^{(n)}$ are calculated recursively

$$\tilde{b}_{t,i}^{(n)} = \tilde{\mathbb{P}}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:t})x_t + \tilde{\mathbb{P}}_{\boldsymbol{\theta}_n}(R_t \neq i|\boldsymbol{x}_{0:t-1})\left(\alpha_i + \phi_i\tilde{b}_{t-1,i}^{(n)}\right). \tag{3.3}$$

Janczura and Weron [60] conduct simulation studies and show that this algorithm seems to work well for the datasets they generate. However, no theoretical results are available that show convergence of, or error bounds for, the EM-like algorithm and we cannot be sure the parameter estimates produced by the EM-like algorithm are consistent.

This is exemplified in Example 3.1, courtesy of Gary Glonek [41].

Example 3.1. *Suppose $Y_1, Y_2, ..., Y_T$ are i.i.d. $N(\mu, \sigma^2)$ where both parameters are unknown. Assume $Y_1, Y_2, ..., Y_t$ are observed and denote their observations $y_1, y_2, ..., y_t$, respectively, and suppose $Y_{t+1}, ..., Y_T$ are missing.*

The true MLEs are

$$\widehat{\mu} = \frac{1}{t}\sum_{i=1}^{t} y_i \text{ and } \widehat{\sigma}^2 = \frac{1}{t}\sum_{i=1}^{t}(y_i - \widehat{\mu})^2. \tag{3.4}$$

The EM algorithm works by replacing the sufficient statistics for the parameters by their conditional expectations. (Note that the conditioning is trivial in this case because the data are independent.) The EM algorithm requires the log of the joint density, $\log f^{\boldsymbol{\theta}}_{\boldsymbol{Y}_{1:t},\boldsymbol{Y}_{t+1:T}}(\boldsymbol{y}_{1:t}, \boldsymbol{Y}_{t+1:T})$, where $\boldsymbol{\theta} = (\mu, \sigma^2)$. We can write this joint density as

$$\begin{aligned}
&\log f^{\boldsymbol{\theta}}_{\boldsymbol{Y}_{1:t},\boldsymbol{Y}_{t+1:T}}(\boldsymbol{y}_{1:t}, \boldsymbol{Y}_{t+1:T}) \\
&= \log\left(\prod_{i=1}^{t}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}\prod_{j=t+1}^{T}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(Y_j-\mu)^2}\right) \\
&= -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{t} y_i^2 + \sum_{i=t+1}^{T} Y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^{t} y_i + \sum_{i=t+1}^{T} Y_i\right) - \frac{T\mu}{2\sigma^2}.
\end{aligned}$$

The E-step corresponds to replacing the missing data terms, $\sum_{i=t+1}^{T} Y_i$ and $\sum_{i=t+1}^{T} Y_i^2$, by their conditional expectations:

$$\begin{aligned}
\sum_{i=t+1}^{T} Y_i &\leftarrow (T-t)\widehat{\mu}_n, \\
\sum_{i=t+1}^{T} Y_i^2 &\leftarrow (T-t)(\widehat{\mu}_n^2 + \widehat{\sigma}_n^2),
\end{aligned}$$

where $\widehat{\mu}_n$ and $\widehat{\sigma}_n$ are the parameters at the n^{th} iteration of the EM algorithm. The M-step can be seen to be

$$\begin{aligned}
\widehat{\mu}_{n+1} &\leftarrow \frac{1}{T}\{t\widehat{\mu} + (T-t)\widehat{\mu}_n\}, \\
\widehat{\sigma}^2_{n+1} &\leftarrow \frac{1}{T}\{t(\widehat{\sigma}^2 + \widehat{\mu}^2) + (T-t)(\widehat{\sigma}_n^2 + \widehat{\mu}_n^2) - 2\widehat{\mu}_n(t\widehat{\mu} + (T-t)\widehat{\mu}_n) + T\widehat{\mu}_n^2\} \\
&= \frac{1}{T}\{t(\widehat{\sigma}^2 + \widehat{\mu}^2) + (T-t)\widehat{\sigma}_n^2 - 2t\widehat{\mu}_n\widehat{\mu} + t\widehat{\mu}_n^2\},
\end{aligned}$$

where $\widehat{\mu}$ and $\widehat{\sigma}^2$ are the true MLEs in Equation (3.4). It can be shown that the iterations will converge to the true MLEs and that they are the fixed point of the iterations.

Consider now an EM-like algorithm in which the missing observations $y_{t+1}, ..., y_T$ are replaced by their conditional expectations,

$$\mathbb{E}\left[Y_j | y_1, y_2, ..., y_t, \theta_n\right] = \widehat{\mu}_n, \qquad \text{for } j = t+1, ..., T,$$

where $\boldsymbol{\theta}_n = (\widehat{\mu}_n, \widehat{\sigma}_n)$, at the E-step of the iterations. This corresponds to the omission of the term $\widehat{\sigma}_n^2$ in the EM algorithm. In the same notation as for the EM algorithm,

$$\sum_{i=t+1}^{T} Y_i \leftarrow (T-t)\widehat{\mu}_n,$$
$$\sum_{i=t+1}^{T} Y_i^2 \leftarrow (T-t)(\widehat{\mu}_n^2),$$

and the M-step is

$$\widehat{\mu}_{n+1} \leftarrow \frac{1}{T}\{t\widehat{\mu} + (T-t)\widehat{\mu}_n^2\},$$
$$\widehat{\sigma}_{n+1}^2 \leftarrow \frac{1}{T}\{t(\widehat{\sigma}^2 + \widehat{\mu}^2) - 2t\widehat{\mu}_n\mu + t\widehat{\mu}_n^2\}.$$

It can be shown that the iterations will converge to $\widehat{\mu}$ and $\dfrac{t}{T}\widehat{\sigma}^2$, respectively, so this EM-like algorithm fails to converge to the true MLE of σ^2.

This suggests that an EM-like approach might work when the log-joint-density is a linear function of the missing data, but not in general. Moreover, since the log-joint-density for MRS models with Gaussian AR(1) regimes is not linear, then we should not expect the EM-like algorithm to perform well, particularly when estimating variance parameters.

Furthermore, examples of MRS models of Type II where the EM-like algorithm fails to get close to the true parameter values can easily be constructed. The first example, Example 3.2, is a 'hard' problem since it is not obvious which observations belong to which regime, as shown in Figure 3.2.

Example 3.2. *Consider the following MRS model of Type II,*

$$X_t = \begin{cases} B_t, & \text{if } R_t = 1, \\ Y_t, & \text{if } R_t = 2, \end{cases}$$

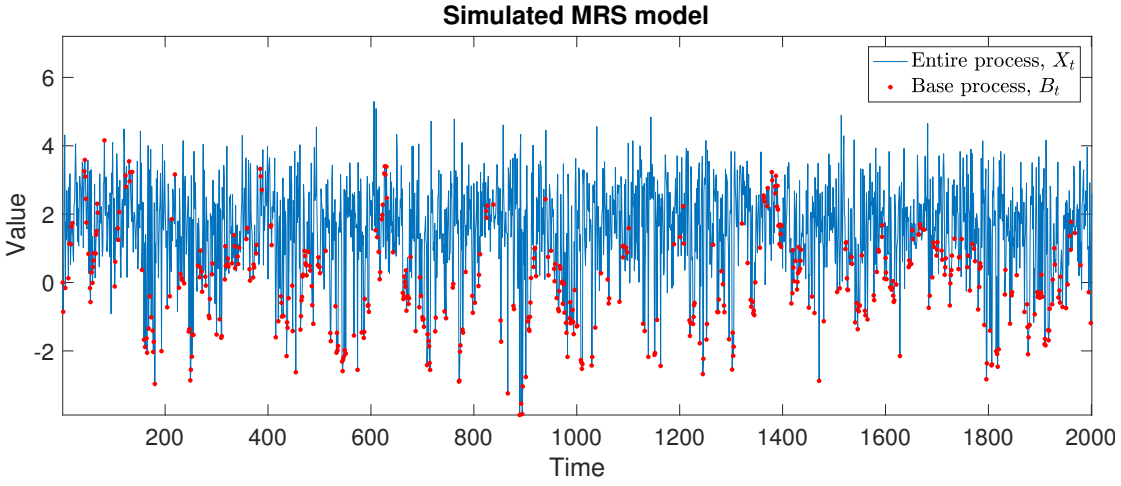


FIGURE 3.2: A plot of a simulated dataset for Example 3.2. The blue line is the entire process and points from Regime 1 are highlighted by red dots. Note that if the observations were not highlighted in red, it would not be obvious which points belong to which regime, this is why we label it a 'hard' problem.

where B_t is an AR(1) process,

$$B_t = 0.95B_{t-1} + 0.2\varepsilon_t,$$

with $\{\varepsilon_t\}$ being a sequence of i.i.d. $N(0,1)$ random variables, Y_t is an i.i.d. sequence of $N(2,1)$ random variables, and $\{R_t\}_{t\in\mathbb{N}}$ is a Markov chain with state space $\mathcal{S} := \{1,2\}$, transition matrix

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix},$$

and initial probability distribution $(1,0)$, so the process always starts in Regime 1. Thus, the true parameter vector is

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\alpha}, \widehat{\phi}, \widehat{\sigma}_1, \widehat{\mu}, \widehat{\sigma}_2, \widehat{p}_{11}, \widehat{p}_{22}\right) = (0, 0.95, 0.2, 2, 1, 0.5, 0.8).$$

We simulated 20 realisations of length $T = 2000$ from this model and used the EM-like algorithm to try to recover the true parameters. Figure 3.2 plots an example of one of these realisations. To give the algorithm the best chance of converging to the true parameters, we initialise the EM-like algorithm at the true parameter values. The parameters recovered by the EM-like algorithm are summarised in Figure 3.3. For comparison, the MLEs obtained using our EM algorithm (Section 3.4) are also shown. Notice, in Figure 3.3, that the EM-like algorithm performs poorly, while our exact method performs much better.

We can also construct 'easier' examples, where the EM-like algorithm fails, such as Example 3.3. This is an 'easier' problem since we can almost eyeball which observations

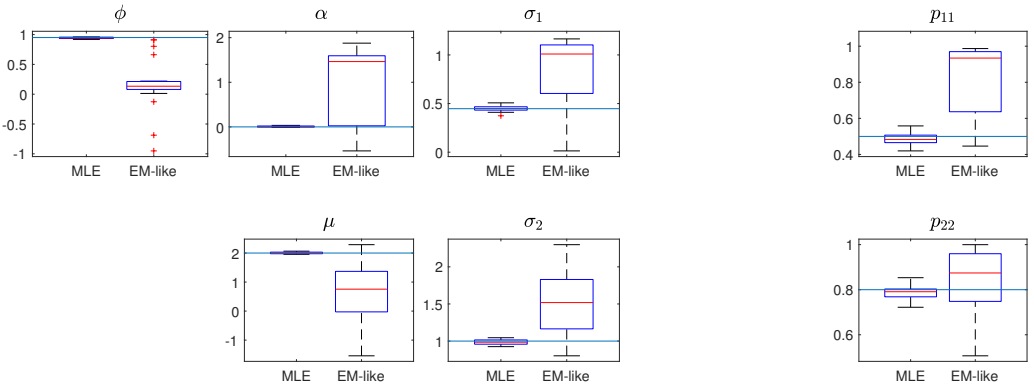


FIGURE 3.3: Boxplots of the parameters recovered by the EM-like algorithm (right) and the MLEs recovered by our EM algorithm (left), for Example 3.2. The blue line represents the true parameter value. Notice that the EM-like algorithm is not able to recover the parameters, while the EM algorithm performs very well.

come from which regime, as shown in Figure 3.4.

Example 3.3. *Consider the following MRS model of Type II,*

$$X_t = \begin{cases} B_t, & \text{if } R_t = 1, \\ Y_t, & \text{if } R_t = 2, \end{cases}$$

where $\{B_t\}$ *is an AR(1) process,*

$$B_t = 0.95B_{t-1} + 0.1\varepsilon_t,$$

with $\{\varepsilon_t\}$ *a sequence of i.i.d.* $N(0,1)$ *random variables,* Y_t *is a i.i.d. sequence of* $N(3,2)$ *random variables, and* $\{R_t\}_{t\in\mathbb{N}}$ *is a Markov chain with state space* $\mathcal{S} := \{1,2\}$*, transition matrix*

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix},$$

and initial probability distribution $(1,0)$*, so the process always starts in Regime 1.*

We simulated 40 realisations of this process were simulated, each of length $T = 2000$*, and the EM-like algorithm used to recover the parameters. For comparison, our EM algorithm was used to obtain the MLEs. One of the simulated datasets is plotted in Figure 3.4. Figure 3.5 summarises the parameter estimates obtained from both algorithms using box plots. Notice the EM-like algorithm struggles to recover the parameters of the i.i.d. regime,* μ *and* σ_2*, and the parameter* p_{22}*.*

We observed in Examples 3.2 and 3.3 the EM-like algorithm was not able to recover the parameters, so the EM-like algorithm cannot be trusted to estimate the true parameters

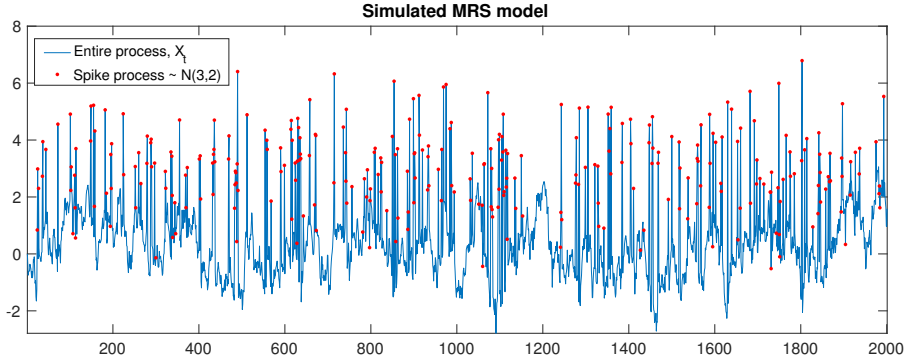


FIGURE 3.4: A dataset simulated from the model in Example 3.3

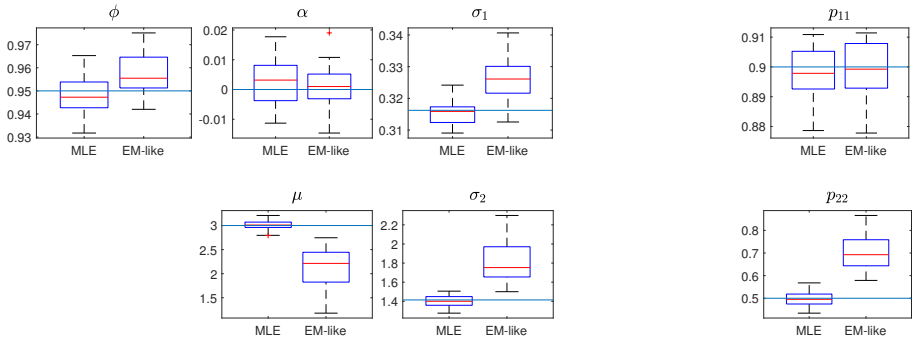


FIGURE 3.5: Box plots of parameter estimates recovered by the EM-like algorithm (right) and the MLEs recovered by the EM algorithm (left), for the model in Example 3.3. Notice the EM-like algorithm struggles to recover the parameters of the i.i.d. regime, μ and σ_2, and the parameter p_{22}.

in a practical problem.

Furthermore, the values $\tilde{b}^{(n)}_{t-1,i}$ are described as the expectation (3.1), however they can be seen to be approximations by the following arguments. First, use linearity of the expectation operator and the definition of B^i_t above to write $\mathbb{E}[B^i_t|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n]$ as

$$\begin{aligned}\mathbb{E}[B^i_t|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] &= \mathbb{E}[\mathbb{I}(R_t=i)x_t + \mathbb{I}(R_t\neq i)\left(\alpha_i+\phi_i B^i_{t-1}+\sigma_i\varepsilon^i_t\right)|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n]\\ &= x_t\mathbb{E}[\mathbb{I}(R_t=i)|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] + \alpha_i\mathbb{E}[\mathbb{I}(R_t\neq i)|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n]\\ &\quad + \phi_i\mathbb{E}[\mathbb{I}(R_t\neq i)B^i_{t-1}|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] + \sigma_i\mathbb{E}[\mathbb{I}(R_t\neq i)\varepsilon^i_t|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n].\end{aligned}$$

As in [60], we have the following

$$\begin{aligned}\mathbb{E}[\mathbb{I}(R_t=i)|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] &= \mathbb{P}^{\boldsymbol{\theta}_n}(R_t=i|\boldsymbol{x}_{0:t}),\\ \mathbb{E}[\mathbb{I}(R_t\neq i)|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] &= \mathbb{P}^{\boldsymbol{\theta}_n}(R_t\neq i|\boldsymbol{x}_{0:t}),\\ \mathbb{E}[\mathbb{I}(R_t\neq i)\varepsilon^i_t|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] &= \mathbb{E}\left[\mathbb{E}[\mathbb{I}(R_t\neq i)\varepsilon^i_t|R_t\neq i,\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n]|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n\right]\\ &= \mathbb{E}\left[\mathbb{I}(R_t\neq i)\mathbb{E}[\varepsilon^i_t|R_t\neq i,\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n]|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n\right]\end{aligned}$$

$$= \mathbb{E}\left[\mathbb{I}(R_t \neq i)0|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n\right]$$
$$= 0,$$

where the last equality holds as $\varepsilon_t^i \sim$ i.i.d. $\mathrm{N}(0,1)$. The last term,

$$\mathbb{E}[\mathbb{I}(R_t \neq i)B_{t-1}^i|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n] = \mathbb{E}\left[\mathbb{E}[\mathbb{I}(R_t \neq i)B_{t-1}^i|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n, R_t \neq i]|\boldsymbol{x}_{0:t};\boldsymbol{\theta}_n\right],$$

cannot be simplified. Since $\{R_t\}_{t\in\mathbb{N}}$ is a Markov chain, knowing $R_t \neq i$ gives us some knowledge of R_{t-1}, and the knowledge of R_{t-1} informs us about the relationship between x_{t-1} and B_{t-1}^i. Thus B_{t-1}^i is not independent of $R_t \neq i$. Also, knowing x_t gives information about which regime x_{t-1} could have come from, so, given x_{t-1}, then B_{t-1}^i is dependent on x_t. If these dependencies are incorrectly ignored, then this term does simplify and we arrive at the expression in Equation (3.3).

3.2 A novel forward algorithm

In this section we develop a computationally feasible forward algorithm to evaluate the log-likelihood for MRS models of Types II and III. We saw in Examples 3.2 and 3.3 that our methods perform well for MRS models of Type II when the EM-like algorithm fails. To our knowledge, no literature exists for exact likelihood methods for MRS models of Type II or III, so our contributions here are completely novel.

The general idea of our algorithm is to augment the hidden Markov chain with counters that record the last time each AR(1) regime was visited. This augmented process is still a Markov chain, which means similar arguments used to construct the forward-backward algorithm for MRS models with dependent regimes can be used to construct a forward and backward algorithms for these models. It turns out that our methods are related to the forward and backward algorithms for hidden semi-Markov models, where the hidden process is also augmented with a counter and the augmented hidden process is then a Markov chain [106], although this link was only realised after the fact. Though similar, the algorithms for hidden semi-Markov models are not applicable to the models considered here.

To describe the algorithm, recall our notation for MRS models. Let $\{R_t\}_{t\in\mathbb{N}}$ be a Markov chain with state space $\mathcal{S} = \{1, 2, \ldots, M\}$ and transition matrix $P = [p_{ij}]_{i,j\in\mathcal{S}}$, where $R_t \in \mathcal{S}$ represents which hidden regime the MRS process is in at time t. Suppose the set of states $\mathcal{S}_{AR} = \{1, \ldots, k < M\}$ are AR(1) regimes and all other regimes are i.i.d.

Define another Markov chain

$$\{\boldsymbol{H}_t\}_{t\in\mathbb{N}} = \{(\boldsymbol{N}_t, R_t)\}_{t\in\mathbb{N}} = \{(N_{t,1}, \ldots, N_{t,k}, R_t)\}_{t\in\mathbb{N}}, \tag{3.5}$$

($\boldsymbol{H}_t$ is for *hidden*) where $N_{t,j} \in \mathbb{N}_+ \cup \Delta_j$ counts the number of lags since the process R_t was last in Regime j before time t, for each AR(1) regime $j = 1, ..., k$, and we define Δ_j to represent when there is no time $\tau \in \{0, 1, ..., t-1\}$ with $R_\tau = j$, for each $j = 1, ..., k$. Furthermore, to help describe the evolution of $\{\boldsymbol{H}_t\}$ succinctly, define $\Delta_i \neq \Delta_j$ for $i \neq j,\ i, j \in \mathcal{S}_{AR}$ and define the operations $\Delta_i + 1 = \Delta_i$, $0 + \Delta_i = \Delta_i$ and $\Delta_i - \Delta_i = 0$ for $i \in \mathcal{S}_{AR}$. The augmented Markov chain $\{\boldsymbol{H}_t\}_{t\in\mathbb{N}}$ lives on the state space

$$\mathcal{T} := (\mathbb{N}_+ \cup \Delta_1) \times ... \times (\mathbb{N}_+ \cup \Delta_k) \times \mathcal{S}.$$

To describe the transitions of the Markov chain $\{\boldsymbol{H}_t\}$, let

$$\boldsymbol{N} := (N_1, ..., N_k) \in (\mathbb{N}_+ \cup \Delta_1) \times ... \times (\mathbb{N}_+ \cup \Delta_k)$$

be an arbitrary vector of counters, with $N_r \neq N_s$ unless $r = s$. Define $\mathbf{1}$ to be a row vector of ones of length k, and $\boldsymbol{e}_i$ to be a row vector of length k with all entries being 0 except the i^{th} entry which is 1. Also, define

$$\boldsymbol{N}^{(-i)} := \boldsymbol{N} - N_i\boldsymbol{e}_i = (N_1, ..., N_{i-1}, 0, N_{i+1}, ..., N_k).$$

The transition probabilities of $\{\boldsymbol{H}_t\}$ are

$$\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_{t+1} = (\boldsymbol{N}_{t+1}, j) | \boldsymbol{H}_t = (\boldsymbol{N}_t, i)) = \begin{cases} p_{ij}, & \text{for } i \in \mathcal{S}_{AR}^c, j \in \mathcal{S}, \boldsymbol{N}_{t+1} = \boldsymbol{N}_t + \mathbf{1}, \\ p_{ij}, & \text{for } i \in \mathcal{S}_{AR}, j \in \mathcal{S}, \boldsymbol{N}_{t+1} = \boldsymbol{N}_t^{(-i)} + \mathbf{1}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

In words, when the current state is $\boldsymbol{H}_t = (\boldsymbol{N}_t, i)$ and i is not an AR(1) regime (so there is no counter associated with state i), then, at time $t+1$, R_t transitions to state j with probability p_{ij} and all the counters are advanced by 1 to $\boldsymbol{N}_{t+1} = \boldsymbol{N}_t + \mathbf{1}$, since there has been one more time step since $\{R_t\}$ was last in any state with a counter (any state in $\mathcal{S}_{AR}$). When the current state is $\boldsymbol{H}_t = (\boldsymbol{N}_t, i)$, where i is an AR(1) regime, then R_t transitions to any state $j \in \mathcal{S}$ with probability p_{ij}, the counter for Regime i, $N_{t+1,i}$, is set to 1, since the last time in state i was t, and all other counters are advanced by 1. All other transition probabilities for $\{\boldsymbol{H}_t\}$ are 0.

The state space of $\{\boldsymbol{H}_t\}$ is countably infinite. However, due to the way $\{\boldsymbol{H}_t\}_{t\in\mathbb{N}}$ is

initialised and evolves, many states in $\mathcal{T}$ are inaccessible for $\{\boldsymbol{H}_t\}_{t\in\mathbb{N}}$ for $t \in \{0, 1, ..., T < \infty\}$, and this makes our algorithm computationally feasible. Specifically, the Markov chain $\boldsymbol{H}_t$ is initialised with a probability distribution

$$\mathbb{P}(\boldsymbol{H}_0 = (N_{0,1}, ..., N_{0,k}, j)) = \begin{cases} \pi_j, & \text{for } N_{0,i} = \Delta_i,\ i \in \mathcal{S}_{AR},\ j \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

The distribution $\boldsymbol{\pi} := (\pi_1, ..., \pi_M)$ can be any proper probability distribution, but is commonly taken to be either the stationary distribution of $\{R_t\}$, or a point mass on a single state, or, when used as part of the EM algorithm, the probabilities $\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{H}_0|\boldsymbol{x}_{0:T})$ calculated at the previous iteration of the EM algorithm. The following lemma gives the number of states that $\{\boldsymbol{H}_t\}$ can be in at time t.

Lemma 3.1. *Define, $\mathcal{S}^{(0)} := (\Delta_1, ..., \Delta_k)$ and $\mathcal{S}^{(t)}$, for $t = 1, 2, ..., T$, as the set of all vectors $\boldsymbol{N} := (N_1, ..., N_k)$ such that*

(i) $N_j \in \{1, 2, ..., t\} \cup \Delta_j$ for all $j \in \mathcal{S}_{AR}$,

(ii) there are at most $\min(t, k)$ elements of $\boldsymbol{N}$ with $N_j \neq \Delta_j$,

(iii) $N_j \neq N_m$ for all $j \neq m$, $j, m \in \mathcal{S}_{AR}$.

Given $\{\boldsymbol{H}_t\}$ is initialised with the distribution in Equation (3.7), it is only possible for $\boldsymbol{H}_t$ to reach states $(\boldsymbol{N}_t, i)$ where $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $i \in \mathcal{S}$. The cardinality of $\mathcal{S}^{(t)}$ is

$$|\mathcal{S}^{(t)}| = \sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m!.$$

Proof. First, we explain why $\mathcal{S}^{(t)}$ contains all possible values of the counters of $\boldsymbol{H}_t$.

At time $t = 0$ the chain, $\{\boldsymbol{H}_t\} = \{(\boldsymbol{N}_t, R_t)\}$, is initialised with the distribution in Equation (3.7), so

$$\mathcal{S}^{(0)} := \{(\Delta_1, ..., \Delta_k)\}.$$

At $t = 1$ the previous regime, $R_0 = i$, was either an element of S_{AR}, in which case $N_{1,i} = 1$, and all other counters keep the value Δ_j, $j \in \mathcal{S}_{AR} \setminus \{i\}$, since the chain is yet to visit the states in $\mathcal{S}_{AR} \setminus \{i\}$. Otherwise, $R_0 = i$ was an element of $\mathcal{S}_{AR}^c$ and the process $\{R_t\}$ is yet to visit any state with a counter, so

$$N_{1,j} = N_{0,j} + 1 = \Delta_j + 1 = \Delta_j, \text{ for } j \in \mathcal{S}_{AR}.$$

Thus

$$\mathcal{S}^{(1)} = \mathcal{S}^{(0)} \cup \{(\Delta_1, ..., \Delta_{i-1}, 1, \Delta_{i+1}, ..., \Delta_k) | i \in \mathcal{S}_{AR}\}.$$

For $t = 2$, first consider the case when the counters at $t-1$ were of the form

$$\boldsymbol{N}_{t-1} = (\Delta_1, ..., \Delta_{i-1}, 1, \Delta_{i+1}, ..., \Delta_k), \text{ for } i \in \mathcal{S}_{AR}.$$

If the regime at time $t-1$, was $R_1 = j \in \mathcal{S}_{AR} \setminus \{i\}$, then $N_{2,j} = 1$, since j was the state just visited, $N_{2,i} = 2$ as it has been one more time step since the chain visited state i, and $N_{2,m} = \Delta_m$ for all $m \in \mathcal{S}_{AR} \setminus \{\{i\}, \{j\}\}$, as the chain is still yet to visit these states.

If the regime at time $t-1$, was $R_1 = i$, then $N_{2,i} = 1$ as the chain just visited state i, and $N_{2,m} = \Delta_m$ for all $m \in \mathcal{S}_{AR} \setminus \{i\}$, as the chain is yet to visit any of these states.

Otherwise, at time $t-1$, the state was $R_1 = j \in \mathcal{S}_{AR}^c$, in which case $N_{t,i} = N_{t-1,i} + 1 = 2$, and $N_{t,m} = \Delta_m$ for $m \in \mathcal{S}_{AR} \setminus \{i\}$ as the chain is still yet to visit these states.

Alternatively, for $t = 2$, the counters at time $t-1$ were of the form

$$\boldsymbol{N}_{t-1} = (\Delta_1, ..., \Delta_k).$$

In this case, if at $t-1$ the regime was $R_1 = j \in \mathcal{S}_{AR}$, then $N_{2,j} = 1$ since the state j was the state just visited and all other counters remain the same, $N_{2,k} = \Delta_k$ for $k \in \mathcal{S}_{AR} \setminus \{j\}$, as these states are yet to be visited. Otherwise the regime at $t-1$ was $j \in \mathcal{S}_{AR}^c$ and all AR(1) states are yet to be visited, so $\boldsymbol{N}_t = (\Delta_1, ..., \Delta_k)$. Thus

$$\mathcal{S}^{(2)} := \mathcal{S}^{(1)} \cup A \cup B,$$

where

$$A := \{(\Delta_1, ..., \Delta_{i-1}, 2, \Delta_{i+1}, ..., \Delta_k) | i \in \mathcal{S}_{AR}\},$$
$$B := \{(\Delta_1, ..., \Delta_{i-1}, 2, \Delta_{i+1}, ..., \Delta_{j-1}, 1, \Delta_{j+1}, ..., \Delta_k) | i \neq j,\ i, j \in \mathcal{S}_{AR}\}.$$

In general, at time t, either $\{R_t\}$ has never visited state $j \in \mathcal{S}_{AR}$, in which case $N_{t,j} = \Delta_j$, or $\{R_t\}$ *last* visited j at time t_j, in which case $N_{t,j} = t - t_j \in \{1, 2, ..., t\}$ (this is part (i) of the definition). Since the process $\{R_t\}$ can only be in one regime at a time, it follows that $N_{t,j} \neq N_{t,m}$ when $j \neq m$ (part (iii) of the definition). Also, at time t, the regime chain $\{R_t\}$ could only possibly have visited $\min(t, k)$ possible states (this is part (ii) of the definition).

Now, to prove the cardinality of $\mathcal{S}^{(t)}$. The elements of $\mathcal{S}^{(t)}$ are of the form $(N_1, ..., N_k)$. At time t, let m be the possible number of counters that are not equal to Δ, so m is an

element of $\{0, 1, ..., \min(t, k)\}$. For each $m \in \{0, 1, ..., \min(t, k)\}$, there are $\binom{k}{m}$ ways of choosing which m of the k counters are not equal to Δ. Next, the value of each of these non-Δ counters needs to be specified. Each counter takes a distinct value in $\{1, ..., t\}$, so there are $\binom{t}{m}$ ways of choosing the value of the m counters. There are $m!$ possible permutations to allocate the chosen values to the counters. So, in total there are

$$\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m!$$

elements in $\mathcal{S}^{(t)}$. □

As a consequence of Lemma 3.1, if $\boldsymbol{N}_t \notin \mathcal{S}^{(t)}$ then $\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j)) = 0$ for any $j \in \mathcal{S}$. So the elements of the set $\mathcal{S}^{(t)}$ partition the space of all counters which the process $\{\boldsymbol{H}_t\}$ possibly has positive probability of reaching. Thus, for any (measurable) set A and any t, the law of total probability can be applied as

$$\mathbb{P}^{\boldsymbol{\theta}}(A) = \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \sum_{j \in \mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j), A). \tag{3.8}$$

We will use this fact multiple times in the following.

To describe our algorithm, first define $\mathcal{L}_i^{(0)} := \Delta_i$, and $\mathcal{L}_i^{(t)} := \{1, 2, ..., t\} \cup \Delta_i$, and notice that the log-likelihood of the data, $\ell(\boldsymbol{\theta}) := \log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}}(\boldsymbol{x}_{0:T})$, can be written as

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log \left\{ \sum_{i \in \mathcal{S}_{AR}} \sum_{m \in \mathcal{L}_i^{(0)}} f^{\boldsymbol{\theta}}_{X_0|R_0,N_{0,i}}(x_0|i,m)\mathbb{P}^{\boldsymbol{\theta}}(R_0 = i) \right\} \\
&\quad + \log \left\{ \sum_{i \in \mathcal{S}^c_{AR}} f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|i)\mathbb{P}^{\boldsymbol{\theta}}(R_0 = i) \right\} \\
&\quad + \sum_{t=1}^{T} \log \left\{ \sum_{i \in \mathcal{S}_{AR}} \sum_{m \in \mathcal{L}_i^{(t)}} f^{\boldsymbol{\theta}}_{X_t|R_t,N_{t,i},\boldsymbol{X}_{0:t-1}}(x_t|i,m,\boldsymbol{x}_{0:t-1})\mathbb{P}^{\boldsymbol{\theta}}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:t-1}) \right\} \\
&\quad + \sum_{t=1}^{T} \log \left\{ \sum_{i \in \mathcal{S}^c_{AR}} f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|i,\boldsymbol{x}_{0:t-1})\mathbb{P}^{\boldsymbol{\theta}}(R_t = i|\boldsymbol{x}_{0:t-1}) \right\} \\
&= \log \left\{ \sum_{i \in \mathcal{S}} \sum_{\boldsymbol{N}_0 \in \mathcal{S}^{(0)}} f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0, i))\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0 = (\boldsymbol{N}_{0,i}, i)) \right\} \\
&\quad + \sum_{t=1}^{T} \log \left\{ \sum_{i \in \mathcal{S}} \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t, i),\boldsymbol{x}_{0:t-1})\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:t-1}) \right\}.
\end{aligned}$$

Our forward algorithm calculates the probabilities $\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:t-1})$, for $t = 1, ..., T$, $i \in \mathcal{S}$ and $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$, by calculating the following

- $\widehat{\alpha}^{(t-1)}_{\boldsymbol{N}_t - \mathbf{1}}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1} = (\boldsymbol{N}_{t-1}, j)|\boldsymbol{x}_{0:t-1})$ for $j \in \mathcal{S}$,
- $\widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) := f^{\boldsymbol{\theta}}_{\boldsymbol{H}_t, X_t|\boldsymbol{X}_{0:t-1}}((\boldsymbol{N}_t, j), x_t|\boldsymbol{x}_{0:t-1})$ for $j \in \mathcal{S}$,
- $c^{(t)} := f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{x}_{0:t-1})$,

for $t = 1, 2, ..., T$. We have to treat the calculations of some quantities differently for $t = 0$, but then can proceed iteratively for $t = 1, 2, ..., T$.

Using the definition of conditional densities, calculate

$$\begin{aligned}\widetilde{\alpha}^{(0)}_{\boldsymbol{N}_0}(j) &:= f^{\boldsymbol{\theta}}_{\boldsymbol{H}_0, X_0}((\boldsymbol{N}_0, j), x_0) \\ &= f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0, j))\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{H}_0 = (\boldsymbol{N}_0, j)),\end{aligned}$$

for $\boldsymbol{N}_0 \in \mathcal{S}^{(0)}$ and $j \in \mathcal{S}$. Here, $f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0, j))$ is known from the model specification. Then using this, the law of total probability, calculate

$$\begin{aligned}c^{(0)} &:= f^{\boldsymbol{\theta}}_{X_0}(x_0) \\ &= \sum_{\boldsymbol{N}_0 \in \mathcal{S}^{(0)}} \sum_{j \in \mathcal{S}} f^{\boldsymbol{\theta}}_{\boldsymbol{H}_0, X_0}((\boldsymbol{N}_0, j), x_0) \\ &= \sum_{\boldsymbol{N}_0 \in \mathcal{S}^{(0)}} \sum_{j \in \mathcal{S}} \widetilde{\alpha}^{(0)}_{\boldsymbol{N_0}}(j)\end{aligned}$$

Next, using the definition of conditional densities, calculate

$$\begin{aligned}\widehat{\alpha}^{(0)}_{\boldsymbol{N}_0}(j) &:= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0 = (\boldsymbol{N}_0, j)|x_0) \\ &= \frac{f^{\boldsymbol{\theta}}_{\boldsymbol{H}_0, X_0}((\boldsymbol{N}_0, j), x_0)}{f^{\boldsymbol{\theta}}_{X_0}(x_0)} \\ &= \frac{\widetilde{\alpha}^{(0)}_{\boldsymbol{N}_0}(j)}{c^{(0)}},\end{aligned} \tag{3.9}$$

for $j \in \mathcal{S}$.

The algorithm then proceeds iteratively for $t = 1, ..., T$. First, for $i \in \mathcal{S}$, calculate the prediction probabilities

$$\begin{aligned}a^{(t)}_{\boldsymbol{N}_t}(i) &:= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:t-1}) \\ &= \sum_{j \in \mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), R_{t-1} = j|\boldsymbol{x}_{0:t-1}).\end{aligned} \tag{3.10}$$

By definition, at time $t-1$, the counters either transition from $\boldsymbol{N}_{t-1}$ to $\boldsymbol{N}_t = \boldsymbol{N}_{t-1} + \mathbf{1}$ when $R_{t-1} \in \mathcal{S}_{AR}^c$, or from $\boldsymbol{N}_{t-1}$ to $\boldsymbol{N}_t = \boldsymbol{N}_{t-1}^{(-j)} + \mathbf{1}$ when $R_{t-1} = j \in \mathcal{S}_{AR}$. In the former case all elements of $\boldsymbol{N}_t$ are different from 1, and in the latter case exactly the j^{th} element of $\boldsymbol{N}_t$ is equal to 1. So, given all elements of $\boldsymbol{N}_t$ are different from 1, then $R_{t-1} \in \mathcal{S}_{AR}^c$, and given the j^{th} element of $\boldsymbol{N}_t$ is 1, then $R_{t-1} = j \in S_{AR}$.

Thus, in the case where all elements of $\boldsymbol{N}_t$ are different from 1, and thus $R_{t-1} \in \mathcal{S}_{AR}^c$, then the prediction probability (3.10) equals

$$\begin{aligned}
&\sum_{j \in \mathcal{S}_{AR}^c} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), R_{t-1} = j | \boldsymbol{x}_{0:t-1}) \\
&= \sum_{j \in \mathcal{S}_{AR}^c} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1} = (\boldsymbol{N}_t - \mathbf{1}, j), R_t = i | \boldsymbol{x}_{0:t-1}), \\
&= \sum_{j \in \mathcal{S}_{AR}^c} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1} = (\boldsymbol{N}_t - \mathbf{1}, j) | \boldsymbol{x}_{0:t-1}) \mathbb{P}^{\boldsymbol{\theta}}(R_t = i | \boldsymbol{H}_{t-1} = (\boldsymbol{N}_t - \mathbf{1}, j), \boldsymbol{x}_{0:t-1}) \\
&= \sum_{j \in \mathcal{S}_{AR}^c} \widehat{\alpha}_{\boldsymbol{N}_t - \mathbf{1}}^{(t-1)}(j) p_{ji},
\end{aligned}$$

where the *forward probabilities*, defined as

$$\widehat{\alpha}_{\boldsymbol{N}_t - \mathbf{1}}^{(t-1)}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_{t-1} = (\boldsymbol{N}_{t-1}, j) | \boldsymbol{x}_{0:t-1}),$$

for $\boldsymbol{N}_{t-1} \in \mathcal{S}^{(t-1)}$ and $j \in \mathcal{S}$, are known from the previous iteration of the algorithm. The last equality holds because

$$\mathbb{P}^{\boldsymbol{\theta}}(R_t = i | \boldsymbol{H}_{t-1} = (\boldsymbol{N}_t - \mathbf{1}, j), \boldsymbol{x}_{0:t-1}) = p_{ji},$$

which comes from the fact R_t is independent of $\boldsymbol{N}_{t-1}$ and $\boldsymbol{x}_{0:t-1}$ given R_{t-1}.

In the other case, where exactly the j^{th} element of $\boldsymbol{N}_t$ is equal to 1, and thus $R_{t-1} = j \in \mathcal{S}_{AR}$ and the counters transition from $\boldsymbol{N}_{t-1}$ to $\boldsymbol{N}_t = \boldsymbol{N}_{t-1}^{(-j)} + \mathbf{1}$, then

$$\begin{aligned}
a_{\boldsymbol{N}_t}^{(t)}(i) &= \mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), R_{t-1} = j | \boldsymbol{x}_{0:t-1}\right), \\
&= \sum_{m \in \mathcal{L}_j^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), R_{t-1} = j, N_{t-1,j} = m | \boldsymbol{x}_{0:t-1}\right). \qquad (3.11)
\end{aligned}$$

When $R_{t-1} = j$, $R_t = i$, $\boldsymbol{N}_t$ and $N_{t-1,j} = m$ are known, then, by definition the value of the counter at $t-1$ is known, $\boldsymbol{N}_{t-1} = \boldsymbol{N}_t - \mathbf{1} + m\boldsymbol{e}_j =: \boldsymbol{N}_{t-1}^{j,m}$, and thus (3.11) is equal to

$$\sum_{m \in \mathcal{L}_j^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t-1} = \left(\boldsymbol{N}_{t-1}^{j,m}, j\right), R_t = i, N_{t,j} = 1 | \boldsymbol{x}_{0:t-1}\right),$$

$$
\begin{aligned}
&= \sum_{m \in \mathcal{L}_j^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}} \left(\boldsymbol{H}_{t-1} = \left(\boldsymbol{N}_{t-1}^{j,m}, j \right) | \boldsymbol{x}_{0:t-1} \right) \\
&\quad \times \mathbb{P}^{\boldsymbol{\theta}} \left(R_t = i, N_{t,j} = 1 | \boldsymbol{H}_{t-1} = \left(\boldsymbol{N}_{t-1}^{j,m}, j \right), \boldsymbol{x}_{0:t-1} \right) \\
&= \sum_{m \in \mathcal{L}_j^{(t-1)}} \widehat{\alpha}_{\boldsymbol{N}_{t-1}^{j,m}}^{(t-1)}(j) p_{ji},
\end{aligned}
$$

for all $i \in \mathcal{S}$ and all $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ with $N_{t,j} = 1$. The last equality holds from the fact that $\mathbb{P}^{\boldsymbol{\theta}}(R_t = i, N_{t,j} = 1 | \boldsymbol{H}_{t-1} = (\boldsymbol{N}_{t-1}^{j,m}, j), \boldsymbol{x}_{0:t-1}) = p_{ji}$, since $N_{t,j} = 1$ with probability 1 given $R_{t-1} = j$, and R_t is independent of $\boldsymbol{N}_{t-1}$ and $\boldsymbol{x}_{0:t-1}$ given R_{t-1}.

Using the definition of conditional density, calculate $\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)$ for $t \geq 1$ as

$$
\begin{aligned}
\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j) &:= f_{\boldsymbol{H}_t, X_t | \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}((\boldsymbol{N}_t, j), x_t | \boldsymbol{x}_{0:t-1}) \\
&= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t-1}) f_{X_t | \boldsymbol{H}_t, \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | (\boldsymbol{N}_t, j), \boldsymbol{x}_{0:t-1}) \\
&= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t-1}) f_{X_t | N_{t,j}, R_t, \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | m, j, \boldsymbol{x}_{0:t-1}) \\
&= a_{\boldsymbol{N}_t}^{(t)}(j) f_{X_t | N_{t,j}, R_t, \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | m, j, \boldsymbol{x}_{0:t-1}),
\end{aligned}
$$

where $f_{X_t | \boldsymbol{N}_{t,j}, R_t, \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | m, j, \boldsymbol{x}_{0:t-1})$ is given by the model specification. Using the values $\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)$, the law of total probability, then calculate for $t \geq 1$

$$
\begin{aligned}
c^{(t)} &:= f_{X_t | \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | \boldsymbol{x}_{0:t-1}) \\
&= \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \sum_{j \in \mathcal{S}} f_{\boldsymbol{H}_t, X_t | \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}((\boldsymbol{N}_t, j), x_t | \boldsymbol{x}_{0:t-1}) \\
&= \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \sum_{j \in \mathcal{S}} \widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j).
\end{aligned}
$$

The terms $c^{(t)}$ are used to calculate the forward probabilities as follows

$$
\begin{aligned}
\widehat{\alpha}_{\boldsymbol{N}_t}^{(t)}(j) &:= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t}) \\
&= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | x_t, \boldsymbol{x}_{0:t-1}) \\
&= \frac{f_{\boldsymbol{H}_t, X_t | \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}((\boldsymbol{N}_t, j), x_t | \boldsymbol{x}_{0:t-1})}{f_{X_t | \boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t | \boldsymbol{x}_{0:t-1})} \\
&= \frac{\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)}{c^{(t)}},
\end{aligned}
$$

for $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$, $j \in \mathcal{S}$ and $t \geq 1$). The algorithm then proceeds to the next iteration, calculating $a_{\boldsymbol{N}_{t+1}}^{(t+1)}(j)$ using the values $\widehat{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)$.

Input: Data, $\boldsymbol{x_{0:T}}$, parameters, $\boldsymbol{\theta}$.
Output: The log-likelihood $\ell(\boldsymbol{\theta}) := \log f^{\boldsymbol{\theta}}_{\boldsymbol{X}}(\boldsymbol{x})$.
Initialise π_j for $j = \{1, ..., M\}$; set $\ell = 0$; $c^{(0)} = 0$;
for $j = 1, 2, ..., M$ **do**
  $\widetilde{\alpha}^{(0)}_{\Delta_1,...,\Delta_k}(j) = f^{\boldsymbol{\theta}}_{\boldsymbol{X}_0|\boldsymbol{H}_0}(\boldsymbol{x}_0|(\Delta_1, ..., \Delta_k, j))\pi(j)$;
end
$c^{(0)} = \sum_{j=1}^{M} \widetilde{\alpha}^{(0)}_{\Delta_1,...,\Delta_k}(j)$;
for $j = 1, 2, ..., M$ **do**
  $\widehat{\alpha}^{(0)}_{\Delta_1,...,\Delta_k}(j) = \dfrac{\widetilde{\alpha}^{(0)}_{\Delta_1,...,\Delta_k}(j)}{c^{(0)}}$;
end
for $t = 1, 2, ..., T$ **do**
  for $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ **do**
    for $i = 1, ..., M$ **do**
      if *any* $N_{t,p} == 1$ **then**
        $a^{(t)}_{\boldsymbol{N}_t}(i) = \sum_{m \in \mathcal{L}^{(t)}_p} \widehat{\alpha}^{(t-1)}_{\boldsymbol{N}^{i,m}_{t-1}}(p)p_{pi}$;
      else
        $a^{(t)}_{\boldsymbol{N}_t}(i) = \sum_{j=k+1}^{M} \widehat{\alpha}^{(t-1)}_{\boldsymbol{N}_t - \boldsymbol{1}}(j)p_{ji}$;
      end
      $\widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(i) = a^{(t)}_{\boldsymbol{N}_t}(i) f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t, i), \boldsymbol{x}_{0:t-1})$;
    end
  end
  $c^{(t)} = \sum_{j=1}^{M} \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)$;
  for *all* i *and all* $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ **do**
    $\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i) = \dfrac{\widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)}{c^{(t)}}$;
  end
  $\ell = \ell + \log c^{(t)}$;
end
return ℓ;

FIGURE 3.6: Pseudo-code implementing our forward algorithm

Finally, the log-likelihood is given by

$$\ell(\theta) = \sum_{t=0}^{T} \log(c^{(t)}). \tag{3.12}$$

The algorithm is presented in pseudo-code in Figure 3.6.

Lemma 3.2. *The complexity of our forward algorithm is*

$$\begin{aligned} C &= 2M + M^2 + M \sum_{m=0}^{\min(T,k)} \binom{T}{m}\binom{k}{m} m! \\ &\quad + \sum_{t=1}^{T-1} \left[M^2 \sum_{m=0}^{\min(t,k-1)} \binom{t}{m}\binom{k-1}{m} m! + 2M \sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] \\ &\leq \mathcal{O}(M^2 T^{k+1} k^k). \end{aligned}$$

Proof. The complexity, C, of this algorithm is calculated by counting all multiplications required. Recall that there are k elements in $\mathcal{S}_{AR}$, M elements in $\mathcal{S}$, $(M-k)$ elements in $\mathcal{S}_{AR}^c$, and $\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m!$ elements in $\mathcal{S}^{(t)}$.

(a) Our forward algorithm first calculates

$$\widetilde{\alpha}_{\boldsymbol{N}_0}^{(0)}(j) = f_{X_0|\boldsymbol{H}_0}^{\boldsymbol{\theta}}(x_0|(\boldsymbol{N}_0, j))\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0),$$

for $\boldsymbol{N}_0 \in \mathcal{S}^{(0)}$ and $j \in \mathcal{S}$. This takes M multiplications, one for each $j \in \mathcal{S}$.

(b) Calculating $c^{(0)}$ takes no multiplications.

(c) Calculating

$$\widehat{\alpha}_{\boldsymbol{N}_0}^{(0)}(j) = \frac{\widetilde{\alpha}_{\boldsymbol{N}_0}^{(0)}(j)}{c^{(0)}},$$

for $\boldsymbol{N}_0 \in \mathcal{S}^{(0)}$, and $j \in \mathcal{S}$, takes M multiplications, one for each $j \in \mathcal{S}$.

Then the iterations for $t = 1, 2, ..., T$ start.

(d) The algorithm calculates

$$a_{\boldsymbol{N}_t}^{(t)}(i) = \sum_{j \in \mathcal{S}_{AR}^c} \widehat{\alpha}_{\boldsymbol{N}_t - \boldsymbol{1}}^{(t-1)}(j) p_{ji},$$

for $t \in \{1, ..., T\}$, $i \in \mathcal{S}$, and $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ such that $N_{t,j} \neq 1$ for all $j \in \mathcal{S}_{AR}$. To calculate these, the multiplication $\widehat{\alpha}_{\boldsymbol{N}_t - \boldsymbol{1}}^{(t-1)}(j) p_{ji}$ needs to be done for every $j \in \mathcal{S}_{AR}$, $i \in \mathcal{S}$, and every $\boldsymbol{N}_t - \boldsymbol{1} \in \mathcal{S}^{(t-1)}$, for each $t = 1, 2, ..., T$. For a given t this requires

$$\overbrace{M}^{\text{no. of } i \in \mathcal{S}} \quad \overbrace{(M-k)}^{\text{no. of } j \in \mathcal{S}_{AR}} \quad \overbrace{\left[\sum_{m=0}^{\min(t-1,k)} \binom{t-1}{m}\binom{k}{m} m! \right]}^{\text{no. of } \boldsymbol{N}_t - \boldsymbol{1} \in \mathcal{S}^{(t-1)}}$$

multiplications. So in total there are

$$M(M-k)\sum_{t=1}^{T}\left[\sum_{m=0}^{\min(t-1,k)}\binom{t-1}{m}\binom{k}{m}m!\right]$$

multiplications for this step.

(e) Our algorithm also calculates

$$a_{\boldsymbol{N}_t}^{(t)}(i)=p_{ji}\sum_{m\in\mathcal{L}_j^{(t-1)}}\widehat{\alpha}_{(\boldsymbol{N}_{t-1}^{j,m})}^{(t-1)}(j),$$

for $t\in\{1,...,T\}$, $i\in\mathcal{S}$, and all $\boldsymbol{N}_t\in\mathcal{S}^{(t)}$ with $N_{t,j}=1$, $j\in\mathcal{S}_{AR}$. For a given $t\geq 1$, there are

$$\underbrace{k}_{\text{no. of } j\in\mathcal{S}_{AR}}\overbrace{\sum_{m=0}^{\min(t-1,k-1)}\binom{t-1}{m}\binom{k-1}{m}m!}^{\text{no. of } \boldsymbol{N}_t\in\mathcal{S}^{(t)} \text{ with } N_{t,j}=1}$$

elements in $\boldsymbol{N}_t\in\mathcal{S}^{(t)}$ with $N_{t,j}=1$, $j=\{1,2,...,k\}$. Since this calculation is executed for each $i\in\mathcal{S}$ and $t=1,2,...,T$ it requires

$$Mk\sum_{t=1}^{T}\left[\sum_{m=0}^{\min(t-1,k-1)}\binom{t-1}{m}\binom{k-1}{m}m!\right]$$

multiplications in total. Thus, calculating all of the $a_{\boldsymbol{N}_t}^{(t)}(j)$ terms, for $t=1,...,T$, $\boldsymbol{N}_t\in\mathcal{S}^{(t)}$ and $j\in\mathcal{S}$, in the algorithm takes

$$M(M-k)\sum_{t=1}^{T}\left[\sum_{m=0}^{\min(t-1,k)}\binom{t-1}{m}\binom{k}{m}m!\right]+Mk\sum_{t=1}^{T}\left[\sum_{m=0}^{\min(t-1,k-1)}\binom{t-1}{m}\binom{k-1}{m}m!\right]$$

multiplications.

(f) Our algorithm then calculates

$$\widetilde{\alpha}_{\boldsymbol{N}_t}^{(t)}(j)=a_{\boldsymbol{N}_t}^{(t)}(j)f_{X_t|R_t,N_{t,j},\boldsymbol{X}_{0:t-1}}^{\boldsymbol{\theta}}(x_t|j,n,\boldsymbol{x}_{0:t-1}),$$

for $t\in\{1,...,T\}$, $j\in\mathcal{S}$, and $\boldsymbol{N}_t\in\mathcal{S}^{(t)}$. This requires

$$M\sum_{t=1}^{T}\left[\sum_{m=0}^{\min(t,k)}\binom{t}{m}\binom{k}{m}m!\right]$$

multiplications in total.

(g) Calculating $c^{(t)}$ requires no multiplications.

(h) Next, the calculation

$$\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) = \frac{\widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(j)}{c^{(t)}}$$

only needs to be executed for $t \in \{1, ..., T-1\}$, $j \in \mathcal{S}$, and $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$. So, in total, this takes

$$M \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right]$$

multiplications.

So the total number of multiplications for the whole algorithm, C, is

$$\overbrace{M}^{\widetilde{\alpha}^{(0)}_{\boldsymbol{N}_0}(j) \text{ terms}} + \overbrace{M}^{\widehat{\alpha}^{(0)}_{\boldsymbol{N}_0}(j) \text{ terms}}$$

$$+ \overbrace{M(M-k) \sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t-1,k)} \binom{t-1}{m}\binom{k}{m} m! \right] + Mk \sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t-1,k-1)} \binom{t-1}{m}\binom{k-1}{m} m! \right]}^{a^{(t)}_{\boldsymbol{N}_t}(j) \text{ terms}}$$

$$+ \overbrace{M \sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right]}^{\widetilde{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) \text{ terms}} + \overbrace{M \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right]}^{\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) \text{ terms}}. \tag{3.13}$$

Noting the following manipulations

$$\sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t-1,k)} \binom{t-1}{m}\binom{k}{m} m! \right] = \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] + 1$$

$$\sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t-1,k-1)} \binom{t-1}{m}\binom{k-1}{m} m! \right] = \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k-1)} \binom{t}{m}\binom{k-1}{m} m! \right] + 1,$$

$$\sum_{t=1}^{T} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] = \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] + \sum_{m=0}^{\min(T,k)} \binom{T}{m}\binom{k}{m} m!,$$

then we can manipulate the sums over t in Expression (3.13) so they all have common limits;

$$M + M + M(M-k) + Mk$$
$$+ M(M-k) \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] + Mk \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k-1)} \binom{t}{m}\binom{k-1}{m} m! \right]$$
$$+ M \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m}\binom{k}{m} m! \right] + M \sum_{m=0}^{\min(T,k)} \binom{T}{m}\binom{k}{m} m!$$

$$
\begin{aligned}
&+ M \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m} \binom{k}{m} m! \right] \\
&= 2M + M^2 + M \sum_{m=0}^{\min(T,k)} \binom{T}{m} \binom{k}{m} m! \\
&+ \sum_{t=1}^{T-1} \left[Mk \sum_{m=0}^{\min(t,k-1)} \binom{t}{m} \binom{k-1}{m} m! + (2M + M^2 - Mk) \sum_{m=0}^{\min(t,k)} \binom{t}{m} \binom{k}{m} m! \right].
\end{aligned}
$$

By replacing $k-1$ in the above by k, we get the upper bound

$$
\begin{aligned}
C &\leq 2M + M^2 + M \sum_{m=0}^{\min(T,k)} \binom{T}{m} \binom{k}{m} m! + \left(M^2 + 2M\right) \sum_{t=1}^{T-1} \left[\sum_{m=0}^{\min(t,k)} \binom{t}{m} \binom{k}{m} m! \right] \\
&\leq 2M + M^2 + M \sum_{m=0}^{\min(T,k)} \binom{T}{m} \binom{k}{m} m! \\
&\quad + \left(M^2 + 2M\right) (T-1) \left[\sum_{m=0}^{\min(T-1,k)} \binom{T-1}{m} \binom{k}{m} m! \right]. \qquad (3.14)
\end{aligned}
$$

Now, noting that k is fixed, observe

$$
\sum_{m=0}^{\min(t,k)} \binom{t}{m} \binom{k}{m} m! \leq \sum_{m=0}^{k} \binom{t}{m} \binom{k}{m} m! \leq \sum_{m=0}^{k} \frac{t^m}{m!} \frac{k^m}{m!} m! \leq k \frac{t^k k^k}{k!} = \frac{t^k k^k}{(k-1)!}, \quad (3.15)
$$

where the second inequality follows from the well-known result for binomial coefficients, $\binom{t}{m} \leq \frac{t^m}{m!}$. So (3.14) is less than

$$
\begin{aligned}
&2M + M^2 + M \frac{T^k k^k}{(k-1)!} + \left(M^2 + 2M\right) (T-1) \left[\frac{(T-1)^k k^k}{(k-1)!} \right] \\
&= \mathcal{O}\left(M^2 T^{k+1} k^k \right). \qquad \square
\end{aligned}
$$

Since k is usually not too large (1 or 2), our algorithm is feasible and is favourable compared to the naive method, where the sum in Equation (2.5) is calculated directly, which is $\mathcal{O}(M^T)$.

The densities As mentioned above, the densities

$$
f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t, \boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t, i), \boldsymbol{x}_{0:t-1}) = f^{\boldsymbol{\theta}}_{X_t|N_{t,i}, R_t, \boldsymbol{X}_{0:t-1}}(x_t|m, i, \boldsymbol{x}_{0:t-1}) \qquad (3.16)
$$

are determined by the model specification. For i.i.d. processes this is trivial,

$$f^{\boldsymbol{\theta}}_{X_t|N_{t,i},R_t,\boldsymbol{X}_{0:t-1}}(x_t|m,i,\boldsymbol{x}_{0:t-1}) = f_{X_t|R_t}(x_t|i)$$

is the density in the i^{th} i.i.d. regime. For AR(1) regimes it depends on whether the regimes are specified to evolve only when they are observed (models of Type III), or at all time points (models of Type II).

The densities for Type III Models Consider the MRS model specification with independent regimes which evolve only at times when they are observed. Let Regime i be an AR(1) process from this model. That is, $\left\{B^{(i)}_{\tau(t)}\right\}_{t\in\mathbb{N}}$ is an AR(1) process, defined by

$$B^{(i)}_{\tau(t)} = \alpha_i + \phi_i B^{(i)}_{\tau(t-1)} + \sigma_i \varepsilon^{(i)}_{\tau(t)},$$

where $\left\{\varepsilon^{(i)}_{\tau(t)}\right\}_{t\in\mathbb{N}}$ is i.i.d. N(0,1) noise, $\tau(t) = \sum_{\ell=0}^{t} \mathbb{I}(R_\ell = i)$ and $\tau(t-1) = \sum_{\ell=0}^{t-1} \mathbb{I}(R_\ell = i)$. Then the distribution function in this regime is

$$f^{\boldsymbol{\theta}}_{X_t|N_{t,i},R_t,\boldsymbol{X}_{0:t-1}}(x_t|m,i,\boldsymbol{x}_{0:t-1}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp -\frac{1}{2\sigma_i^2}(x_t - \alpha_i - \phi_i x_{t-m})^2. \tag{3.17}$$

The densities for Type II Models Consider the MRS model specification with independent regimes which evolve at all times points (so AR(1) processes in this model evolve regardless of whether they are observed or not) as in Figure 3.1. Let Regime i be an AR(1) process from such a model, $\left\{B^{(i)}_t\right\}_{t\in\mathbb{N}}$, defined by

$$B^{(i)}_t = \alpha_i + \phi_i B^{(i)}_{t-1} + \sigma_i \varepsilon^{(i)}_t,$$

where $\left\{\varepsilon^{(i)}_t\right\}_{t\in\mathbb{N}}$ is i.i.d. N(0,1) noise. A recursive argument gives

$$\begin{aligned}
B^{(i)}_t &= \alpha_i + \phi_i B^{(i)}_{t-1} + \sigma_i \varepsilon^{(i)}_t \\
&= \alpha_i + \phi_i(\alpha_i + \phi_i B^{(i)}_{t-2} + \sigma_i \varepsilon^{(i)}_{t-1}) + \sigma_i \varepsilon^{(i)}_t \\
&= \alpha_i(1+\phi_i) + \phi_i^2 B^{(i)}_{t-2} + \phi_i\sigma_i\varepsilon^{(i)}_{t-1} + \sigma_i\varepsilon^{(i)}_t \\
&= \alpha_i(1+\phi_i) + \phi_i^2(\alpha_i + \phi_i B^{(i)}_{t-3} + \sigma_i\varepsilon^{(i)}_{t-2}) + \phi_i\sigma_i\varepsilon^{(i)}_{t-1} + \sigma_i\varepsilon^{(i)}_t \\
&= \alpha_i(1+\phi_i+\phi_i^2) + \phi_i^3 B^{(i)}_{t-3} + \sigma_i\varepsilon^{(i)}_t + \phi_i\sigma_i\varepsilon^{(i)}_{t-1} + \phi_i^2\sigma_i\varepsilon^{(i)}_{t-2} \\
&\;\;\vdots \\
&= \alpha_i \sum_{k=0}^{m-1} \phi_i^k + \phi_i^m B^{(i)}_{t-m} + \sigma_i \sum_{k=0}^{m-1} \phi_i^k \varepsilon^{(i)}_{t-k}.
\end{aligned}$$

Now, since $\varepsilon_t^{(i)}$ is i.i.d. N(0,1) then $\sum_{k=0}^{m-1} \phi_i^k \varepsilon_{t-k}^{(i)} \sim \mathrm{N}\left(0, \sum_{k=0}^{m-1} \phi_i^{2k}\right)$. Hence

$$\begin{aligned} B_t^{(i)} | B_{t-m}^{(i)} &\sim \mathrm{N}\left(\alpha_i \sum_{k=0}^{m-1} \phi_i^k + \phi_i^m B_{t-m}^{(i)}, \sigma_i^2 \sum_{k=0}^{m-1} \phi_i^{2k}\right) \\ &= \mathrm{N}\left(\alpha_i \left(\frac{1-\phi_i^m}{1-\phi_i}\right) + \phi_i^m B_{t-m}^{(i)}, \sigma_i^2 \frac{1-\phi_i^{2m}}{1-\phi_i^2}\right), \end{aligned} \tag{3.18}$$

for any $m \in \mathbb{N}_+$. So the distribution function when i is an AR(1) regime is

$$\begin{aligned} &f^{\boldsymbol{\theta}}_{X_t | N_{t,i}, R_t, \boldsymbol{X}_{0:t-1}}(x_t | m, i, \boldsymbol{x}_{t-1}) \\ &= \frac{1}{\left(2\pi\sigma_i^2 \left(\frac{1-\phi_i^{2m}}{1-\phi_i^2}\right)\right)^{(1/2)}} \exp\left\{ - \frac{\left(x_t - \alpha_i \frac{1-\phi_i^m}{1-\phi_i} - \phi_i^m x_{t-m}\right)^2}{2\sigma_i^2 \left(\frac{1-\phi_i^{2m}}{1-\phi_i^2}\right)} \right\}. \end{aligned} \tag{3.19}$$

Other useful results from our forward algorithm As a byproduct of the forward algorithm the prediction probabilities

$$a^{(t)}_{\boldsymbol{N}_t}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t-1}), \tag{3.20}$$

and the filtered probabilities,

$$\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t}), \tag{3.21}$$

for $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $j \in \mathcal{S}$ are calculated. We refer to the prediction (3.20) and filtered (3.21) probabilities again in Section 3.3 and use them as part of our backward algorithm.

The filtered probabilities, $\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t})$, can be used to calculate the probabilities

$$P^{\boldsymbol{\theta}}(R_t = j | \boldsymbol{x}_{0:t}) = \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t}).$$

Similarly, the prediction probabilities, $a^{(t)}_{\boldsymbol{N}_t}(j) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t-1})$, can be used to calculate the probabilities

$$P^{\boldsymbol{\theta}}(R_t = j | \boldsymbol{x}_{0:t-1}) = \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, j) | \boldsymbol{x}_{0:t-1}).$$

Remark 3.1. Note, when evaluating the likelihood using our forward algorithm, calculating $\widehat{\alpha}^{(T)}_{\boldsymbol{N}_T}(j)$ is unnecessary. This is taken into account in the calculation of complexity

above. However, if the forward algorithm is to be used as a precursor to our backward algorithm in Section 3.3, then this step is necessary and adds to the computational complexity but the complexity remains at worst $\mathcal{O}(M^2T^{k+1}k^k)$.

Remark 3.2. The likelihood could also be evaluated using the following. Define

$$\alpha^{(t)}_{\boldsymbol{N}_t}(j) := f^{\boldsymbol{\theta}}_{\boldsymbol{H}_t,\boldsymbol{X}_{0:t}}((\boldsymbol{N}_t, j), \boldsymbol{x}_{0:t})$$

for $t = 0, 1, ..., T$. First calculate

$$\alpha^{(0)}_{\boldsymbol{N}_0}(j) = f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0, j))\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{H}_0 = (\boldsymbol{N}_0, j)).$$

Then for $t = 1, 2, \ldots, T$ and for $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$, $j \in \mathcal{S}$ calculate

$$\alpha^{(t)}_{\boldsymbol{N}_t}(j) \tag{3.22}$$

$$= \begin{cases} \sum\limits_{i \in \mathcal{S}^c_{AR}} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t, j), \boldsymbol{x}_{0:t-1}) p_{ij} \alpha^{(t-1)}_{\boldsymbol{N}_t - \boldsymbol{1}}(i), & N_{t,\ell} \neq 1, \forall \ell \in \mathcal{S}_{AR}, \\ \sum\limits_{m=1}^{t} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t, j), \boldsymbol{x}_{0:t-1}) p_{\ell j} \alpha^{(t-1)}_{\boldsymbol{N}^{\ell m}_{t-1}}(\ell), & N_{t,\ell} = 1, \text{ for some } \ell \in \mathcal{S}_{AR} \end{cases} \tag{3.23}$$

where $\boldsymbol{N}^{\ell m}_{t-1} = (N_{t,1} - 1, \ldots, N_{t,\ell-1} - 1, m, N_{t,\ell+1} - 1, \ldots, N_{t,k} - 1)$. The likelihood is given by

$$L(\boldsymbol{\theta}) = \sum_{j \in \mathcal{S}} \sum_{\boldsymbol{N}_T \in \mathcal{S}^{(T)}} \alpha^{(T)}_{\boldsymbol{N}_T}(j). \tag{3.24}$$

While this is a more pleasant object than our algorithm above, it is not practical as it can suffer from underflow.

3.3 A novel backward algorithm

In this section a *new* backward algorithm is presented, analogous to Baum's backward algorithm for HMMs [9–11], and Kim's backward algorithm for MRS models with dependent regimes [67] (see also Section 2.3.3). Our new backward algorithm gives a computationally feasible method to calculate *smoothed probabilities* for MRS models with independent regimes. Smoothed probabilties are of interest since, as we shall see in Section 3.4, they can be used as part of the EM algorithm for MRS models with independent regimes.

Recall our notation for the forward algorithm (Section 3.2) and assume the prediction probabilities $a^{(t)}_{\boldsymbol{N}_t}(i) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:t-1})$ and filtered probabilities $\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:t})$ are known after running the forward algorithm. The goal is to

calculate the smoothed probabilities

$$\gamma^{(t)}_{\boldsymbol{N}_t}(i) := \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T}),$$

for $t = 0, 1, ..., T$, $\boldsymbol{N}_t = (N_{1,t}, ..., N_{k,t}) \in \mathcal{S}^{(t)}$ and $i \in \mathcal{S}$.

Lemma 3.3. *The smoothed probabilities can be calculated using the following. Set*

$$\gamma^{(T)}_{\boldsymbol{N}_T}(i) = \widehat{\alpha}^{(T)}_{\boldsymbol{N}_T}(i), \text{ for all } i \in \mathcal{S}, \ \boldsymbol{N}_T \in \mathcal{S}^{(T)}.$$

Then, for $t = T-1, T-2, ..., 0$, calculate

$$\gamma^{(t)}_{\boldsymbol{N}_t}(i) = \begin{cases} \widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i) \sum\limits_{j\in\mathcal{S}} p_{ij} \dfrac{\gamma^{(t+1)}_{\boldsymbol{N}_t+\boldsymbol{1}}(j)}{a^{(t+1)}_{\boldsymbol{N}_t+\boldsymbol{1}}(j)} & \text{for } i \in \mathcal{S}^c_{AR}, \ \boldsymbol{N}_t \in \mathcal{S}^{(t)} \\ \widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i) \sum\limits_{j\in\mathcal{S}} p_{ij} \dfrac{\gamma^{(t+1)}_{\boldsymbol{N}_t^{(-i)}+\boldsymbol{1}}(j)}{a^{(t+1)}_{\boldsymbol{N}_t^{(-i)}+\boldsymbol{1}}(j)} & \text{for } i \in \mathcal{S}_{AR}, \ \boldsymbol{N}_t \in \mathcal{S}^{(t)}. \end{cases}$$

This requires $(2k+1)|\mathcal{S}^{(t)}|k$ multiplications for each $t \in \{0, 1, ..., T\}$ and each $i \in \mathcal{S}$, and the total complexity of the algorithm, as measured by the total number of multiplications is

$$C \leq (M^2 + 2M)\frac{T^{k+1}k^k}{(k-1)!} = \mathcal{O}(M^2T^{k+1}k^k).$$

Proof. Consider the event $\{\boldsymbol{H}_t = (\boldsymbol{N}_t, i)\}$, by the definition of $\{\boldsymbol{N}_t\}$, when $i \in \mathcal{S}_{AR}$, then $\boldsymbol{N}_{t+1} = \boldsymbol{N}_t^{(-i)} + \boldsymbol{1}$, and when $i \in \mathcal{S}^c_{AR}$, then $\boldsymbol{N}_{t+1} = \boldsymbol{N}_t + \boldsymbol{1}$. Thus, when $\boldsymbol{H}_t = (\boldsymbol{N}_t, i)$ is known, then $\boldsymbol{N}_{t+1}$ is also known. As a result,

$$\begin{aligned} \gamma^{(t)}_{\boldsymbol{N}_t}(i) &= \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T}) \\ &= \sum_{j\in\mathcal{S}} \sum_{\boldsymbol{N}_{t+1}\in\mathcal{S}^{(t+1)}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), \boldsymbol{H}_{t+1} = (\boldsymbol{N}_{t+1}, j)|\boldsymbol{x}_{0:T}) \\ &= \begin{cases} \sum\limits_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), \boldsymbol{H}_{t+1} = \left(\boldsymbol{N}_t^{(-i)} + \boldsymbol{1}, j\right)|\boldsymbol{x}_{0:T}\right), & \text{for } i \in \mathcal{S}_{AR}, \\ \sum\limits_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_t = (\boldsymbol{N}_t, i), \boldsymbol{H}_{t+1} = (\boldsymbol{N}_t + \boldsymbol{1}, j)|\boldsymbol{x}_{0:T}\right), & \text{for } i \in \mathcal{S}^c_{AR}, \end{cases} \end{aligned} \tag{3.25}$$

for $t = 0, 1, ..., T-1$, $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $i \in \mathcal{S}$. Since the following arguments are the same for both cases, $i \in \mathcal{S}_{AR}$ and $i \in \mathcal{S}^c_{AR}$, for notational convenience, let $\boldsymbol{N}$ take the value $\boldsymbol{N}_t^{(-i)} + \boldsymbol{1}$ when $i \in \mathcal{S}_{AR}$ and the value $\boldsymbol{N}_t + \boldsymbol{1}$ when $i \in \mathcal{S}^c_{AR}$.

Using the definition of conditional densities multiple times, the right hand side of (3.25) can be written as

$$\begin{aligned}
&\sum_{j\in\mathcal{S}} \frac{f^{\boldsymbol{\theta}}_{\boldsymbol{H}_t,\boldsymbol{H}_{t+1},\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}\left((\boldsymbol{N}_t,i),(\boldsymbol{N},j),\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t}\right)}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t})} \\
&= \sum_{j\in\mathcal{S}} \Bigg[\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{H}_t=(\boldsymbol{N}_t,i),\boldsymbol{x}_{0:t}\right) \\
&\qquad \times \frac{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{H}_t,\boldsymbol{H}_{t+1},\boldsymbol{X}_{0:t}}\left(\boldsymbol{x}_{t+1:T}|(\boldsymbol{N}_t,i),(\boldsymbol{N},j),\boldsymbol{x}_{0:t}\right)}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t})} \Bigg] \\
&= \sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})p_{ij}\frac{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{H}_t,\boldsymbol{H}_{t+1},\boldsymbol{X}_{0:t}}\left(\boldsymbol{x}_{t+1:T}|(\boldsymbol{N}_t,i),(\boldsymbol{N},j),\boldsymbol{x}_{0:t}\right)}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}\left(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t}\right)},
\end{aligned} \tag{3.26}$$

where the last equality holds since $\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{H}_t=(\boldsymbol{N}_t,i),\boldsymbol{x}_{0:t}\right)=p_{ij}$, from the definition of $\{\boldsymbol{H}_t\}_{t\in\mathbb{N}}$ and since $\boldsymbol{H}_{t+1}$ is independent of $\boldsymbol{x}_{0:t}$ given $\boldsymbol{H}_t$. Now, noting that $\boldsymbol{x}_{t+1:T}$ is independent of $\boldsymbol{H}_t$ given $\boldsymbol{H}_{t+1}$ and $\boldsymbol{x}_{0:t}$, then the right hand side of (3.26) equals

$$\begin{aligned}
&\sum_{j\in\mathcal{S}} \frac{\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})p_{ij}}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t})} f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{H}_{t+1},\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|(\boldsymbol{N},j),\boldsymbol{x}_{0:t}), \\
&= \sum_{j\in\mathcal{S}} \frac{\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})p_{ij}}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}\left(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t}\right)} \frac{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T},\boldsymbol{H}_{t+1}|\boldsymbol{X}_{0:t}}\left(\boldsymbol{x}_{t+1:T},\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:t}\right)}{\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:t}\right)} \\
&= \sum_{j\in\mathcal{S}} \frac{\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})p_{ij}}{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t})} \frac{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{t+1:T}|\boldsymbol{X}_{0:t}}(\boldsymbol{x}_{t+1:T}|\boldsymbol{x}_{0:t})\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:T}\right)}{\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:t}\right)} \\
&= \sum_{j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t=(\boldsymbol{N}_t,i)|\boldsymbol{x}_{0:t})p_{ij}\frac{\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:T}\right)}{\mathbb{P}^{\boldsymbol{\theta}}\left(\boldsymbol{H}_{t+1}=(\boldsymbol{N},j)\,|\boldsymbol{x}_{0:t}\right)} \\
&= \sum_{j\in\mathcal{S}} \widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)p_{ij}\frac{\gamma^{(t)}_{\boldsymbol{N}}(j)}{a^{(t)}_{\boldsymbol{N}}(j)}.
\end{aligned}$$

Writing out $\boldsymbol{N}$ explicitly for the two cases, then

$$\gamma^{(t)}_{\boldsymbol{N}_t}(i) = \begin{cases} \sum\limits_{j\in\mathcal{S}} \widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)p_{ij}\dfrac{\gamma^{(t)}_{\boldsymbol{N}_t+\mathbf{1}}(j)}{a^{(t)}_{\boldsymbol{N}_t+\mathbf{1}}(j)}, & \text{for } i\in\mathcal{S}^c_{AR}, \\ \sum\limits_{j\in\mathcal{S}} \widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)p_{ij}\dfrac{\gamma^{(t)}_{\boldsymbol{N}_t^{(-i)}+\mathbf{1}}(j)}{a^{(t)}_{\boldsymbol{N}_t^{(-i)}+\mathbf{1}}(j)}, & \text{for } i\in\mathcal{S}_{AR}. \end{cases}$$

Now, to prove the complexity result, first consider t fixed. We need to calculate the ratio $\frac{\gamma^{(t+1)}_{\boldsymbol{N}_{t+1}}(j)}{a^{(t+1)}_{\boldsymbol{N}_{t+1}}(j)}$ for every corresponding $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $j \in \mathcal{S}$. This costs $M|\mathcal{S}^{(t)}|$ multiplications. This quantity is independent of i, thus only needs to be done once for a given t if we save the resulting quantities.

Now consider t, i and $\boldsymbol{N}_t$ fixed. The calculation $p_{ij}\frac{\gamma^{(t+1)}_{\boldsymbol{N}_{t+1}}(j)}{a^{(t+1)}_{\boldsymbol{N}_{t+1}}(j)}$ is done for every $j \in \mathcal{S}$ which costs M multiplications (the division has already been executed and saved). The sum over $j \in \mathcal{S}$ results in a single term, which is then multiplied by the corresponding $\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)$, and this costs 1 multiplication. So, for fixed t, i and $\boldsymbol{N}_t$, (assuming the ratio has already been calculated and saved) we require $M+1$ multiplications. We do this for all $i \in \mathcal{S}$ and $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ which costs $(M+1)M|\mathcal{S}^{(t)}|$ multiplications.

So, for a given t we execute $M|\mathcal{S}^{(t)}| + (M+1)M|\mathcal{S}^{(t)}|$ multiplications. This is done for every $t = 0, \ldots, T-1$, so the total number of multiplications is

$$\begin{aligned}
&\sum_{t=0}^{T-1} M|\mathcal{S}^{(t)}| + (M+1)M|\mathcal{S}^{(t)}| \\
&= (M^2+2M)\sum_{t=0}^{T-1}|\mathcal{S}^{(t)}| \leq (M^2+2M)\frac{T(T-1)^k k^{k-1}(k+1)}{(k-1)!} \\
&= \mathcal{O}(M^2 T^{k+1} k^k),
\end{aligned}$$

where we have used similar arguments as those in the proof of Lemma 3.2 to bound the complexity. □

Remark 3.3. The smoothed probabilities $\mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T})$ can be used to calculate the probabilities

$$\mathbb{P}^{\boldsymbol{\theta}}(R_t = i, N_{t,i} = \ell|\boldsymbol{x}_{0:T}) = \sum_{\substack{\boldsymbol{N}_t \in \mathcal{S}^{(t)}: \\ N_{t,i}=\ell}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T}),$$

which are used in the EM algorithm in Section 3.4, and

$$\mathbb{P}^{\boldsymbol{\theta}}(R_t = i|\boldsymbol{x}_{0:T}) = \sum_{\boldsymbol{N}_t \in \mathcal{S}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T}),$$

which may also be of interest.

3.4 The EM algorithm for independent regimes models

In Sections 3.2 and 3.3 we presented new forward and backward algorithms to calculate the probabilities $\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{H}_t = (\boldsymbol{N}_t, i)|\boldsymbol{x}_{0:T})$, which can be used to calculate the smoothed probabilities $\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = \ell|\boldsymbol{x}_{0:T})$ as in Remark 3.3. Here we show how these probabilities can be used to implement an exact, computationally feasible EM algorithm for MRS models with independent regimes.

3.4.1 The E-step

Recall that the EM algorithm is an iterative procedure, alternating between an expectation step and a maximisation step. In the expectation step the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ is constructed as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) = \mathbb{E}[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T},\boldsymbol{R}}(\boldsymbol{x}_{0:T}, \boldsymbol{R})|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n] \tag{3.27}$$

$$= \mathbb{E}[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T},\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}, \boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n], \tag{3.28}$$

where $\boldsymbol{R} = (R_0, ..., R_T)$ is a sequence of the hidden Markov chain $\{R_t\}$, and $\boldsymbol{H}_0, ..., \boldsymbol{H}_T$ is a sequence of the corresponding augmented hidden process $\{\boldsymbol{H}_t\}$. The information contained in the sequences $\boldsymbol{R}$ and $\boldsymbol{H}_0, ..., \boldsymbol{H}_T$ is entirely equivalent but we opt for the latter representation to remain consistent with, and emphasise the place of, the work in the previous sections. In the M-step of the algorithm, the maximisers $\arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ are found. To describe the EM algorithm for MRS models with independent regimes recall our notation from our forward and backward algorithms in Sections 3.2 and 3.3.

First observe that, for MRS models, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ can be written as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) \tag{3.29}$$

$$= \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T},\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}, \boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right]$$

$$= \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}|\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}|\boldsymbol{H}_0, ..., \boldsymbol{H}_T) + \log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right]$$

$$= \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}|\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}|\boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right] + \mathbb{E}\left[\log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right]. \tag{3.30}$$

Now, using the augmented hidden Markov chain, $\{\boldsymbol{H}_t\}_{t\in\mathbb{N}}$ from Equation (3.5), Equation (3.30) can be written in such a way that the function Q is computationally feasible. First note that, given $\boldsymbol{H}_t$, x_t is independent of $\boldsymbol{H}_\tau$ for $\tau \neq t$ which, along with the definition of conditional densities, allows the function $\log\{f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}|\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}|\boldsymbol{H}_0, ..., \boldsymbol{H}_T)\}$ to be

written as

$$
\begin{aligned}
&\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}|\boldsymbol{H}_0,...,\boldsymbol{H}_T}(\boldsymbol{x}_{0:T}|\boldsymbol{H}_0,...,\boldsymbol{H}_T) \\
&= \log\left\{ f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|\boldsymbol{H}_0) \prod_{t=1}^{T} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{H}_t,\boldsymbol{x}_{0:t-1}) \right\} \\
&= \log f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|\boldsymbol{H}_0) + \sum_{t=1}^{T} \log f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|\boldsymbol{H}_t,\boldsymbol{x}_{0:t-1}) \\
&= \log\left\{ \prod_{j\in\mathcal{S}} \prod_{\boldsymbol{N}_0\in\mathcal{S}^{(0)}} f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0,j))^{\mathbb{I}(\boldsymbol{H}_0=(\boldsymbol{N}_0,j))} \right\} \\
&\quad + \sum_{t=1}^{T} \log\left\{ \prod_{j\in\mathcal{S}} \prod_{\boldsymbol{N}_t\in\mathcal{S}^{(t)}} f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t,j),\boldsymbol{x}_{0:t-1})^{\mathbb{I}(\boldsymbol{H}_t=(\boldsymbol{N}_t,j))} \right\} \\
&= \sum_{j\in\mathcal{S}} \sum_{\boldsymbol{N}_0\in\mathcal{S}^{(0)}} \mathbb{I}(\boldsymbol{H}_0=(\boldsymbol{N}_0,j)) \log f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_0,j)) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}} \sum_{\boldsymbol{N}_t\in\mathcal{S}^{(t)}} \mathbb{I}(\boldsymbol{H}_t=(\boldsymbol{N}_t,j)) \log f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{X}_{0:t-1}}(x_t|(\boldsymbol{N}_t,j),\boldsymbol{x}_{0:t-1}).
\end{aligned}
$$

Lastly, since $f^{\boldsymbol{\theta}}_{X_t|\boldsymbol{H}_t,\boldsymbol{x}_{0:t-1}}(x_t|(\boldsymbol{N}_t,j),\boldsymbol{x}_{0:t-1}) = f^{\boldsymbol{\theta}}_{X_t|N_{t,j},R_t,\boldsymbol{X}_{0:t-1}}(x_t|m,j,\boldsymbol{x}_{0:t-1})$, and similarly for $f^{\boldsymbol{\theta}}_{X_0|\boldsymbol{H}_0}(x_0|(\boldsymbol{N}_{t,j},j))$, this expression simplifies to

$$
\begin{aligned}
&\sum_{j\in\mathcal{S}} \sum_{m\in\mathcal{L}^{(0)}} \mathbb{I}(N_{0,j}=m,R_0=j) \log f^{\boldsymbol{\theta}}_{X_0|N_{0,j},R_0}(x_0|m,j) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}} \sum_{m\in\mathcal{L}^{(t)}} \mathbb{I}(N_{t,j}=m,R_t=j) \log f^{\boldsymbol{\theta}}_{X_t|N_{t,j},R_t,\boldsymbol{X}_{0:t-1}}(X_t|m,j,\boldsymbol{x}_{0:t-1}). \qquad (3.31)
\end{aligned}
$$

Taking the expectation of (3.31) with respect to the distribution $f^{\boldsymbol{\theta}_n}_{\boldsymbol{H}_0,...,\boldsymbol{H}_T|\boldsymbol{X}_{0:T}}$ (equivalently the distribution $f^{\boldsymbol{\theta}_n}_{\boldsymbol{R}|\boldsymbol{X}_{0:T}}$) gives

$$
\begin{aligned}
&\mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T},\boldsymbol{H}_0,...,\boldsymbol{H}_T|\boldsymbol{X}_{0:T}}(\boldsymbol{x}_{0:T},\boldsymbol{H}_0,...,\boldsymbol{H}_T)|\boldsymbol{x}_{0:T};\boldsymbol{\theta}_n\right] \\
&= \sum_{j\in\mathcal{S}_{AR}} \sum_{m\in\mathcal{L}^{(0)}} \mathbb{P}^{\boldsymbol{\theta}_n}(N_{0,j}=m,R_0=j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_0|N_{0,j},R_0}(x_0|m,j) \\
&\quad + \sum_{j\in\mathcal{S}^c_{AR}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0=j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}_{AR}} \sum_{m\in\mathcal{L}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(N_{t,j}=m,R_t=j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_t|N_{t,j},R_t,\boldsymbol{X}_{0:t-1}}(x_t|m,j,\boldsymbol{x}_{0:t-1}) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}^c_{AR}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t=j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_t|R_t,\boldsymbol{X}_{0:t-1}}(x_t|j,\boldsymbol{x}_{0:t-1}).
\end{aligned}
$$

Using similar arguments $\mathbb{E}\left[\log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right]$ is found to be

$$\begin{aligned}
\mathbb{E}\left[\log \mathbb{P}^{\boldsymbol{\theta}}(\boldsymbol{H}_0, ..., \boldsymbol{H}_T)\middle|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] &= \mathbb{E}\left[\log\left\{\prod_{i\in\mathcal{S}} \pi_i^{\mathbb{I}(R_0=i)} \prod_{i,j\in\mathcal{S}} p_{ij}^{\eta_{ij}}\right\}\middle|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{S}} \mathbb{I}(R_0 = i)\log \pi_i + \sum_{i,j\in\mathcal{S}} \eta_{ij}\log p_{ij}\middle|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] \\
&= \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = i|\boldsymbol{x}_{0:T})\log \pi_i + \sum_{i,j\in\mathcal{S}} \mathbb{E}\left[\eta_{ij}|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right]\log p_{ij},
\end{aligned}$$

where η_{ij} is the number of transitions from state $R_{t-1} = i$ to state $R_t = j$ in the sequence $\boldsymbol{R} = (R_0, R_1, ..., R_T)$. The expectation $\mathbb{E}\left[\eta_{ij}|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right]$ can be calculated as

$$\begin{aligned}
\mathbb{E}\left[\eta_{ij}|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(R_{t-1} = i, R_t = j)\middle|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{I}(R_{t-1} = i, R_t = j)|\boldsymbol{x}_{0:T}, \boldsymbol{\theta}_n\right] \\
&= \sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1} = i, R_t = j|\boldsymbol{x}_{0:T}\right).
\end{aligned}$$

So, the function Q is

$$\begin{aligned}
&Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n) \\
&= \mathbb{E}\left[\log f^{\boldsymbol{\theta}}_{\boldsymbol{X}_{0:T}, \boldsymbol{H}_0, ..., \boldsymbol{H}_T|\boldsymbol{X}_{0:T}}(\boldsymbol{x}_{0:T}, \boldsymbol{H}_0, ..., \boldsymbol{H}_T)|\boldsymbol{x}_{0:T}; \boldsymbol{\theta}_n\right] \\
&= \sum_{j\in\mathcal{S}_{AR}} \sum_{m\in\mathcal{L}^{(0)}} \mathbb{P}^{\boldsymbol{\theta}_n}(N_{0,j} = m, R_0 = j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_0|N_{0,j}, R_0}(x_0|m, j) \\
&\quad + \sum_{j\in\mathcal{S}^c_{AR}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_0|R_0}(x_0|j) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}_{AR}} \sum_{m\in\mathcal{L}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(N_{t,j} = m, R_t = j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_t|N_{t,j}, R_t, \boldsymbol{X}_{0:t-1}}(x_t|m, j, \boldsymbol{x}_{0:t-1}) \\
&\quad + \sum_{t=1}^{T} \sum_{j\in\mathcal{S}^c_{AR}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{x}_{0:T}) \log f^{\boldsymbol{\theta}}_{X_t|R_t, \boldsymbol{X}_{0:t-1}}(x_t|j, \boldsymbol{x}_{0:t-1}) \\
&\quad + \sum_{i\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = i|\boldsymbol{x}_{0:T}) \log \pi_i + \sum_{i,j\in\mathcal{S}} \sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1} = i, R_t = j|\boldsymbol{x}_{0:T}\right) \log p_{ij}. \qquad (3.32)
\end{aligned}$$

Lemma 3.4. *The joint probabilities are given by*

$$\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1} = i, R_t = j|\boldsymbol{x}_{0:T}\right) = \mathbb{P}^{\boldsymbol{\theta}_n}\left(N_{t,i} = 1, R_t = j|\boldsymbol{x}_{0:T}\right)$$

when $i \in \mathcal{S}_{AR}$, *and*

$$\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i, R_t=j|\boldsymbol{x}_{0:T}\right)$$
$$= \sum_{\boldsymbol{n}_t-\mathbf{1}\in\mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=j, \boldsymbol{N}_t=\boldsymbol{n}_t|\boldsymbol{x}_{0:T}\right) \frac{p_{ij}^{(n)}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i, \boldsymbol{N}_{t-1}=\boldsymbol{n}_t-\mathbf{1}|\boldsymbol{x}_{0:t-1}\right)}{\sum\limits_{\ell\in\mathcal{S}_{AR}^c} p_{\ell j}^{(n)}\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1}=\ell, \boldsymbol{N}_{t-1}=\boldsymbol{n}_t-\mathbf{1}|\boldsymbol{x}_{0:t-1})} \tag{3.33}$$

when $i \in \mathcal{S}_{AR}^c$.

Proof. This proof follows similar arguments to those in Kim [67] which develops algorithms for MRS models of Type I (with dependent regimes).

For the case $i \in \mathcal{S}_{AR}$, note that $N_{t,i} = 1$ if and only if $R_{t-1} = i$ and we are done.

When $i \in \mathcal{S}_{AR}^c$ all counters in $\boldsymbol{N}_t$ are different from 1, so $\boldsymbol{N}_t - \mathbf{1} \in \mathcal{S}^{(t-1)}$. Thus

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i, R_t=j|\boldsymbol{x}_{0:T}\right)\\
&= \sum_{\boldsymbol{n}_t-\mathbf{1}\in\mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}\left(\boldsymbol{N}_t=\boldsymbol{n}_t, R_{t-1}=i, R_t=j|\boldsymbol{x}_{0:T}\right)\\
&= \sum_{\boldsymbol{n}_t-\mathbf{1}\in\mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}\left(\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j|\boldsymbol{x}_{0:T}\right)\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:T}\right)\\
&= \sum_{\boldsymbol{n}_t-\mathbf{1}\in\mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}\left(\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j|\boldsymbol{x}_{0:T}\right)\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:t-1}\right).
\end{aligned} \tag{3.34}$$

The last equality uses

$$\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:T}\right) = \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:t-1}\right),$$

which is not obvious, but holds since, given R_t and $\boldsymbol{N}_t$, then $\boldsymbol{x}_{t:T}$ is independent of R_{t-1}. Focusing on the term $\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:t-1}\right)$ in Equation (3.34),

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j, \boldsymbol{x}_{0:t-1}\right)\\
&= \frac{\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=j|\boldsymbol{N}_t=\boldsymbol{n}_t, R_{t-1}=i, \boldsymbol{x}_{0:t-1}\right)\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, \boldsymbol{x}_{0:t-1}\right)}{\mathbb{P}^{\boldsymbol{\theta}_n}(R_t=j|\boldsymbol{N}_t=\boldsymbol{n}_t, \boldsymbol{x}_{0:t-1})}\\
&= \frac{\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=j|R_{t-1}=i\right)\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_{t-1}=i|\boldsymbol{N}_t=\boldsymbol{n}_t, \boldsymbol{x}_{0:t-1}\right)}{\mathbb{P}^{\boldsymbol{\theta}_n}(R_t=j|\boldsymbol{N}_t=\boldsymbol{n}_t, \boldsymbol{x}_{0:t-1})},\\
&= \frac{p_{ij}^{(n)}\mathbb{P}^{\boldsymbol{\theta}_n}\left(\boldsymbol{N}_t=\boldsymbol{n}_t, R_{t-1}=i|\boldsymbol{x}_{0:t-1}\right)}{\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t=\boldsymbol{n}_t, R_t=j|\boldsymbol{x}_{0:t-1})},
\end{aligned}$$

where $\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|R_{t-1} = i) =: p_{ij}^{(n)}$ is the p_{ij} parameter in $\boldsymbol{\theta}_n$. In the second equality we have used the fact that $\mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|\boldsymbol{N}_t = \boldsymbol{n}_t, R_{t-1} = i, \boldsymbol{x}_{0:t-1}) = \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j|R_{t-1} = i)$, which holds since, given R_{t-1}, then R_t is independent of $\boldsymbol{N}_t$ and $\boldsymbol{x}_{0:t-1}$. Now, notice that

$$\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t = \boldsymbol{n}_t, R_{t-1} = i|\boldsymbol{x}_{0:t-1}) = \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_{t-1} = \boldsymbol{n}_t - \mathbf{1}, R_{t-1} = i|\boldsymbol{x}_{0:t-1}),$$

since $i \in \mathcal{S}_{AR}^c$. Thus, continuing from the right hand side of Equation (3.34),

$$\begin{aligned}
&\sum_{\boldsymbol{n}_t - \mathbf{1} \in \mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t = \boldsymbol{n}_t, R_t = j|\boldsymbol{x}_{0:T})\, \mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1} = i|\boldsymbol{N}_t = \boldsymbol{n}_t, R_t = j, \boldsymbol{x}_{0:t-1}) \\
&= \sum_{\boldsymbol{n}_t - \mathbf{1} \in \mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t = \boldsymbol{n}_t, R_t = j|\boldsymbol{x}_{0:T}) \frac{p_{ij}^{(n)} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_{t-1} = \boldsymbol{n}_t - \mathbf{1}, R_{t-1} = i|\boldsymbol{x}_{0:t-1})}{\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t = \boldsymbol{n}_t, R_t = j|\boldsymbol{x}_{0:t-1})} \\
&= \sum_{\boldsymbol{n}_t - \mathbf{1} \in \mathcal{S}^{(t-1)}} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_t = \boldsymbol{n}_t, R_t = j|\boldsymbol{x}_{0:T}) \frac{p_{ij}^{(n)} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_{t-1} = \boldsymbol{n}_t - \mathbf{1}, R_{t-1} = i|\boldsymbol{x}_{0:t-1})}{\sum_{k \in \mathcal{S}_{AR}} p_{kj}^{(n)} \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_{t-1} = \boldsymbol{n}_t - \mathbf{1}, R_{t-1} = k|\boldsymbol{x}_{0:t-1})}.
\end{aligned}$$

The sum in the denominator is over $k \in \mathcal{S}_{AR}$ since only when $k \in \mathcal{S}_{AR}$ is $\boldsymbol{N}_t - \mathbf{1}$ defined. □

Remark 3.4. Care should be taken when calculating (3.33) in Lemma 3.4 since the probabilities $\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{N}_{t-1} = \boldsymbol{n}_t - \mathbf{1}, R_{t-1} = k|\boldsymbol{x}_{0:t-1})$ can be small and computational errors may occur.

3.4.2 The M-step

Next, the maximisers, $\boldsymbol{\theta}_{n+1} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$, are derived. Conveniently, for the parameters p_{ij}, $i, j \in \mathcal{S}$, we can use the work of Hamilton [45],

$$p_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = j, R_{t-1} = i|\boldsymbol{x}_{0:T})}{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1} = i|\boldsymbol{x}_{0:T})}.$$

However note that to get this analytic update for the $p_{ij}^{(n+1)}$ parameters, terms involving π_j in Equation (3.32) have been treated as if they are unrelated to p_{ij}, $i, j \in \mathcal{S}$, which is not true when π_j is specified as the stationary distribution of the process $\{R_t\}$, but holds for other cases, such as when π_j is some predetermined distribution, or when the π_js are specified as seperate parameters to be inferred. Nonetheless, this simplification is appropriate if we assume that, as the sample size grows, the contribution of terms involving R_0 become insignificant.

M-step for i.i.d. regimes The updates for parameters of the i.i.d. regimes can often be found analytically too.

Lemma 3.5. *Suppose Regime i is i.i.d. $N(\mu_i, \sigma_i^2)$. The M-step updates $\mu_i^{(n+1)}$ and $\left(\sigma_i^{(n+1)}\right)^2$ are found to be*

$$\mu_i^{(n+1)} = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) x_t}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})},$$
$$\left(\sigma_i^{(n+1)}\right)^2 = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \left(x_t - \mu_i^{(n+1)}\right)^2}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}.$$

Proof. Differentiate Q given by Equation (3.32) with respect to μ_i and find when the derivative is zero:

$$\frac{\partial Q}{\partial \mu_i} = \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) (x_t - \mu_i) = 0,$$

which leads to

$$\mu_i = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) x_t}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}. \tag{3.35}$$

The second derivative test shows that (3.35) is a maximiser. Now differentiate Q with respect to σ_i^2 and find when the derivative is zero:

$$\frac{\partial Q}{\partial \sigma_i^2} = \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \left(\frac{1}{2\sigma_i^2} - \frac{1}{2\sigma_i^4} (x_t - \mu_i)^2 \right) = 0.$$

Thus,

$$\sigma_i^2 = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) (x_t - \mu_i)^2}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}. \tag{3.36}$$

Now, to show that (3.36) is indeed a global maximum, define $\widehat{\sigma}_i^2$ to be the value in (3.36) and $\tilde{\sigma}_i^2 = \widehat{\sigma}_i^2 + \varepsilon$ for some $\varepsilon \in \mathbb{R}$. The goal is to show $Q(\widehat{\boldsymbol{\theta}}; \boldsymbol{\theta}_n) - Q(\widehat{\boldsymbol{\theta}} + \varepsilon \boldsymbol{e}_{\sigma_i}; \boldsymbol{\theta}_n) > 0$ for any $\varepsilon \neq 0$, where $\widehat{\boldsymbol{\theta}}$ is any parameter vector with $\sigma_i = \widehat{\sigma}_i$ and $\boldsymbol{e}_{\sigma_i}$ is a vector of zeros with 1 in the position corresponding to σ_i. Since the only terms in Q involving σ_i are

the terms involving the density of Regime i,

$$
\begin{aligned}
&Q(\widehat{\boldsymbol{\theta}};\boldsymbol{\theta}_n) - Q(\widehat{\boldsymbol{\theta}} + \boldsymbol{e}_{\sigma_i}\varepsilon;\boldsymbol{\theta}_n) \\
&= \sum_{t=0}^{T} \sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \left[\log\left\{\frac{1}{\sqrt{2\pi\widehat{\sigma}_i^{\,2}}} e^{\frac{-1}{2\widehat{\sigma}_i^2}(x_t-\mu_i)^2}\right\}\right] \\
&\quad - \sum_{t=0}^{T} \sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \left[\log\left\{\frac{1}{\sqrt{2\pi\tilde{\sigma}_i^{\,2}}} e^{\frac{-1}{2\tilde{\sigma}_i^{\,2}}(x_t-\mu_i)^2}\right\}\right] \\
&= \sum_{t=0}^{T} \sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \left[\frac{1}{2}\log\left\{\frac{\tilde{\sigma}_i^{\,2}}{\widehat{\sigma}_i^{\,2}}\right\} - \frac{1}{2}(x_t-\mu_i)^2\left(\frac{1}{\widehat{\sigma}_i^{\,2}} - \frac{1}{\tilde{\sigma}_i^{\,2}}\right)\right] \\
&= \frac{1}{2}\log\left\{\frac{\tilde{\sigma}_i^{\,2}}{\widehat{\sigma}_i^{\,2}}\right\} \sum_{t=0}^{T} \sum_{m\in\mathcal{L}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \\
&\quad - \frac{1}{2}\left(\frac{1}{\widehat{\sigma}_i^{\,2}} - \frac{1}{\tilde{\sigma}_i^{\,2}}\right) \sum_{t=0}^{T} \sum_{m\in\mathcal{L}^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T})\,(x_t-\mu_i)^2 .
\end{aligned}
$$

Multiplying the second sum by $\dfrac{\sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T})}{\sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T})}$, and recalling the definition of $\widehat{\sigma}_i^{\,2}$ gives

$$
\begin{aligned}
&\frac{1}{2}\log\left\{\frac{\tilde{\sigma}_i^{\,2}}{\widehat{\sigma}_i^{\,2}}\right\} \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \\
&\quad - \frac{1}{2}\left(\frac{1}{\widehat{\sigma}_i^{\,2}} - \frac{1}{\tilde{\sigma}_i^{\,2}}\right) \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})\,(x_t-\mu_i)^2}{\sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T})} \sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T}) \\
&= \frac{1}{2}\log\left\{\frac{\tilde{\sigma}_i^{\,2}}{\widehat{\sigma}_i^{\,2}}\right\} \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) - \frac{1}{2}\left(\frac{1}{\widehat{\sigma}_i^{\,2}} - \frac{1}{\tilde{\sigma}_i^{\,2}}\right)\widehat{\sigma}_i^2 \sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T}) \\
&= \frac{1}{2}\log\left\{\frac{\tilde{\sigma}_i^{\,2}}{\widehat{\sigma}_i^{\,2}}\right\} \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) - \frac{1}{2}\left(1 - \frac{\widehat{\sigma}_i^2}{\tilde{\sigma}_i^{\,2}}\right) \sum_{s=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_s = i|\boldsymbol{x}_{0:T}) \\
&\geq \frac{1}{2}\left(1 - \frac{\widehat{\sigma}_i^2}{\tilde{\sigma}_i^{\,2}}\right) \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) - \frac{1}{2}\left(1 - \frac{\widehat{\sigma}_i^2}{\tilde{\sigma}_i^{\,2}}\right) \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}), \\
&= 0,
\end{aligned}
$$

where the last inequality holds since $1 - \frac{1}{y} \leq \log\{y\}$ with equality if and only if $y = 1$. Thus (3.36) is a maximiser. □

Lemma 3.6. *Suppose Regime i follows i.i.d. shifted-log-normal dynamics, that is, if X_t is from regime i, then $\log(X_t - q_i) \sim N(\mu_i, \sigma_i^2)$, and suppose the parameter q_i is known.*

The M-step updates $\mu_i^{(n+1)}(q_i)$ *and* $\left(\sigma_i^{(n+1)}\right)^2$ *are found to be*

$$\mu_i^{(n+1)} = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \log(x_t - q_i)}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})},$$

$$\left(\sigma_i^{(n+1)}\right)^2 = \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \left(\log(x_t - q_i) - \mu_i^{(n+1)}\right)^2}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}.$$

Proof. The proof is similar to the proof of Lemma 3.5. □

Lemma 3.7. *Suppose Regime* i *follows an i.i.d. shifted-Gamma distribution, that is, if* X_t *is from Regime* i*, then* $(X_t - q_i) \sim Gamma(\mu_i, \sigma_i^2)$*, and suppose the parameter* q_i *is known. The M-step update for the scale parameter* $\left(\sigma_i^{(n+1)}\right)^2$ *as a function of* μ_i *is*

$$\left(\sigma_i^{(n+1)}(\mu_i)\right)^2 = \mu_i \frac{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})(x_t - q_i)}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}.$$

The update for μ_i *is then found by finding*

$$\begin{aligned}\mu_i^{(n+1)} = \underset{\mu\in(0,\infty)}{\arg\max}\Bigg\{ &- \mu \log\left(\sigma_i^{(n+1)}(\mu)\right)^2 \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \\ &- \log\Gamma(\mu) \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \\ &+ (\mu - 1) \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \log(x_t - q_i) \\ &- \mu \sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T}) \Bigg\},\end{aligned}$$

where $\Gamma(\cdot)$ *is the Gamma function.*

Proof. The result follows after differentiating Q with respect to σ_i^2, and solving for the stationary point, which is a maximum by the second derivative test. □

Note Lemmas 3.6 and 3.7 assume the parameter q_i is known. This is necessary for the shifted-log-normal distribution, and the shifted-Gamma distribution when the shape parameter μ_i is less than 1. We elaborate on this more in Section 3.5.3.

M-step for AR(1) regimes of MRS models of Type III Consider an MRS model with independent regimes and let Regime i be an AR(1) regime, with

$$B_\tau^{(i)} = \alpha_i + \phi_i B_{\tau-1}^{(i)} + \sigma_i \varepsilon_\tau^{(i)},$$

for $\tau \in \mathbb{N}$. Suppose that $\{B_\tau^{(i)}\}$ only evolves when it is observed; that is, $\{B_\tau^{(i)}\}$ only evolves at times where the hidden Markov chain, $\{R_t\}$, is in state i. The EM algorithm is most useful when analytic E- and M-steps can be derived. To obtain analytic updates for parameters of models of this type, a slight simplification of the function Q is required: we suppose that

$$f^{\boldsymbol{\theta}}_{X_t|N_{t,i},R_t,\boldsymbol{X}_{0:t-1}}(x_t|\Delta_i, i, \boldsymbol{x}_{0:t-1}) = g_i(x_t), \tag{3.37}$$

for all $\boldsymbol{\theta} \in \Theta$. That is, the first observation from each regime has the same density for all possible parameter values, which allows us to ignore these terms when finding the M-step updates. The benefit of this assumption stems from the fact that for $N_{t,i} = \Delta_i$, the density $f^{\boldsymbol{\theta}}_{X_t|N_{t,i},R_t,\boldsymbol{X}_{0:t-1}}(x_t|\Delta_i, i, \boldsymbol{x}_{0:t-1})$ is the stationary distribution of the process in Regime i, whereas when $N_{t,i} \in \{1, 2, ..., t\}$, then the densities $f^{\boldsymbol{\theta}}(x_t|\Delta_i, i, \boldsymbol{x}_{0:t-1})$ have the same form and therefore closed expressions for the M-step updates can be derived. The same benefit is not achieved for MRS models of Type II since no closed form updates are available, even with the assumption in Equation (3.37). As the number of observations, T, gets large, we expect this assumption will have a vanishing impact on the shape of the likelihood, and therefore the MLEs will be unaffected, asymptotically.

Lemma 3.8. *Suppose the simplification in Equation (3.37) holds, and Regime i is an AR(1) regime of a MRS model of Type III, then the updates in the M-step of the EM algorithm are*

$$\alpha_i^{(n+1)} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right) x_t \left(A_2 - A_1 x_{t-m}\right)}{A_2 \sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right) - A_1^2}, \tag{3.38}$$

$$\text{where } A_1 = \sum_{s=1}^{T}\sum_{\ell=1}^{s} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_s = i, N_{s,i} = \ell|\boldsymbol{x}_{0:T}\right) x_{s-\ell}, \tag{3.39}$$

$$\text{and } A_2 = \sum_{s=1}^{T}\sum_{\ell=1}^{s} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_s = i, N_{s,i} = \ell|\boldsymbol{x}_{0:T}\right) x_{s-\ell}^2, \tag{3.40}$$

$$\phi_i^{(n+1)} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right) x_t x_{t-m} - \alpha_i^{(n+1)} A_1}{A_2}, \tag{3.41}$$

$$\left(\sigma_i^{(n+1)}\right)^2 = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right) \left(x_t - \alpha_i^{(n+1)} - \phi_i^{(n+1)} x_{t-m}\right)^2}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)}. \tag{3.42}$$

Proof. Differentiating Q, given by Equation (3.32) and conditional densities

$$f^{\boldsymbol{\theta}_n}_{X_t|R_t,N_{t,j},\boldsymbol{X}_{0:t-1}}(x_t|j,m,\boldsymbol{x}_{0:t-1})$$

given by Equation (3.17), with respect to α_i, ϕ_i and σ_i^2 gives

$$\begin{aligned}
\frac{\partial Q}{\partial \alpha_i} &= \sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(\frac{1}{\sigma_i^2}\left(x_t - \alpha_i - \phi_i x_{t-m}\right)\right),\\
\frac{\partial Q}{\partial \phi_i} &= \sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(\frac{1}{\sigma_i^2}\left(x_t - \alpha_i - \phi_i x_{t-m}\right)x_{t-m}\right),\\
\frac{\partial Q}{\partial \sigma_i^2} &= \sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(\frac{-1}{2\sigma_i^2} + \frac{\left(x_t - \alpha_i - \phi_i x_{t-m}\right)^2}{2\sigma_i^4}\right).
\end{aligned}$$

Setting these derivatives to zero and solving gives the following system of equations

$$\alpha_i = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(x_t - \phi_i x_{t-m}\right)}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)}, \tag{3.43}$$

$$\phi_i = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(x_t - \alpha_i\right)x_{t-m}}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)x_{t-m}^2}, \tag{3.44}$$

$$\sigma_i^2 = \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(x_t - \alpha_i - \phi_i x_{t-m}\right)^2}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)}. \tag{3.45}$$

Substituting Equation (3.44) into Equation (3.43) gives

$$\begin{aligned}
\alpha_i &= \frac{\sum_{t=1}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)x_t - \phi_i A_1}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)}\\
&= \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)x_t}{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)}\\
&\quad - \frac{\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)\left(x_t - \alpha_i\right)x_{t-m}A_1}{A_2\sum_{t=1}^{T}\sum_{m=1}^{t} \mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}\right)},
\end{aligned}$$

and multiplying both sides by $\dfrac{\sum_{s=1}^{T}\sum_{\ell=1}^{s}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_s=i,N_{s,i}=\ell|\boldsymbol{x}_{0:T}\right)}{A_1}$ gives

$$\begin{aligned}
&\alpha_i\frac{\sum_{s=1}^{T}\sum_{\ell=1}^{s}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_s=i,N_{s,i}=\ell|\boldsymbol{x}_{0:T}\right)}{A_1}\\
&=\frac{\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)x_t}{A_1}\\
&\quad-\frac{\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)(x_t-\alpha_i)\,x_{t-m}}{A_2}.
\end{aligned}$$

Then rearrange to get α_i on the left hand side and a common denominator,

$$\begin{aligned}
&\alpha_i\frac{\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)(A_2-A_1x_{t-m})}{A_1A_2}\\
&=\frac{\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)x_t(A_2-A_1x_{t-m})}{A_1A_2},
\end{aligned}$$

and thus

$$\alpha_i=\frac{\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)x_t(A_2-A_1x_{t-m})}{A_2\sum_{t=1}^{T}\sum_{m=1}^{t}\mathbb{P}^{\boldsymbol{\theta}_n}\left(R_t=i,N_{t,i}=m|\boldsymbol{x}_{0:T}\right)-A_1^2},$$

which is Equation (3.38). Equations (3.41) and (3.42) follow after solving (3.44) and (3.45) by substitution. The fact that α_i and ϕ_i are maximisers can be proved by the second derivative test. To show (3.45) is a maximiser, a similar argument to the independent case can be used (see the proof of Lemma 3.5). □

M-step for AR(1) regimes of MRS models of Type II Consider now an MRS model with independent regimes with AR(1) regimes which evolve at all time points. Let Regime i be an AR(1) process, $B_t^{(i)}=\alpha_i+\phi_iB_{t-1}^{(i)}+\sigma_i\varepsilon_t^{(i)}$. Here, things do not work out as nicely as they do for MRS models of Type III. The näive way to proceed would be to find the maximisers of Q numerically in 3-dimensional space, $(\alpha_i,\phi_i,\sigma_i^2)$. However, the dimension of the optimisation can be reduced by deriving expressions for $\alpha_i^{(n+1)}$ and $\sigma_i^{(n+1)}$ in terms of $\phi_i^{(n+1)}$.

Lemma 3.9. *If Regime i is an AR(1) regime of a MRS model of Type II, the M-step of the EM algorithm can be executed by the following. The updates $\alpha_i^{(n+1)}$ and $\left(\sigma_i^{(n+1)}\right)^2$*

as functions of $\phi_i^{(n+1)}$ *are*

$$\alpha_i^{(n+1)}\left(\phi_i^{(n+1)}\right) = \frac{\sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T})B_{t,m}}{\sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T})A_{t,m}}, \tag{3.46}$$

where

$$A_{t,m} = \left(\frac{1-\left(\phi_i^{(n+1)}\right)^m}{1-\phi_i^{(n+1)}}\right)\left(\frac{1+\left(\phi_i^{(n+1)}\right)}{1+\left(\phi_i^{(n+1)}\right)^m}\right),$$

$$B_{t,m} = \left(x_t - \left(\phi_i^{(n+1)}\right)^m x_{t-m}\right)\frac{1+\left(\phi_i^{(n+1)}\right)}{1+\left(\phi_i^{(n+1)}\right)^m},$$

and

$$\left(\sigma_i^{(n+1)}\left(\phi_i^{(n+1)}\right)\right)^2 = \frac{\sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T})C_{t,m}}{\sum_{t=0}^{T} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i|\boldsymbol{x}_{0:T})}, \tag{3.47}$$

where

$$C_{t,m} = \frac{\left(x_t - \alpha_i^{(n+1)}(\phi_i^{(n+1)})\left(\frac{1-\left(\phi_i^{(n+1)}\right)^m}{1-\phi_i^{(n+1)}}\right) - \left(\phi_i^{(n+1)}\right)^m x_{t-m}\right)^2}{\left(\frac{1-\left(\phi_i^{(n+1)}\right)^{2m}}{1-\left(\phi_i^{(n+1)}\right)^2}\right)}.$$

The M-step update for $\phi_i^{(n+1)}$ *is given by*

$$\phi_i^{(n+1)} = \underset{\phi_i\in(-1,1)}{\arg\max}\, g(\phi_i),$$

where

$$g(\phi_i) := \sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T})L_{t,m}\left(\phi_i, \sigma_i^{(n+1)}\right)$$

and

$$L_{t,m}\left(\phi_i, \sigma_i^{(n+1)}\right) := \frac{1}{2}\log\left\{\frac{1-\phi_i^2}{1-\phi_i^{2m}}\right\} - \log\left\{\sigma_i^{(n+1)}(\phi_i)\right\}.$$

Proof. Differentiate Q in Equation (3.32) with conditional densities $f^{\boldsymbol{\theta}}(x_t|R_t = i, N_{t,i} = m, \boldsymbol{x}_{0:t-1})$ given by Equation (3.19), with respect to α_i and solve for when the derivative is zero:

$$\frac{\partial Q}{\partial \alpha_i} = \sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \frac{\left(x_t - \alpha_i\left(\frac{1-\phi_i^m}{1-\phi_i}\right) - \phi_i^m x_{t-m}\right)\left(\frac{1-\phi_i^m}{1-\phi_i}\right)}{\sigma_i^2\left(\frac{1-\phi_i^{2m}}{1-\phi_i^2}\right)} = 0,$$

which, after some simplifications, gives Equation (3.46), and can be shown to be a maximiser by the second derivative test. Now,

$$\frac{\partial Q}{\partial \sigma_i^2} = \sum_{t=0}^{T}\sum_{m\in\mathcal{L}_i^{(t)}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, N_{t,i} = m|\boldsymbol{x}_{0:T}) \left(\frac{-1}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}\frac{\left(x_t - \alpha_i\left(\frac{1-\phi_i^m}{1-\phi_i}\right) - \phi_i^m x_{t-n}\right)^2}{\left(\frac{1-\phi_i^{2m}}{1-\phi_i^2}\right)}\right).$$

Setting this equal to zero, substituting in $\alpha_i^{(n+1)}(\phi_i)$ and simple manipulations gives (3.47). To show $\left(\sigma_i^{(n+1)}(\phi_i)\right)^2$ is a maximiser we use a similar argument to the independent regime case in the proof of Lemma 3.5.

Finally, substitute the maximisers in Equations (3.46) and (3.47) into Equation (3.32) with conditional densities given by Equation (3.19) and collect all terms involving ϕ_i, to give the function g. That we only need to search for the global maximiser of g on the interval $(-1, 1)$ comes from the fact that we have assumed Regime i is a stationary or mean-reverting process, in which case $|\phi_i| < 1$ is a necessary condition.

□

A pseudo-code implementation of our EM algorithm in given in Figure 3.7.

Remark 3.5. When using the EM algorithm, at the $(n+1)^{\text{th}}$ iteration, the forward algorithm can be initialised with probabilities $\mathbb{P}^{\boldsymbol{\theta}_n}(R_0 = i|\boldsymbol{x}_{0:T})$ calculated as part of the final step of the backward algorithm, rather than initialising the forward algorithm with probabilities π_j.

Remark 3.6. We terminate our implementation of the EM algorithm when the step size is small, i.e. when $|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n|_\infty < \varepsilon$ where ε is a specified tolerance (we choose $\varepsilon = 1.5 \times 10^{-6}$).

Remark 3.7. The memory and time complexity of our algorithms can be too large to be practical for models with two or more AR(1) regimes. A solution is to truncate the

Input: Data, $\boldsymbol{x_{0:T}}$, termination condition, ε, starting value, $\boldsymbol{\theta}_0$.
Output: MLEs, $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \log f^{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T})$.
Set $error = \varepsilon + 1$; $n = 0$;
Calculate π_j, the stationary distribution of $\{R_t\}$, for $j = \{1, ..., M\}$;
Set $\mathbb{P}_{\boldsymbol{\theta}_0}(\boldsymbol{H}_0 = (\boldsymbol{\Delta}, j)) = \pi_j$;
while $error > \varepsilon$ **do**
  `E-step`;
  Run the forward algorithm, initialising it with $\mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{H}_0 = (\boldsymbol{\Delta}, j))$, and store $\widehat{\alpha}^{(t)}_{\boldsymbol{N}_t}(i)$, for $t = 0, ..., T$, $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $i \in \mathcal{S}$;
  Run the backward algorithm and store $\gamma^{(t)}_{\boldsymbol{N}_t}(i)$, for $t = 0, ..., T$, $\boldsymbol{N}_t \in \mathcal{S}^{(t)}$ and $i \in \mathcal{S}$;
  Set $\mathbb{P}_{\boldsymbol{\theta}_{n+1}}(\boldsymbol{H}_0 = (\boldsymbol{\Delta}, j)) = \mathbb{P}^{\boldsymbol{\theta}_n}(\boldsymbol{H}_0 = (\boldsymbol{\Delta}, j)|\boldsymbol{x}_{0:T})$
  `M-step`;
  Set $\boldsymbol{\theta}_{n+1} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$;
  Set $error = \max_i |(\theta_i)_{n+1} - (\theta_i)_n|$
  Set $n = n + 1$;
end
return $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_n$

FIGURE 3.7: Pseudo-code implementing our EM algorithm

memory within Regime i at some lag ℓ, so that the observation x_t may only depend on the values $x_{t-1}, ..., x_{t-\ell}$, else x_t comes from the stationary distribution in Regime i. For Type II models we expect this approximation to have minimal affect on the accuracy of the inferences, since within regime processes decay to stationarity exponentially with the counter. Type III models do not have this property and more analysis is needed to determine the effects of this approximation. In practice, many of the smoothed probabilities are very close to zero and a smart implementation of this algorithm could easily take advantage of this.

3.5 Discussion

3.5.1 Convergence of EM and 'black-box' optimisation methods

MRS models of Type II To explore the practice of implementing maximum likelihood estimation we considered the following 2-regime model.

$$X_t = \begin{cases} B_t, & \text{when } R_t = 1, \\ S_t, & \text{when } R_t = 2. \end{cases} \tag{3.48}$$

where $B_t = \alpha + \phi B_{t-1} + \sigma_1 \varepsilon_t$ and $\{\varepsilon_t\}$ is i.i.d. N(0,1) noise and $\{S_t\}$ is an i.i.d. N(μ, σ_2^2) process, and assume that B_t evolves at all time points regardless of whether it is observed or not.

Realisations of the process in Equation (3.48) were simulated for a range of parameter sets and MATLAB's `fmincon` function (with the default settings) was used to find the maximisers of the likelihood, as given in Equation (3.12). We observed the points that `fmincon` converged to were highly dependent on the starting point of the algorithm. This was unsurprising as we know that `fmincon` converges to local and not necessarily global minima (see MATLAB's documentation and references therein). When initialised at random starting points, this method converged to the global maximum of the likelihood function about 40% of the time. The other 60% of the time this method either converged to a local maximum or to a point where only one regime of the model was capturing any of the data.

We also implemented our EM algorithm using the same simulated datasets and found that the EM algorithm showed similar behaviour to `fmincon`, although the EM algorithm showed some evidence that it was more stable since it converged to the true maximum more often. Example 3.4 illustrates the types of behaviour shown by `fmincon` and the EM algorithm.

Example 3.4. *Consider the MRS model given by (3.48) with parameters* $\alpha = 0$, $\phi = 0.7$, $\sigma_1 = 1$, $\mu = 5$, $\sigma_2 = \sqrt{2}$, $p_{11} = 0.9$ *and* $p_{22} = 0.5$. *We simulated a single realisation of length* $T = 2000$ *from this model and used* `fmincon` *and the EM algorithm to search for the MLE. We randomly sampled 20 parameter sets to initialise the optimisation algorithms and observed their terminating points. Table 3.1 details the starting and terminating points for the* `fmincon` *algorithm, while Table 3.2 details the terminating points for the EM algorithm.*

In Table 3.1 notice there are five different terminating behaviours. Terminating behaviour A occurs when `fmincon` finds the MLEs. Terminating behaviour B occurs when `fmincon` terminates at a local maximum corresponding to modelling the data generated by Regime 2 with Regime 1 and the data generated by Regime 1 with Regime 2.

Terminating behaviour C occurs when `fmincon` terminates at a point where $p_{11} = 0$, $p_{22} < 1$ and $\phi < 0$, which is a local maximum also. It seems counter intuitive that $\phi < 0$, could produce a local maximum for this simulated model since we specify an AR(1) process with positive correlation. However, since $p_{11} = 0$ the process cannot show two consecutive observations from Regime 1, so no two consecutive points are negatively correlated. Moreover, the corresponding lag 2 process of the AR(1) process has positive correlation ϕ^2. More generally, at even lags the AR(1) process has positive correlation while at odd lags the AR(1) process shows negative correlation. In this case the i.i.d. regime and the AR(1) regime capture data generated from both regimes.

Terminating behaviour D is when `fmincon` finds $p_{22} = 1$, and the algorithm has converged to a point where the data is being modelled by Regime 2 only. As a result, there is no contribution to the likelihood from Regime 1 and the algorithm is flat with respect to the parameters from Regime 1, thus `fmincon` terminates at a point where the likelihood is at a local maximum in the dimension of Regime-2 parameters, but is random in the Regime-1 parameters.

Terminating behaviour E is when `fmincon` finds $p_{11} = 1$, and the algorithm has converged to a point that uses only Regime 1 to model the data. This is similar to behaviour D, and we see the algorithm terminates at a point where the likelihood is at a local maximum in the dimension of Regime-1 parameters, but is random in the Regime-2 parameters.

Table 3.2 shows that the EM algorithm displays four terminating behaviours. We see a local maximum corresponding to the true MLE (behaviour A), the case when Regime 2 models data generated from Regime 1, and Regime 1 models data generated from Regime 2 (behaviour B), and the case when $p_{11} = 0$, $p_{22} < 1$ and $\phi < 0$ (behaviour C). The fourth terminating behaviour is convergence of the algorithm to 'Inf' values, which correspond to terminating behaviours D and E of the `fmincon` method. Since the EM algorithm does not rely on the gradient, it does not care that the function may be locally flat in some directions, rather the EM algorithm knows it is optimal to set one of the variance terms to zero and this produces Inf values. For example, in the case where the EM algorithm finds $p_{22} = 1$ (which corresponds to the terminating behaviour D), then the data are only being modelled by Regime 2, and Regime 1 models no data points. Therefore the EM algorithm updates the variance parameter to $\sigma_1 = 0$ and produces Inf values. This is technically the global maximum of the likelihood since the value of the likelihood at this point is infinite, however it is clearly not useful and does not give a sensible estimate of the parameters. While this behaviour is not too problematic for this model, as we can just ignore cases where this happens or restrict parameters to avoid these maxima, it does become problematic if we want to fit MRS models with shifted-log-normal regimes or shifted-Gamma regimes (see Section 3.5.3).

To attempt to stop `fmincon` from terminating at boundaries where $p_{jj} = 1$ or $p_{jj} = 0$, we constrained the probabilities p_{ij} to the interval $[0.05, 0.95]$ and observed that this increased the proportion of time that `fmincon` converged to the true MLE, but still observed convergence to the new boundaries. We implemented this in Example 3.5. The moral of the story is, we need to be careful when maximising the likelihood as there can be multiple local maxima in the likelihood function. Constraining parameters can help to avoid some sub-optimal terminating behaviours and can be implemented in practice if, for example, we have reason to believe that the parameters p_{jj} are not

TABLE 3.1: Terminating points and exit flags of the `fmincon` algorithm in MATLAB from 20 random starting points for the simulation in Example 3.4. Behaviour A – MLEs; B – a local maximum where the regimes are switched; C – a local maximum with $p_{11} = 0$, $p_{22} < 1$, $\phi < 0$; D – local maximum $p_{22} = 1$; E – local maximum $p_{11} = 1$. ℓ is the value of the log-likelihood at these terminating points. Exit flag 1 occurs when `fmincon` terminates because the first-order optimality measure is less than the specified tolerance (10^{-6}) and the constraints are not violated. This means that the gradient at the terminating point is close to zero and the algorithm thinks it is at a stationary point of the likelihood. Exit flag 2 occurs when the constraints are not violated and the norm of the step size is below the tolerance (10^{-10}).

		Terminating point of `fmincon`							Initial point							
Behaviour	ℓ	α	ϕ	σ_1	μ	σ_2	p_{11}	p_{22}	α	ϕ	σ_1	μ	σ_2	p_{11}	p_{22}	Exit flag
A	-3710.1	-0.01	0.73	0.98	4.84	1.59	0.90	0.54	3.60	0.17	1.54	7.43	8.66	0.08	0.38	2
									2.03	0.89	6.85	3.15	0.04	0.66	0.26	2
									8.25	-0.48	8.89	9.90	8.30	0.02	0.89	2
									3.25	0.32	6.35	12.99	1.47	0.18	0.87	2
									0.44	-0.90	4.94	17.63	0.48	0.07	0.30	2
B	-4119.6	0.59	0.72	1.93	-0.29	1.04	0.79	0.76	0.30	0.86	8.24	3.06	3.08	0.34	0.31	2
									0.96	0.79	8.64	18.98	1.09	0.33	0.66	2
									2.05	-0.77	2.69	18.70	6.94	0.74	0.96	2
									1.90	-0.37	0.99	18.65	1.93	0.90	0.58	2
									5.30	0.90	0.84	10.69	1.35	0.76	0.58	2
C	-4457.2	0.18	-0.85	0.77	1.30	2.65	0.00	0.45	6.31	0.24	2.37	14.63	1.56	0.24	0.55	2
									5.17	0.14	0.89	3.33	2.44	0.61	0.87	2
									7.48	-0.68	8.74	6.40	4.24	0.44	0.97	2
D	-4575.2	7763.50	-0.86	5638.10	0.84	2.38	0.00	1.00	8.94	-0.50	7.74	6.36	9.47	0.00	0.22	1
		-143.79	-0.86	220.38	0.84	2.38	0.00	1.00	6.46	0.97	3.51	8.49	2.00	0.87	0.51	1
		-60.81	0.03	415.42	0.84	2.38	0.00	1.00	2.84	0.67	5.56	0.17	0.05	0.62	0.44	1
E	-4306.3	0.43	0.49	2.08	18.98	1.22	1.00	0.55	1.21	0.10	7.73	10.86	6.51	0.81	0.03	1
		0.43	0.49	2.08	18.80	1.74	1.00	0.48	5.11	-0.17	1.04	11.22	8.19	0.59	0.99	1
		0.43	0.49	2.08	17.63	0.92	1.00	0.48	7.28	-0.79	9.04	18.80	1.54	0.49	0.39	1
		0.43	0.49	2.08	-328.20	995.11	1.00	0.70	5.23	0.28	9.84	0.24	9.80	0.77	0.82	1

TABLE 3.2: Terminating points of our EM algorithm from 20 random starting points in Example 3.4. Notice that there are four distinct terminating behaviours. Behaviour A – the MLE; B – a local maximum where the regimes are switched; C – a local maximum with $p_{11} = 0$, $p_{22} < 1$, $\phi < 0$; D/E – a maximum with either $p_{11} = 1$ and $\sigma_2 = 0$, or $p_{22} = 1$ and $\sigma_1 = 0$. The terminating point Inf are analogous to terminating behaviours D and E discussed for the `fmincon` method. ℓ is the value of the log-likelihood.

		Terminating point							Initial point						
Behaviour	ℓ	α	ϕ	σ_1	μ	σ_2	p_{11}	p_{22}	α	ϕ	σ_1	μ	σ_2	p_{11}	p_{22}
A	-3710.1	-0.01	0.73	0.98	4.84	1.59	0.90	0.54	6.46	0.97	3.51	8.49	2.00	0.87	0.51
									8.25	-0.48	8.89	9.90	8.30	0.02	0.89
									3.25	0.32	6.35	12.99	1.47	0.18	0.87
									6.31	0.24	2.37	14.63	1.56	0.24	0.55
									7.48	-0.68	8.74	6.40	4.24	0.44	0.97
									1.21	0.10	7.73	10.86	6.51	0.81	0.03
									5.23	0.28	9.84	0.24	9.80	0.77	0.82
									0.30	0.86	8.24	3.06	3.08	0.34	0.31
									2.05	-0.77	2.69	18.70	6.94	0.74	0.96
									1.90	-0.37	0.99	18.65	1.93	0.90	0.58
									5.30	0.90	0.84	10.69	1.35	0.76	0.58
B	-4119.6	0.59	0.72	1.93	-0.29	1.04	0.79	0.76	2.84	0.67	5.56	0.17	0.05	0.62	0.44
									3.60	0.17	1.54	7.43	8.66	0.08	0.38
									5.17	0.14	0.89	3.33	2.44	0.61	0.87
									5.11	-0.17	1.04	11.22	8.19	0.59	0.99
C	-4457.2	0.18	-0.85	0.77	1.30	2.65	0.00	0.45	8.94	-0.50	7.74	6.36	9.47	0.00	0.22
D/E	-	Inf	Inf	Inf	Inf	Inf	Inf	Inf	2.03	0.89	6.85	3.15	0.04	0.66	0.26
									0.44	-0.90	4.94	17.63	0.48	0.07	0.30
									7.28	-0.79	9.04	18.80	1.54	0.49	0.39
									0.96	0.79	8.64	18.98	1.09	0.33	0.66

close to 1. Furthermore, we recommend initialising these optimisation algorithms from a range of starting points so that they do not get stuck at a local maxima.

Example 3.5. *Again consider the model in Equation (3.48) with the same parameters as in Example 3.4. We simulated a single realisation of this process of length* $T = 2000$ *and use* `fmincon` *to find the MLE, initialising the algorithm from 20 random starting points. This time we also restricted* `fmincon` *away from the boundaries by specifying* $p_{11}, p_{22} \in [0.05, 0.95]$. *The initial and terminating points for this example are in Table 3.3. Notice the higher proportion of terminating points corresponding to the true MLEs.*

MRS models of Type III We also studied terminating points of optimisation algorithms for MRS models with independent regimes that evolve only when observed. Consider the following two regime model of Type III

$$X_t = \begin{cases} B_{\tau(t)}, & \text{when } R_t = 1, \\ S_t, & \text{when } R_t = 2. \end{cases} \tag{3.49}$$

where $\tau(t) = \sum_{t=0}^{t} \mathbb{I}(R_t = 1)$, $B_{\tau(t)} = \alpha + \phi B_{\tau(t-1)} + \sigma_1 \varepsilon_{\tau(t)}$ and $\varepsilon_{\tau(t)}$ is i.i.d. N(0,1) noise, S_t is i.i.d. $\mathrm{N}(\mu, \sigma_2^2)$ and $\{R_t\}$ is a Markov chain on the state space $\{1, 2\}$. That is, assume that $B_{\tau(t)}$ evolves only when it is observed. Again, using simulations, the terminating points of `fmincon` and the EM algorithm were investigated. We observed similar behaviours to those in Example 3.4. Example 3.6 demonstrates this.

Example 3.6. *Consider the model in Equation (3.49) with parameters* $\alpha = 0$, $\phi = 0.7$, $\sigma_1 = 1$, $\mu = 5$, $\sigma_2 = \sqrt{2}$, $p_{11} = 0.9$ *and* $p_{22} = 0.5$. *A single dataset was simulated from this model, and* `fmincon` *and the EM algorithm were used to find the MLEs from 20 random starting points. We report the starting points and terminating points for the* `fmincon` *method in Table 3.4, and those for the EM algorithm in Table 3.5.*

In Table 3.4 we see that `fmincon` displays four types of convergence behaviour corresponding to behaviours A, B, D and E from before. Similarly to our simulation in Example 3.4, for convergence behaviour D (respectively, behaviour E), `fmincon` finds local maximisers for Regime 2 (respectively, Regime 1), but not for Regime 1 (respectively, Regime 2). This is because the likelihood is flat in the dimensions of Regime 2's parameters (respectively, Regime 1's parameters). Notice in Table 3.4 that we do not see convergence behaviour C for Type III models. For Type III models, if $\phi < 0$, consecutive observations from the process $\{B_{\tau(t)}\}$ must be negatively correlated, regardless of the distance between observations from the process $\{B_{\tau(t)}\}$. Thus, Type III models do not have the flexibility of Type II models, where correlations between successive observed

TABLE 3.3: Terminating points of `fmincon` for Example 3.5 where we have restricted the parameters $p_{11}, p_{22} \in [0.05, 0.95]$. Behaviour A – MLEs; B – a local maximum where the regimes are switched; C – a local maximum with $p_{11} = 0$, $p_{22} < 1$, $\phi < 0$; D – local maximum $p_{22} = 0.95$. ℓ is the value of the log-likelihood. Exit flag 1 occurs when `fmincon` terminates because the first-order optimality measure is less than the specified tolerance (10^{-6}) and the constraints are not violated. This means that the gradient at the terminating point is close to zero and the algorithm thinks it is at a stationary point of the likelihood. Exit flag 2 occurs when the constraints are not violated and the norm of the step size is below the tolerance (10^{-10}). Notice that restricting the parameters has increased the proportion of times that the algorithm finds the true MLE compared to Table 3.1, but has not eliminated unwanted terminating points.

		Terminating point							Initial point							
Behaviour	ℓ	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}	Exit flag
A	-3710.1	-0.01	0.73	0.98	4.84	1.59	0.90	0.54	2.71	-0.36	5.36	7.00	9.61	0.40	0.12	2
									2.84	0.67	5.56	0.17	0.05	0.62	0.44	2
									5.72	-0.76	4.51	9.44	9.75	0.80	0.98	2
									2.05	-0.77	2.69	18.70	6.94	0.74	0.96	2
									0.30	-0.47	8.19	12.60	1.05	0.90	0.20	2
									3.25	0.32	6.35	12.99	1.47	0.18	0.87	2
									1.21	0.10	7.73	10.86	6.51	0.81	0.03	2
									8.85	1.00	0.32	19.11	4.62	0.64	0.34	2
									9.76	0.19	2.72	16.18	0.35	0.39	0.24	2
B	-4119.6	0.59	0.72	1.93	-0.29	1.04	0.79	0.76	5.41	0.66	1.38	1.34	3.80	0.86	0.94	2
									5.23	0.28	9.84	0.24	9.80	0.77	0.82	2
									1.62	-0.77	5.61	2.63	4.41	0.11	0.90	2
									0.30	0.86	8.24	3.06	3.08	0.34	0.31	2
									9.10	0.13	0.70	8.12	6.33	0.70	0.07	2
									0.02	0.24	5.11	2.46	4.54	0.17	0.90	2
									7.84	0.42	0.18	12.05	7.88	0.32	0.79	2
									6.31	0.24	2.37	14.63	1.56	0.24	0.55	2
C	-4461.7	0.11	-0.85	0.75	1.32	2.66	0.05	0.47	8.94	-0.50	7.74	6.36	9.47	0.00	0.22	2
D	-4614.7	-44.29	-1.00	0.16	0.81	2.35	0.05	0.95	9.66	-0.22	2.70	16.98	8.05	0.67	0.35	1
	-4677.8	15.58	0.93	0.05	0.84	2.38	0.05	0.95	8.94	-0.13	9.50	6.45	6.32	0.63	0.92	1

values from an AR(1) process can change sign depending on whether the distance between the two observations is odd or even. The data that we simulated from the model in Example 3.4 has a large proportion of consecutive, positively correlated observations, thus we hypothesise that it is unlikely for a Type III model with $\phi < 0$ to fit the data well, and that the domain of attraction for behaviour C is much smaller in this case.

In Table 3.5 we see three types of convergence behaviour. The EM algorithm either finds the true MLE (behaviour A), or converges to a point where Regime 1 is used to model data generated from Regime 2 and Regime 2 is used to model data generated from Regime 1 (behaviour B), or the algorithm produces 'Inf' values (behaviour D/E). In the latter case, the EM algorithm has actually converged to a point where $p_{11} = 1$ and $\sigma_2 = 0$ or $p_{22} = 1$ and $\sigma_1 = 0$, which causes the algorithm to produce NaN values. This behaviour is analogous to behaviours D and E of the `fmincon` method in Table 3.4.

In practice, choosing which optimisation method to use is problem-specific. We have seen that the EM algorithm appears to converge to the true MLE more often; however, the memory requirements of the EM algorithm are much larger than the forward algorithm. For the code used in the previous examples, implementing the EM algorithm for the simple two-regime models discussed in this section for datasets of size $T = 20,000$ was infeasible on an Apple iMac with 8 GB 1867 MHz DDR3 RAM, and we had to resort to the University's high performance computing facilities. Compare this to implementing the forward algorithm and `fmincon`, which was easily executed on the Apple iMac.

In terms of computation time, for MRS model of Type III, the EM and `fmincon` methods performed similarly, while for models of Type II the `fmincon` method was quicker. However, this is probably not a fair comparison, since MATLAB has optimised `fmincon`'s code, while the code we use for the EM algorithm has not had the same treatment.

Lastly, when applying these maximisation techniques to real electricity price datasets we can impose constraints on parameters to help these algorithms converge to the true MLE as we did in Example 3.5. We can also consider other logical constraints such as a constraint to ensure that low prices cannot be modelled by a spike regime, or a constraint to ensure that the majority of points in the dataset belong to the base regime, or constrain $\phi > 0$ since we expect positive correlation between prices.

3.5.2 Bias and consistency of the MLE

To study the bias and consistency of our algorithm, we conducted a simulation study which showed bias was small for sample sizes greater than 200 observations. As our datasets are an order of magnitude larger than this, then we can expect negligible bias

TABLE 3.4: Terminating points of `fmincon` for Example 3.6. Behaviour A – MLEs; B – a local maximum where the regimes are switched; D – local maximum $p_{22} = 1$; E – local maximum $p_{11} = 1$. ℓ is the value of the log-likelihood. Exit flag 1 occurs when `fmincon` terminates because the first-order optimality measure is less than the specified tolerance (10^{-6}) and the constraints are not violated. This means that the gradient at the terminating point is close to zero and the algorithm thinks it is at a stationary point of the likelihood. Exit flag 2 occurs when the constraints are not violated and the norm of the step size is below the tolerance (10^{-10}). Exit flag 0 occurs when `fmincon` has exceeded the maximum number of iterations allowed, in this case 3000.

		Terminating point							Initial point							
Behaviour	ℓ	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}	Exit flag
A	-3661.4	0.06	0.70	1.01	5.18	1.27	0.90	0.50	8.85	1.00	0.32	19.11	4.62	0.64	0.34	2
									7.80	-0.03	8.69	10.75	5.90	0.40	0.93	2
									6.19	0.17	7.87	10.95	7.07	0.48	0.56	2
									8.11	0.47	4.85	16.99	5.02	0.50	0.04	2
									1.07	0.96	4.04	7.95	1.09	0.94	0.96	2
									0.30	-0.47	8.19	12.60	1.05	0.90	0.20	2
									8.76	0.13	5.81	17.99	2.42	0.00	0.35	2
									4.51	-0.51	9.65	15.06	2.23	0.93	0.59	2
B	-4096.6	1.00	0.61	1.98	-0.24	1.09	0.79	0.82	7.46	0.73	4.79	9.03	6.18	0.13	0.63	2
									9.19	0.69	7.28	7.33	9.10	0.18	0.96	2
									1.07	-0.17	1.94	2.17	0.86	0.29	0.42	2
									9.10	0.13	0.70	8.12	6.33	0.70	0.07	2
									0.02	0.24	5.11	2.46	4.54	0.17	0.90	2
									8.36	0.17	7.74	1.77	9.01	0.07	0.78	2
D	-4517.2	-0.48	-0.99	0.00	1.04	2.32	0.00	1.00	6.91	0.30	5.87	2.47	8.17	0.20	0.79	2
	-4522.2	2655900	0.35	1080700	1.04	2.32	0.00	1.00	4.00	-0.96	4.31	2.23	3.59	0.41	0.19	1
	-4522.2	-6.97	0.67	14.34	1.04	2.32	0.00	1.00	6.31	0.24	2.37	14.63	1.56	0.24	0.55	1
	-4522.2	-20.21	-0.14	27.67	1.04	2.32	0.00	1.00	5.59	0.31	3.21	1.67	3.91	0.86	0.33	1
	-4522.2	-10.50	0.00	64.80	1.04	2.32	0.00	1.00	8.44	0.08	7.04	10.32	2.40	0.60	0.52	1
E	-4291.5	0.55	0.47	2.04	4564.10	7895.70	1.00	0.03	5.06	-0.51	1.98	2.88	0.87	0.38	0.61	0

TABLE 3.5: Terminating points of EM for Example 3.6. Notice the three convergence behaviours. Behaviour A – MLEs; B – a local maximum where the regimes are switched; D/E – maximum with $p_{22} = 1$/maximum with $p_{11} = 1$.

		Terminating points							Initial points						
Behaviour	ℓ	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}	α_1	ϕ_1	σ_1	μ_2	σ_2	p_{11}	p_{22}
A	-3661.4	0.06	0.70	1.01	5.18	1.27	0.90	0.50	7.80	-0.03	8.69	10.75	5.90	0.40	0.93
									6.19	0.17	7.87	10.95	7.07	0.48	0.56
									8.11	0.47	4.85	16.99	5.02	0.50	0.04
									1.07	0.96	4.04	7.95	1.09	0.94	0.96
									8.76	0.13	5.81	17.99	2.42	0.00	0.35
									4.51	-0.51	9.65	15.06	2.23	0.93	0.59
									7.46	0.73	4.79	9.03	6.18	0.13	0.63
									9.19	0.69	7.28	7.33	9.10	0.18	0.96
									0.02	0.24	5.11	2.46	4.54	0.17	0.90
									4.00	-0.96	4.31	2.23	3.59	0.41	0.19
									8.44	0.08	7.04	10.32	2.40	0.60	0.52
B	-4096.6	1.00	0.61	1.98	-0.24	1.09	0.79	0.82	1.07	-0.17	1.94	2.17	0.86	0.29	0.42
									9.10	0.13	0.70	8.12	6.33	0.70	0.07
									8.36	0.17	7.74	1.77	9.01	0.07	0.78
									6.91	0.30	5.87	2.47	8.17	0.20	0.79
									5.59	0.31	3.21	1.67	3.91	0.86	0.33
									5.06	-0.51	1.98	2.88	0.87	0.38	0.61
D/E	-	Inf	Inf	Inf	Inf	Inf	Inf	Inf	8.85	1.00	0.32	19.11	4.62	0.64	0.34
									0.30	-0.47	8.19	12.60	1.05	0.90	0.20
									6.31	0.24	2.37	14.63	1.56	0.24	0.55

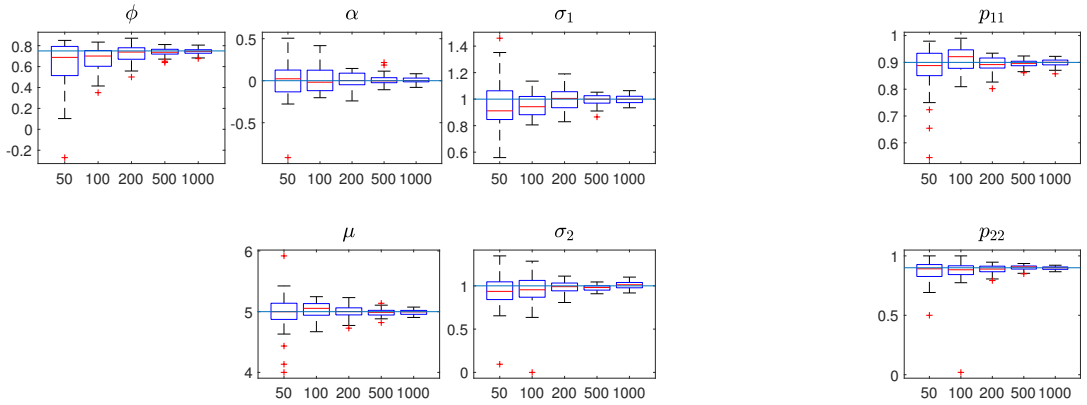


FIGURE 3.8: Boxplots of MLEs for simulated MRS model from Example 3.7. Each boxplot contains 40 MLEs from simulated datasets of length given on the x-axis. The solid horizontal blue lines represent the true parameter value, $\boldsymbol{\theta} = (\alpha, \phi, \sigma_1, \mu, \sigma_2, p_{11}, p_{22}) = (0, 0.75, 1, 5, 1, 0.9, 0.9)$. Notice the small and decreasing bias in the parameters ϕ, σ_1 and σ_2.

in our MLEs. Moreover, our simulation study suggests that the MLEs are asymptotically consistent. These conclusions hold for both Type II and Type III MRS models. Example 3.7 illustrates these conclusions.

Example 3.7. *Consider the MRS model of Type II in Equation (3.48) with parameters* $\boldsymbol{\theta} = (\alpha, \phi, \sigma_1, \mu, \sigma_2, p_{11}, p_{22}) = (0, 0.75, 1, 5, 1, 0.9, 0.9)$. *We simulated 40 realisations of this process for each length* $T = 50, 100, 200, 500$ *and* 1000 *(a total of 200 independent simulations). Our methods were then used to find the MLEs for each simulation, with the algorithms initialised at the true value of the parameters. Figure 3.8 shows box-plots that summarise the MLEs for these simulated datasets. Notice the bias is small for these datasets, and that the MLEs seem to converge to the true parameter values as the sample size increases.*

Related models have been proven to produce consistent MLEs. Robinson [92] proves MLEs are consistent estimators for parameters of AR(1) processes observed at discrete, not necessarily equally spaced times; Leroux [69] proves consistency of the MLEs for hidden Markov models with general (not necessarily discrete) observation distribution; and Francq and Roussignol [36] prove consistency of the MLEs for MRS models of Type I (with dependent regimes). There are also other papers in the literature proving consistency for the cases above, which modify and relax some of the assumptions in these papers. It would be useful to prove the consistency of our algorithms, but this is not the focus of this thesis.

3.5.3 The difficulties of shifted-log-normal and shifted-Gamma distributions

As we eluded to earlier, there are difficulties in estimating the shifting parameter q_i using maximum likelihood for the shifted-log-normal and shifted-Gamma distributions. This is due to the likelihood approaching ∞ as the shifting parameter approaches the MLE.

Shifted-log-normal distributions The shifted-log-normal distribution has density function given by

$$f(x;\mu,\sigma,q) = \frac{1}{\sqrt{2\pi}\sigma}(x-q)^{-1}e^{-\frac{1}{2\sigma^2}(\log(x-q)-\mu)^2}.$$

Suppose $\boldsymbol{x} := (x_1, ..., x_n)$, with $x_1 < x_2 < ... < x_n$, are observations from a shifted-log-normal distribution, then the likelihood is

$$L(\mu,\sigma,q;\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n}(x_i-q)^{-1}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(x_i-q)-\mu)^2},$$

for $q < x_1$, and the log-likelihood is

$$\ell(\mu,\sigma,q;\boldsymbol{x}) = \left(-n\log\sqrt{2\pi} - n\log\sigma - \sum_{i=1}^{n}(x_i-q) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(x_i-q)-\mu)^2\right),$$

for $q < x_1$ also.

The MLEs for μ and σ are

$$\widehat{\mu}(q) = \frac{1}{n}\sum_{i=1}^{n}\log(x_i-q),$$
$$\widehat{\sigma}(q)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log(x_i-q)-\mu\right]^2.$$

Substituting $\widehat{\mu}(q)$ and $\widehat{\sigma}(q)$ into the likelihood equation gives

$$L^*(q;\boldsymbol{x}) \propto \left(\frac{1}{\widehat{\sigma}(q)}\right)^n \prod_{i=1}^{n}(x_i-q)^{-1}.$$

Hill [50] notes that the following limits

$$\lim_{\mu\to-\infty} L(\mu,\sigma,q;\boldsymbol{x}) = \lim_{\mu\to+\infty} L(\mu,\sigma,q;\boldsymbol{x}) = 0,$$
$$\lim_{\sigma\to-\infty} L(\mu,\sigma,q;\boldsymbol{x}) = \lim_{\sigma\to+\infty} L(\mu,\sigma,q;\boldsymbol{x}) = 0,$$
$$\lim_{q\to-\infty} L(\mu,\sigma,q;\boldsymbol{x}) = \lim_{q\to x_1} L(\mu,\sigma,q;\boldsymbol{x}) = 0,$$

are all nicely behaved, but the function L^* is not:

$$\lim_{q\to-\infty} L^*(q;\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^{-n/2}, \tag{3.50}$$

$$\lim_{q\to x_1} L^*(q;\boldsymbol{x}) = \infty. \tag{3.51}$$

To see Equation (3.51), Hill shows $\widehat{\sigma}(q)^2 < \log^2(x_1 - q)$ for $q \in (x_1-\varepsilon, x_1)$ for sufficiently small $\varepsilon > 0$, hence

$$L^*(q;\boldsymbol{x}) = \left(\frac{1}{\widehat{\sigma}(q)}\right)^n \prod_{i=1}^{n}(x_i-q)^{-1} \geq |\log(x_1-q)|^{-n}(x_1-q)^{-1}\prod_{i=2}^{n}(x_i-q)^{-1}$$

for $q \in (x_1-\varepsilon, x_1)$. Noting that $\prod_{i=2}^{n}(x_i-q) \to \prod_{i=2}^{n}(x_i-x_1)$ as $q \to x_1$, then the right hand side goes to ∞ as $q \to x_1$, and the MLE is therefore $(q,\mu,\sigma) = (x_1, -\infty, \infty)$.

Hill then comments that both of these results are surprising. We expect the likelihood to be very small in remote regions of the likelihood, but Equation (3.50) shows this is not the case, and Equation (3.51) says arbitrarily large likelihood values can be achieved by allowing q to converge to x_1 along the path $(\widehat{\mu}(q), \widehat{\sigma}(q), q)$. The function L^* is a very interesting function indeed.

A common workaround is to instead use a *local* maximum likelihood estimate for q. However, when we extend this to the MRS model there are yet more problems. Firstly, when we try to estimate the parameters of a shifted-log-normal distribution embedded in an MRS model, the parameter q is no longer restricted to being less than the smallest observation, and the likelihood is infinite at any point where $q = x_i$. Furthermore, local maxima of the likelihood exist which correspond to the shifted-log-normal collapsing to a point mass on a single observation (where the variance of the shifted-log-normal is small and the mode of the distribution is located at some observation x_i) so there are many more points where the likelihood is infinite.

Even if a local maximum likelihood method estimate is used instead, there is still a problem: there exist many local maxima of the likelihood, all with similar likelihood

values, but the estimates of q obtained from these local maxima can vary wildly. An example of this is shown in Table 3.6, where we simulated a single realisation of length 400 from an MRS model of Type II, with one AR(1) regime and a shifted-log-normal regime, and used our EM algorithm to find the maxima of the likelihood.

The shifted-Gamma distribution The shifted-Gamma distribution has density function

$$f(x;\mu,\sigma^2,q) = \begin{cases} \dfrac{1}{\sigma^{2\mu}\Gamma(\mu)}(x-q)^{\mu-1}\exp\left(-\dfrac{x-q}{\sigma^2}\right) & x > q, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $\boldsymbol{x} := (x_1, ..., x_n)$, with $x_1 < x_2 < ... < x_n$, are *ordered* observations from a shifted-Gamma distribution, then the likelihood is

$$L(\mu,\sigma^2,q;\boldsymbol{x}) = \left(\frac{1}{\sigma^{2\mu}\Gamma(\mu)}\right)^n \left[\prod_{i=1}^{n}(x_i-q)^{\mu-1}\right] \exp\left(-\sum_{i=1}^{n}\frac{x_i-q}{\sigma^2}\right),$$

for $q < x_1$, and the log-likelihood is

$$\ell(\mu,\sigma^2,q;\boldsymbol{x}) = -2n\mu\log\sigma - n\log\Gamma(\mu) + (\mu-1)\sum_{i=1}^{n}\log(x_i-q) - \sum_{i=1}^{n}\frac{x_i-q}{\sigma^2},$$

for $q < x_1$ also.

One issue with the shifted-Gamma distribution is, when $\mu < 1$, the MLE is to set $q = x_1$, the likelihood is infinite, and the MLE for the other two parameters does not exist. Johnson and Kotz [64] observe that related issues can arise when μ is near 1, and advise against maximum likelihood estimation when $\mu < 2.5$. Simulations suggest this is also good advice when fitting MRS models with shifted-Gamma regimes.

After restricting $\mu > 2.5$, one more issue stills exits. As was the case for shifted-log-normal regimes, there are often many local maxima. However, unlike the shifted-log-normal distribution, these local maxima are better behaved; there is typically a single maxima, with the highest likelihood, and the parameters are 'close' to their true values. Although, when $\mu < 2.5$, simulations suggest this behaviour is not so nice, and there can be local maxima with parameter values that do not relate the true parameters.

The restriction $\mu > 1$ limits the shape of the Gamma distribution, by requiring the density function to be zero at $x = q$. Moreover, the mode of the Gamma distribution is $(\mu-1)\sigma^2$, so the restriction $\mu > 2.5$ ensures that the mode of the Gamma distribution is away from the boundary at q (provided that $\sigma > 0$, which it should be, otherwise the

ℓ	α	ϕ	σ_1	q	μ	σ_2	p_{11}	p_{22}	Mode
-1842.75	5.36	0.35	583.61	228.61	-7.50	113.88	0.99	0.00	228.61
-1842.10	5.36	0.35	583.61	228.61	-7.86	121.62	0.99	0.00	228.61
-1842.01	5.36	0.35	583.61	228.61	-7.91	122.69	0.99	0.00	228.61
-1805.69	4.85	0.34	450.17	179.39	-2.13	99.28	0.99	0.00	179.51
-1778.38	4.39	0.37	381.16	173.71	-1.24	82.38	0.99	0.00	174.00
-1778.31	4.39	0.37	381.16	173.71	-1.26	83.02	0.99	0.00	173.99
-1777.74	4.39	0.37	381.16	173.71	-1.40	88.40	0.99	0.00	173.95
-1742.89	3.73	0.38	275.87	115.19	3.89	0.64	0.98	0.00	163.98
-1742.89	3.73	0.38	275.87	115.19	3.89	0.64	0.98	0.00	163.98
-1742.89	3.73	0.38	275.87	115.19	3.89	0.64	0.98	0.00	163.98
-1715.45	2.29	0.45	181.80	4.82	4.38	0.46	0.94	0.00	84.75
-1715.44	2.29	0.45	181.76	5.24	4.37	0.46	0.94	0.00	84.65
-1715.43	2.29	0.45	181.71	5.74	4.37	0.47	0.94	0.00	84.54
-1667.02	2.48	0.32	141.33	52.85	3.19	2.61	0.96	0.27	77.03
-1667.02	2.48	0.32	141.33	52.85	3.19	2.61	0.96	0.27	77.03
-1646.80	2.13	0.34	122.33	50.58	2.97	2.99	0.95	0.32	70.00
-1623.08	1.66	0.33	96.97	39.97	3.14	2.04	0.95	0.38	63.04
-1607.71	1.22	0.36	78.36	28.03	3.52	0.94	0.93	0.34	61.72
-1607.71	1.22	0.36	78.36	28.03	3.52	0.94	0.93	0.34	61.72
-1595.44	0.69	0.40	61.47	21.42	3.27	1.45	0.91	0.37	47.80
-1583.56	0.26	0.44	51.33	11.96	3.33	1.17	0.91	0.47	39.90
-1582.11	0.09	0.46	48.49	5.29	3.41	1.01	0.90	0.51	35.50
-1581.18	0.11	0.51	47.22	-17.76	3.73	0.60	0.89	0.66	23.98
-1580.85	0.02	0.51	47.29	-11.52	3.54	0.78	0.89	0.66	22.94
-1580.85	0.02	0.51	47.29	-11.52	3.54	0.78	0.89	0.66	22.94
-1580.85	0.02	0.51	47.28	-11.40	3.53	0.79	0.89	0.66	22.84
NaN	6.33	0.28	723.25	245.37	1.94	0.00	1.00	0.00	252.36
NaN	6.33	0.28	723.25	234.91	2.86	0.00	1.00	0.00	252.36
NaN	6.33	0.28	723.25	251.34	0.01	0.00	1.00	0.00	252.36
NaN	6.33	0.28	723.25	234.62	2.88	0.00	1.00	0.00	252.36
True value	0.00	0.55	53.00	12.00	3.50	1.00	0.90	0.50	45.12

TABLE 3.6: Local maximisers found by `EM` for a simulated realisation of an MRS model of Type II with a shifted-log-normal regime. Each row in the table corresponds to a terminating point of our EM algorithm, and each run of the EM algorithm was initialised from a randomly sampled starting point. The column ℓ corresponds to the value of the log-likelihood at the terminating point. The rightmost column titled 'mode' is the mode of the shifted-log-normal distribution, which often is located exactly on an observation, which is the case for the points at 252.36, 228.61 and 84.54. When this is the case, either μ is negative and σ large and positive, or σ is close to zero. Both of these behaviours occur when the EM algorithm is in the domain of attraction of one of the points when the likelihood tends to infinity. Another case when the likelihood is tending to infinity occurs when q is at one of the observation, which is the case, for example, when $q = 252.36$, 228.62, 179.39, and 173.71. All the other terminating points are local maxima, and notice that the value of q varies greatly between them.

Gamma distribution is a point mass). When modelling electricity price data, recall that the shifted-Gamma distribution is typically used to capture large price spikes, while an AR(1) process is used to capture 'base' prices. It is logical that the mode of the spike distribution is away from the shifting parameter q, as this ensures the majority of the mass of the spike regime is away from the base regime. The spike regime should capture those extreme prices, as well as some prices relatively near q that are not suitably modelled by the AR(1) process. In summary, restricting $\mu > 2.5$ is not as limiting as it may seem, and is likely to make the model easier to interpret.

3.5.4 Applications/Extensions for more complex models

A natural extension to the time-homogeneous models considered here is to introduce exogenous variables into the switching probabilities. This can be achieved by modelling the switching probabilities using multinomial logistic regression. Let $\boldsymbol{z}_t$, $t = 1, ..., T$, be row vectors of predictor variables with length r. Then, for example, the switching probability can be modelled as

$$\mathbb{P}(R_t = i | \boldsymbol{z}_t, R_{t-1} = j) = \frac{e^{\boldsymbol{\beta}_{j,i}\boldsymbol{z}_t'}}{1 + \sum_{k=1}^{M-1} e^{\boldsymbol{\beta}_{j,k}\boldsymbol{z}_t'}}, \tag{3.52}$$

where the superscript $'$ is the transpose and $\boldsymbol{\beta}_{j,k}$ $j, k \in \mathcal{S}$ are row vectors of regression coefficients with length r. If the sequence $\{R_t\}_{t\in\{1,...,T\}}$ is known, the log-likelihood of the switching process can be written as

$$\sum_{t=1}^{T} \sum_{i\in\mathcal{S}} \mathbb{I}(R_t = i, R_{t-1} = j) \log \frac{e^{\boldsymbol{\beta}_{j,i}\boldsymbol{z}_t'}}{1 + \sum_{k=1}^{M-1} e^{\boldsymbol{\beta}_{j,k}\boldsymbol{z}_t'}}.$$

As the sequence $\{R_t\}$ is not known for MRS models, then applying the EM methodology to this log-likelihood expression (taking the expectation of this expression given the observed prices, the predictor variables $\boldsymbol{z}_t$, and the current parameters $\boldsymbol{\theta}_n$) yields

$$\sum_{t=1}^{T} \sum_{i,j\in\mathcal{S}} \mathbb{P}^{\boldsymbol{\theta}_n}(R_t = i, R_{t-1} = j | \boldsymbol{x}_{0:T}, \boldsymbol{z}_{0:T}) \log \frac{e^{\boldsymbol{\beta}_{j,i}\boldsymbol{z}_t'}}{1 + \sum_{k=1}^{M-1} e^{\boldsymbol{\beta}_{j,k}\boldsymbol{z}_t'}}. \tag{3.53}$$

Replacing the terms $p_{ij}^{(n)}$ in the forward and backward algorithms with (3.52) evaluated with regression parameters from the previous iteration of the EM algorithm, and

replacing the sum

$$\sum_{i,j\in\mathcal{S}}\sum_{t=1}^{T}\mathbb{P}^{\boldsymbol{\theta}_n}(R_{t-1}=i,R_t=j|\boldsymbol{x}_{0:T})\log p_{ij}$$

in Equation (3.32) with the expression (3.53) gives the appropriate function Q to implement the EM algorithm for an MRS model with independent regimes and dependent switching parameter. So, our forward algorithm or EM algorithm can be used to find the MLEs for this model also.

Another possible extension would be to include exogenous variables in the mean for each regime. For example, suppose Regime j is an AR(1) regime, then the parameter α_j could be replaced by the linear function

$$\alpha_j(\boldsymbol{\beta}_j,\boldsymbol{z}_t)=\boldsymbol{\beta}_j\boldsymbol{z}_t',$$

where $\boldsymbol{\beta}_j$ is a row vector of regression coefficients, and $\boldsymbol{z}_t$ is exogenous data. However, in this case the M-step of the EM algorithm may have to be performed numerically. These types of model are explored in [80], although they rely on the EM-like algorithm for approximate parameter estimation.

Our forward and backward algorithms can also be extended to cope with higher-order autoregressive processes by adding more counters to the hidden Markov chain, for example, for AR(2) processes, by augmenting the hidden Markov chain with last visit counters and second-to-last visit counters. However, memory requirements and complexity would greatly increase.

So far we have assumed there is at least one AR(1) regime and one i.i.d. regime, however this is not necessary and our algorithms can be easily be modified for these models.

Chapter 4

Bayesian inference methods for independent-regime MRS models

In this chapter, we describe a data-augmented Bayesian method for estimating the parameters of MRS models. We implement a sophisticated Markov Chain Monte Carlo algorithm with an automatic parameter tuning aspect. The motivation for this was to have a method that would adapt to various models, without the need for manually editing each time.

4.1 The Bayesian framework

Recall from Section 2.2.4 that parameters, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, are treated as random variables and the goal of Bayesian inference is to infer the posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{x})$, where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, and $\boldsymbol{x}$ is a vector of observed data. Depending on the MRS model being investigated, the vector $\boldsymbol{\theta}$ contains the parameters $\alpha_\ell, \phi_\ell, \sigma_\ell^2$ for each $\ell \in \mathcal{S}_{AR}$, q_k, μ_k and σ_k^2 for each $k \in \mathcal{S}_{AR}^c$, and the switching parameters p_{ij} for $i, j \in \mathcal{S}$, $i \neq j$. The posterior distribution is accessed via Bayes' Theorem,

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{x})}, \tag{4.1}$$

where $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood, $\pi(\boldsymbol{\theta})$ the prior distribution, and $f(\boldsymbol{x})$ a normalising constant (with respect to $\boldsymbol{\theta}$). Until our algorithm in Section 3.2 there was no method to evaluate the likelihood for MRS models with independent regimes, and thus evaluating the posterior using Equation (4.1) was computationally infeasible. The likelihood can

be written as a marginal distribution

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{R}\in\mathcal{R}} f(\boldsymbol{x}, \boldsymbol{R}|\boldsymbol{\theta}) = \sum_{\boldsymbol{R}\in\mathcal{R}} f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{\theta}),$$

where $\mathcal{R}$ is the space of all possible regime sequences. Using this, the conditional distributions $f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{R})$ and $f(\boldsymbol{R}|\boldsymbol{\theta})$ are relatively straightforward to evaluate for any MRS model, and so a natural way to proceed is using *data-augmented* methods.

In a data-augmented framework the goal is to infer the joint posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x})$. Bayes' Theorem states that

$$f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{R}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{x})},$$

where we call $f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{R})$ the *conditional likelihood*, and $f(\boldsymbol{R}|\boldsymbol{\theta})$ is the probability of the regime sequence $\boldsymbol{R}$, given the transition probabilities in $\boldsymbol{\theta}$. The normalising constant, $f(\boldsymbol{x})$, is intractable and we can only use the proportional relationship

$$f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

and must resort to numerical methods to approximate posterior distributions. Our method of choice is a *blockwise data-augmented adaptive Metropolis-Hastings* algorithm, as presented in Section 4.2.

Recall from Section 3.2, the definition of the event

$$\{N_{t,i} = k\} = \{R_{t-1} \neq i, ..., R_{t-k+1} \neq i, R_{t-k} = i\}.$$

Then the conditional likelihood can be written as

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{R}, \boldsymbol{\theta}) = &\prod_{i\in\mathcal{S}} f(x_0|R_0 = i, \boldsymbol{\theta})^{\mathbb{I}(R_0=i)} \prod_{i\in\mathcal{S}_{AR}^c} \prod_{t=1}^{T} f(x_t|R_t = i, \boldsymbol{\theta})^{\mathbb{I}(R_t=i)} \\ &\times \prod_{i\in\mathcal{S}_{AR}} \prod_{t=1}^{T} \prod_{k=1}^{t} f(x_t|\boldsymbol{x}_{0:t-1}, R_t = i, N_{t,i} = k, \boldsymbol{\theta})^{\mathbb{I}(R_t=i, N_{t,i}=k)}, \end{aligned}$$

where the densities $f(x_0|R_0 = i, \boldsymbol{\theta})$, $i \in \mathcal{S}$, $f(x_t|R_t = i, \boldsymbol{x}_{0:t-1}, \boldsymbol{\theta})$, $i \in \mathcal{S}_{AR}^c$ and $f(x_t|\boldsymbol{x}_{0:t-1}, R_t = i, N_{t,i} = k, \boldsymbol{\theta})$, $i \in \mathcal{S}_{AR}$, are given by the MRS model specification (for example, Equation (3.19) or Equation (3.17)). The likelihood of the hidden sequence $\boldsymbol{R}$ is

$$f(\boldsymbol{R}|\boldsymbol{\theta}) = \mathbb{P}(R_0 = i) \prod_{i,j\in\mathcal{S}} p_{ij}^{\eta_{ij}},$$

where $\eta_{ij} = \sum_{t=1}^{T} \mathbb{I}(R_{t-1} = i, R_t = j)$, $i, j \in \mathcal{S}$, is the number of transitions from state i to state j in the sequence $\boldsymbol{R}$. Assuming $\{R_t\}$ is stationary, then $\mathbb{P}(R_0 = i)$ is given by the stationary probabilities.

Remark 4.1. Even with the forward algorithm of Section 3.2, there is still value in employing a data-augmented algorithm since the forward algorithm has complexity $\mathcal{O}(T^{k+1})$ (where T is the length of $\boldsymbol{x}$ and k is the number of AR(1) regimes), whereas data-augmented MCMC methods can be implemented with complexity $\mathcal{O}(T)$.

Prior distributions

In this thesis we choose to use a uniform prior distribution, $\pi(\boldsymbol{\theta}) \propto 1$, which is a type of *objective* prior distribution. Specifying a uniform prior distribution can be interpreted as, 'we are making no prior assumptions about the parameters before we have seen the data.' However, recall our note from Chapter 2, that this is not an entirely correct interpretation due to the fact that the uniform prior is not invariant to transformations. For example, if $\theta \sim$Uniform$(0, 1)$, then $-\log(\theta) \sim$Exponential(1). Note that our prior distributions are not always proper, since the support of $\pi(\cdot)$ can be unbounded. Table 4.1 details the uniform prior distributions we use.

Parameter	Prior distribution
α_i	$U(-\infty, \infty)$
ϕ_i	$U(-1, 1)$
σ_i^2	$U(0, \infty)$
q_i	$U(c_i, \infty)$, $c_i = \frac{1}{3}$, $\frac{2}{3}$ or 0.98 quantile of the data
μ_i	$U(-\infty, \infty)$
p_{ij}	$U(0, 1)$

TABLE 4.1: Prior distributions for the parameters of our MRS models. U denotes the uniform distribution. For q_i, when Regime i is the first shifted-log-normal distribution in the model, c_i is the $\frac{2}{3}$ quantile of the data, when Regime i is the second shifted-log-normal distribution in the model, c_i is the 0.98 quantile of the data, when Regime i is a frops regime and follows a distribution of the form $\log(q_i - X) \sim N(\mu_i, \sigma_i^2)$, then c_i is the $\frac{1}{3}$ quantile of the data.

4.2 MCMC implementation

As mentioned above, an approximate or numerical method must be used to obtain posterior inferences since the normalising constant $f(\boldsymbol{x})$ is not computable, and we resort to data-augmented MCMC algorithms for this. Recall from Section 2.2.5 that the idea behind data-augmented MCMC is to construct a Markov chain that has the joint

posterior distribution, $f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x})$, as its stationary distribution. Then, by simulating a long realisation of this process so that the process is close to stationary, samples towards the end of the simulated chain will be approximately distributed as the joint posterior distribution, $f(\boldsymbol{\theta}, \boldsymbol{R}|\boldsymbol{x})$. Denote the MCMC chain as $\left\{\boldsymbol{\psi}^{(n)}\right\}_{n\in\mathbb{N}} = \left\{\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right)\right\}_{n\in\mathbb{N}}$.

Now we describe the development of our MCMC algorithms, including unsuccessful attempts, since we believe these are still instructive.

Our initial MCMC algorithm Our first attempt was a data-augmented Metropolis-Hasting algorithm which proposes moves for $\boldsymbol{\theta}$ from the multivariate normal distribution $\mathrm{N}_p\left(\boldsymbol{\theta}^{(n)}, diag(\boldsymbol{s}^2\right)$, where $diag\left(\boldsymbol{s}^2\right)$ is a $p\times p$ diagonal matrix with s_i^2 along the diagonal. The terms s_i are tuning parameters to be specified. Moves for the hidden regime sequence $\boldsymbol{R}$ are proposed by simulating the process $\{R_t\}_{t\in\mathbb{N}}$ for $t = 0, 1, ..., T$, using the transition probabilities $p_{ij}^{(n)}$, $i, j \in \mathcal{S}$, which are elements of $\boldsymbol{\theta}^{(n)}$, and initial distribution $\mathbb{P}(R_0 = 1) = 1$, for simplicity. Moves are proposed to the entire vector $\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right)$ at once. The problem with this method is that the acceptance probability of proposed moves is small, with high probability, and the chain rarely moves.

A second MCMC implementation To overcome this we implemented a blockwise (element-at-a-time) Metropolis-Hastings (MH) algorithm, also known as a Metropolis-within-Gibbs algorithm. This algorithm iterates over elements of $\boldsymbol{\theta}$ and $\boldsymbol{R}$ sequentially, as a Gibbs sampler does.

Moves for elements of $\boldsymbol{\theta}$ are still made using a MH-style rule. That is, at the n^{th} iteration of the algorithm, to update the i^{th} element of $\boldsymbol{\theta}$, a move is proposed from a Normal distribution with mean $\theta_i^{(n)}$ and variance s_i^2, and the proposed move is accepted or rejected with an MH acceptance rule.

To sample the hidden sequence $\boldsymbol{R}$, we obtain the conditional posteriors

$$\mathbb{P}\left(R_t \middle| R_0, ..., R_{t-1}, R_{t+1}, ..., R_T, \boldsymbol{x}_{0:T}, \boldsymbol{\theta}^{(n)}\right), \tag{4.2}$$

in a similar way to Henneke *et al.* [47], and use a Gibbs sampler to sample the components R_t directly from these conditional posteriors.

This is different from our first implementation as at each step only one element of $\{\boldsymbol{\psi}^{(n)}\}$ is able to change, rather than trying to change the whole chain at once, and the Gibbs sampler accepts moves with probability 1. When implementing this algorithm on real data, we found constructing and sampling from the conditional posteriors in Equation (4.2) was taking the majority of the run time. This motivated us to try MH-style updates

for the components of $\boldsymbol{R}$. More specifically, to update the t^{th} component of $\boldsymbol{R}$, we sample uniformly from the set $\mathcal{S} \setminus \{R_t\}\}$ and accept or reject this with a MH acceptance rule.

In a component-wise algorithm such as this, the acceptance probability of proposed moves to the MCMC chain, $\{\boldsymbol{\psi}^{(n)}\}$, is generally higher than that in update-all-elements-at-a-time MH algorithms, so this algorithm is more practical than our first implementation.

The issue with this implementation is finding suitable tuning parameters s_i^2. It is not difficult to find tuning parameters s_i^2 that are adequate for a specific dataset, or model, using a trial and error approach. However, it is tedious. Over the course of this research we investigated numerous real and simulated datasets to validate our methods, and the algorithm had to be re-tuned often. This motivated us to use a Gibbs proposal for the parameters p_{ij}, and an adaptive-MH algorithm for the rest of the elements of the MCMC chain.

A third MCMC implementation A Gibbs proposal for the parameters p_{ij}, $i, j \in \mathcal{S}$, was chosen since the construction of the conditional proposal distributions for the parameters p_{ij} is rapid, and this eliminates any need to manually specify the proposal distribution. For MRS models our conditional proposal distribution is the same as that given in Henneke *et al.* [47]. The proposal distribution for the i^{th} row of the transition probability matrix of the hidden regime sequence is

$$f\left(p_{i1}, ..., p_{iM} \middle| \boldsymbol{R}^{(n)}, \boldsymbol{x}, \boldsymbol{\theta}_n\right) = f\left(p_{i1}, ..., p_{iM} \middle| \boldsymbol{R}^{(n)}\right) \sim \text{Dirichlet}\left(\eta_{i1} + 1, ..., \eta_{iM} + 1\right),$$

where M is the number of regimes in the model and, as before,

$$\eta_{ij} = \sum_{t=1}^{T} \mathbb{I}\left(R_{t-1}^{(n)} = i, R_t^{(n)} = j\right) \quad \text{for } i, j \in \mathcal{S}.$$

The rest of the parameter updates are executed using a blockwise-MH algorithm as before, except the tuning parameters s_i^2 are adaptively determined by the algorithm. The hidden sequence $\boldsymbol{R}$, is still updated as before, using blockwise-MH steps.

Adaptive steps The adaptive algorithm we employ is the *Adaptive-Metropolis* algorithm of Roberts and Rosenthal [90]. There is limited literature surrounding the optimal acceptance rate of MH algorithms for general posterior distributions. In [90], Roberts and Rosenthal provide an example of an adaptive scheme, which automatically adjusts the parameters s_i^2 to asymptotically reach a given acceptance rate while maintaining the

necessary ergodic properties. In [89], Roberts and Rosenthal prove, for an idealised version of our blockwise Metropolis-Hasting algorithm, an optimal acceptance rate is 0.44. Note that the theoretical results in [89] are derived for posterior distributions that are multivariate Normal. However, for our problems, it is unlikely that the posterior distributions are normal, and we additionally have to sample the hidden regime sequence, $\boldsymbol{R}$. Nonetheless, we use both their adaptive scheme and the proposed optimal acceptance rate, as these works well for our purposes.

Our implementation of Roberts and Rosenthal's adaptive scheme [90] is as follows. For each parameter, we initialise the standard deviation of the proposal distributions to $s_\ell = 1$, $\ell = 1, ..., 3M$, and begin our block-Metropolis-Hastings algorithm. After each batch of 50 iterations of the block-Metropolis-Hastings algorithm, we update s_ℓ, by multiplying by $\exp(\delta(n))$ if the acceptance rate for that parameter is above 0.44, or by $\exp(-\delta(n))$ if the acceptance rate is less than 0.44. Following the ideas in [90], we define

$$\delta(n) = \min\left(\frac{2}{\sqrt{n}}, \frac{10}{n}, \frac{10000}{n^2}\right).$$

Note that to satisfy the conditions for convergence of this algorithm outlined in [90], we also need to specify a bound $K < \infty$ and restrict $\log(s_i)$ to $[-K, K]$. In our implementation this bound is not needed, since we stop the adaptive iterations after some specified burn-in. We observe that after a sufficient number of iterations the sequence of s_i created by this algorithm converges to a fixed value, and that the acceptance rate is close enough to 0.44 for our purpose. Stopping the adaptive steps after some burn-in period also has the advantage of making the algorithm output easier to interpret. The stochastic process produced by the algorithm during the adaptive steps is no longer a Markov chain, since transitions depend on all previous values of the chain. By stopping the adaptive steps and only considering the process from this point on, the resulting process is a time-homogeneous Markov chain.

One last improvement We also found computational savings could be made when sampling the hidden regime sequence. When the characteristics of each regime in the model are sufficiently different the hidden regime sequence, $\boldsymbol{R}$, is relatively obvious, in that the posterior probability $\mathbb{P}(R_t = i|\boldsymbol{x}, \boldsymbol{\theta})$, is close to 1 or 0 for most values of $\boldsymbol{\theta}$, and the variance of $\boldsymbol{R}$ in the posterior distribution is low. We found it much more computationally efficient to update only a subsample of the hidden regimes at each iteration of the algorithm. Although this makes the chain mix slower, this is typically outweighed by the computational savings made. For some datasets, we found that we needed to update as little as 1% of the hidden regimes and the algorithm still converged within $\approx 100,000$ iterations. We settled on updating (a conservative) 10% of the hidden

regime at each iteration of the MCMC algorithm, as this performed well for most of the datasets we investigated.

The end result is a flexible and efficient MCMC algorithm to sample from the posterior distribution of MRS models, which is a *blockwise data-augmented adaptive Metropolis-Hastings* algorithm. We assess convergence of our algorithm by comparing trace plots of four independent chains and observe when they show stationary characteristics. In the following, we give more details behind the efficient computation in certain aspects of the algorithm.

Simplifying the MH ratios for fast computation

Here we detail some intricacies for efficient computation of the MH-ratio when updating some elements of the MCMC chain $\boldsymbol{\psi}^{(n)}$. We show that evaluating the whole conditional likelihood can be avoided when computing the MH-ratio.

Within-regime parameter updates By within-regime we mean the parameter α_i, ϕ_i, σ_i^2, q_i or μ_i. In our algorithm, the MCMC chain updates for all of the within-regime parameters are made by first proposing a move from the one-dimensional Normal distribution, centred around the current value and with variance s_ℓ^2. First we note that, the proposal distributions are symmetric, so they cancel out in the MH acceptance ratio.

Now, suppose the current state of the MCMC chain is $\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right)$, and we are to update the ℓ^{th} element of $\boldsymbol{\theta}$, which belongs to Regime j. If the algorithm proposes a move to $\boldsymbol{\theta}' = \left(\theta_1^{(n)}, ..., \theta_{\ell-1}^{(n)}, \theta', \theta_{\ell+1}^{(n)}, ..., \theta_p^{(n)}\right)$, then the MH-ratio is

$$
\begin{aligned}
\alpha(x, y) &= \frac{f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}', \boldsymbol{R}^{(n)}\right) f\left(\boldsymbol{R}^{(n)}\middle|\boldsymbol{\theta}'\right) \pi\left(\boldsymbol{\theta}'\right)}{f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right) f\left(\boldsymbol{R}^{(n)}\middle|\boldsymbol{\theta}^{(n)}\right) \pi\left(\boldsymbol{\theta}^{(n)}\right)} \\
&= \frac{f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}', \boldsymbol{R}^{(n)}\right) \pi\left(\boldsymbol{\theta}'\right)}{f\left(\boldsymbol{x}\middle|\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right) \pi\left(\boldsymbol{\theta}^{(n)}\right)} \\
&= \frac{f\left(x_0\middle|\boldsymbol{\theta}', R_0^{(n)} = j\right)^{\mathbb{I}\left(R_0^{(n)}=j\right)} \prod_{t=1}^{T} f\left(x_t\middle|\boldsymbol{x}_{0:t-1}, \boldsymbol{\theta}', \boldsymbol{R}^{(n)}\right)^{\mathbb{I}\left(R_t^{(n)}=j\right)}}{f\left(x_0\middle|\boldsymbol{\theta}_n, R_0^{(n)} = j\right)^{\mathbb{I}\left(R_0^{(n)}=j\right)} \prod_{t=1}^{T} f\left(x_t\middle|\boldsymbol{x}_{0:t-1}, \boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right)^{\mathbb{I}\left(R_t^{(n)}=j\right)}} \frac{\pi\left(\boldsymbol{\theta}'\right)}{\pi\left(\boldsymbol{\theta}^{(n)}\right)},
\end{aligned}
$$

where the second equality holds since the parameters p_{ij} are the same in both $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}^{(n)}$. Depending on the form on the densities $f\left(x_0\middle|R_0 = j, \boldsymbol{\theta}\right)$ and $f\left(x_t\middle|\boldsymbol{x}_{0:t-1}, \boldsymbol{\theta}, \boldsymbol{R}\right)$ this can possibly be simplified further, but that is model specific.

Hidden regime sequence updates Suppose the current value of the MCMC chain is $\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{R}^{(n)}\right)$. Updates to $\boldsymbol{R}^{(n)}$ are proposed by first sampling $r := \lceil 0.1T + 1 \rceil$ indices from $\{0, 1, ..., T\}$. Label this sample of indices $\boldsymbol{\tau} = \{\tau_1, ..., \tau_r\}$. Let τ be any element of $\boldsymbol{\tau}$. For each τ, a move is proposed to the τ^{th} element of $\boldsymbol{R}$ in the following way. Suppose the τ^{th} element of $\boldsymbol{R}^{(n)}$ is $R_\tau^{(n)} = i$, then sample an alternative regime, j, uniformly from $\mathcal{S} \setminus \{i\}$. Set $\boldsymbol{R}' = \left(R_0^{(n)}, ..., R_{\tau-1}^{(n)}, j, R_{\tau+1}^{(n)}, ..., R_T^{(n)}\right)$, and let $m = R_{\tau-1}$ and $\ell = R_{\tau+1}$.

The most complex case is when $i, j \in \mathcal{S}_{AR}$, and the other cases are simplifications of this, so we treat this first. When $i, j \in \mathcal{S}_{AR}$ the relevant terms in the conditional likelihoods concern AR(1) processes, which require knowledge of last visit and next visit times. We use the notation $t - b_j^t$, $j \in \mathcal{S}$, to denote the time of the last visit to state j, before time t, in the sequence $\boldsymbol{R}^{(n)}$, and we use $t + a_j^t$, $j \in \mathcal{S}$, to denote the time of the next visit to state j, after time t, in the sequence $\boldsymbol{R}^{(n)}$ (a for *after*, b for *before*). If there is no last visit time, b_i^τ, then set $b_i^\tau = t + 1$. If there is no next visit time, a_i^τ, then define

$$\begin{aligned}
&f(x_{\tau+a_i^\tau} | R_{\tau+a_i^\tau} = i, N_{\tau+a_i^\tau,i} = a_i^\tau + b_i^\tau, \boldsymbol{x}_{0:\tau+a_i^\tau-1}, \boldsymbol{\theta}^{(n)}) \\
&= f(x_{\tau+a_i^\tau} | R_{\tau+a_i^\tau} = i, N_{\tau+a_i^\tau,i} = a_i^\tau, \boldsymbol{x}_{0:\tau+a_i^\tau-1}, \boldsymbol{\theta}^{(n)}) \\
&= 1.
\end{aligned}$$

For simplicity, first assume $\tau \neq 0$ and $\tau \neq T$, then the MH ratio is

$$\begin{aligned}
\alpha(x, y) = &\frac{f\left(x_\tau \middle| R_\tau = j, N_{\tau,j} = b_j^\tau, \boldsymbol{x}_{0:\tau-1}, \boldsymbol{\theta}^{(n)}\right)}{f\left(x_\tau \middle| R_\tau = i, N_{\tau,i} = b_i^\tau, \boldsymbol{x}_{0:\tau-1}, \boldsymbol{\theta}^{(n)}\right)} \\
&\times \frac{f\left(x_{\tau+a_j^\tau} \middle| R_{\tau+a_j^\tau} = j, N_{\tau+a_j^\tau,j} = a_j^\tau, \boldsymbol{x}_{0:\tau+a_j^\tau-1}, \boldsymbol{\theta}^{(n)}\right)}{f\left(x_{\tau+a_i^\tau} \middle| R_{\tau+a_i^\tau} = i, N_{\tau+a_i^\tau,i} = a_i^\tau, \boldsymbol{x}_{0:\tau+a_i^\tau-1}, \boldsymbol{\theta}^{(n)}\right)} \\
&\times \frac{f\left(x_{\tau+a_i^\tau} \middle| R_{\tau+a_i^\tau} = i, N_{\tau+a_i^\tau,i} = a_i^\tau + b_i^\tau, \boldsymbol{x}_{0:\tau+a_i^\tau-1}, \boldsymbol{\theta}^{(n)}\right)}{f\left(x_{\tau+a_j^\tau} \middle| R_{\tau+a_j^\tau} = j, N_{\tau+a_j^\tau,j} = a_j^\tau + b_j^\tau, \boldsymbol{x}_{0:\tau+a_j^\tau-1}, \boldsymbol{\theta}^{(n)}\right)} \frac{p_{mj}^{(n)} p_{j\ell}^{(n)}}{p_{mi}^{(n)} p_{i\ell}^{(n)}}.
\end{aligned} \tag{4.3}$$

The idea behind the derivation of Equation (4.3) is, since both i and j are AR(1) regimes, then $f\left(\boldsymbol{x} \middle| \boldsymbol{R}', \boldsymbol{\theta}^{(n)}\right)$ and $f\left(\boldsymbol{x} \middle| \boldsymbol{R}^{(n)}, \boldsymbol{\theta}^{(n)}\right)$ differ for terms that involve $x_{\tau-b_j^\tau}$, $x_{\tau-b_i^\tau}$, x_τ, $x_{\tau+a_i^\tau}$ and $x_{\tau+a_j^\tau}$ only. Specifically, given $\boldsymbol{R}'$, $x_{\tau+a_j^\tau}$ depends on x_τ, x_τ depends on $x_{\tau-b_j^\tau}$, and $x_{\tau+a_i^\tau}$ depends on $x_{\tau-b_i^\tau}$, and the relevant densities for these are

$$f\left(x_{\tau+a_j^\tau} \middle| R_{\tau+a_j^\tau} = j, N_{\tau+a_j^\tau,j} = a_j^\tau, \boldsymbol{x}_{0:\tau+a_j^\tau-1}, \boldsymbol{\theta}^{(n)}\right),$$

and

$$f\left(x_\tau \middle| R_\tau = j, N_{\tau,j} = b_j^\tau, \boldsymbol{x}_{0:\tau-1}, \boldsymbol{\theta}^{(n)}\right)$$

and

$$f\left(x_{\tau+a_i^\tau}\middle|R_{\tau+a_i^\tau}=i,N_{\tau+a_i^\tau,i}=a_i^\tau+b_i^\tau,\boldsymbol{x}_{0:\tau+a_i^\tau-1},\boldsymbol{\theta}^{(n)}\right).$$

Given $\boldsymbol{R}^{(n)}$, $x_{\tau+a_i^\tau}$ depends on x_τ, x_τ depends on $x_{\tau-b_i^\tau}$, and $x_{\tau+a_j^\tau}$ depends on $x_{\tau-b_j^\tau}$, and the relevant densities for these are

$$f\left(x_{\tau+a_i^\tau}\middle|R_{\tau+a_i^\tau}=i,N_{\tau+a_i^\tau,i}=a_i^\tau,\boldsymbol{x}_{0:\tau+a_i^\tau-1},\boldsymbol{\theta}^{(n)}\right),$$

and

$$f\left(x_\tau\middle|R_\tau=i,N_{\tau,i}=b_i^\tau,\boldsymbol{x}_{0:\tau-1},\boldsymbol{\theta}^{(n)}\right)$$

and

$$f\left(x_{\tau+a_j^\tau}\middle|R_{\tau+a_j^\tau}=j,N_{\tau+a_j^\tau,j}=a_j^\tau+b_j^\tau,\boldsymbol{x}_{0:\tau+a_j^\tau-1},\boldsymbol{\theta}^{(n)}\right).$$

When $\tau=0$ or T, a modifications to the terms p_{ij} is needed. When $\tau=0$ the ratio becomes

$$\begin{aligned}\alpha(x,y)=&\frac{f\left(x_\tau\middle|R_\tau=j,N_{\tau,j}=b_j^\tau,\boldsymbol{x}_{0:\tau-1},\boldsymbol{\theta}^{(n)}\right)}{f\left(x_\tau\middle|R_\tau=i,N_{\tau,i}=b_i^\tau,\boldsymbol{x}_{0:\tau-1},\boldsymbol{\theta}^{(n)}\right)}\\&\times\frac{f\left(x_{\tau+a_j^\tau}\middle|R_{\tau+a_j^\tau}=j,N_{\tau+a_j^\tau,j}=a_j^\tau,\boldsymbol{x}_{0:\tau+a_j^\tau-1},\boldsymbol{\theta}^{(n)}\right)}{f\left(x_{\tau+a_i^\tau}\middle|R_{\tau+a_i^\tau}=i,N_{\tau+a_i^\tau,i}=a_i^\tau,\boldsymbol{x}_{0:\tau+a_i^\tau-1},\boldsymbol{\theta}^{(n)}\right)}\\&\times\frac{f\left(x_{\tau+a_i^\tau}\middle|R_{\tau+a_i^\tau}=i,N_{\tau+a_i^\tau,i}=a_i^\tau+b_i^\tau,\boldsymbol{x}_{0:\tau+a_i^\tau-1},\boldsymbol{\theta}^{(n)}\right)}{f\left(x_{\tau+a_j^\tau}\middle|R_{\tau+a_j^\tau}=j,N_{\tau+a_j^\tau,j}=a_j^\tau+b_j^\tau,\boldsymbol{x}_{0:\tau+a_j^\tau-1},\boldsymbol{\theta}^{(n)}\right)}\frac{p_j^{(n)}p_{j\ell}^{(n)}}{p_i^{(n)}p_{i\ell}^{(n)}},\end{aligned}\tag{4.4}$$

where $p_j^{(n)}$, $j\in\mathcal{S}$, is the stationary distribution of $\boldsymbol{R}$ given $\boldsymbol{\theta}^{(n)}$. When $\tau=T$ the ratio becomes

$$\alpha(x,y)=\frac{f\left(x_\tau\middle|R_\tau=j,N_{\tau,j}=b_j^\tau,\boldsymbol{x}_{0:\tau-1},\boldsymbol{\theta}^{(n)}\right)}{f\left(x_\tau\middle|R_\tau=i,N_{\tau,i}=b_i^\tau,\boldsymbol{x}_{0:\tau-1},\boldsymbol{\theta}^{(n)}\right)}\frac{p_{mj}^{(n)}}{p_{mi}^{(n)}}.\tag{4.5}$$

Now, note that for $k\notin S_{AR}$, then

$$\begin{aligned}&f(x_{\tau+a_k^\tau}|R_{\tau+a_k^\tau}=k,N_{\tau+a_k^\tau,k}=a_k^\tau+b_k^\tau,\boldsymbol{x}_{0:\tau+a_k^\tau-1},\boldsymbol{\theta}^{(n)})\\&=f(x_{\tau+a_k^\tau}|R_{\tau+a_k^\tau}=k,N_{\tau+a_k^\tau,k}=a_k^\tau,\boldsymbol{x}_{0:\tau+a_k^\tau-1},\boldsymbol{\theta}^{(n)})\\&=f(x_{\tau+a_k^\tau}|R_{\tau+a_k^\tau}=k,\boldsymbol{\theta}^{(n)}),\end{aligned}$$

since k is an i.i.d. regime, thus, given the regime is k, then $x_{\tau+a_k^\tau}$ is independent of $N_{\tau+a_k^\tau,k}=a_k^\tau$, and $\boldsymbol{x}_{0:\tau+a_k^\tau-1}$, where $N_{t,k}$ for $k\in\mathcal{S}$, and $t=0,...,T$, are random variables denoting the time since the last visit to state k at time t. So when either one,

or both, of i and j are in $\mathcal{S}_{AR}^{c}$ then the ratios (4.3), (4.4) and (4.5) simplify, since some, or all, of the terms involving $x_{\tau+a^\tau}$ terms cancel out.

4.3 Posterior predictive checks

The general idea behind posterior predictive checks (PPCs) is that data replicated under the *Bayesian posterior predictive distribution* should look similar to the observed data. By comparing statistics generated under the posterior predictive distribution, to statistics calculated from the observed data, we can see where a model fails. The Bayesian posterior predictive distribution is

$$f(\boldsymbol{x}_{\text{new}}|\boldsymbol{x}) = \int_{\Theta} \sum_{\boldsymbol{R}\in\mathcal{R}} f(\boldsymbol{x}_{\text{new}}|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{R}) f(\boldsymbol{\theta},\boldsymbol{R}|\boldsymbol{x}) d\boldsymbol{\theta},$$

where $\boldsymbol{x}$ is the observed data, $\boldsymbol{R}$ is the hidden regime sequence, and $\boldsymbol{x}_{\text{new}}$ is data generated independently of $\boldsymbol{x}$. The statistics used to compare observed data to the posterior predictive distribution in PPCs are typically related to characteristics of the data that we want to capture. That is, if an important aspect of the model is to capture the variance of the data, then the sample variance of the observed data should be compared to the variance of the posterior predictive distribution, or, compared to the sample variance of data replicated by the posterior predictive distribution in the case that the variance of the posterior predictive distribution is not known.

For our MRS models, suppose we are interested in some statistic $T(\boldsymbol{x},(\boldsymbol{\theta},\boldsymbol{R}))$ which depends on observed data, parameters and the hidden regime sequence. To construct a PPC:

Step 1. Sample $\boldsymbol{\theta}^*$ and $\boldsymbol{R}^*$ from the posterior distribution.

Step 2. Calculate $T(\boldsymbol{x},(\boldsymbol{\theta}^*,\boldsymbol{R}^*))$ from observed data.

Step 3. Determine the statistic's true value under the distribution $f(\cdot|\boldsymbol{\theta}^*,\boldsymbol{R}^*)$. If this cannot be done analytically then we can approximate the true value by simulating $\boldsymbol{x}_{\text{new}}$ from $f(\cdot|\boldsymbol{\theta}^*,\boldsymbol{R}^*)$, and use this sample to calculate the relevant statistic, $T(\boldsymbol{x}_{\text{new}},(\boldsymbol{\theta}^*,\boldsymbol{R}^*))$.

Step 4. Compare the statistics $T(\boldsymbol{x}_{\text{new}},(\boldsymbol{\theta}^*,\boldsymbol{R}^*))$ and $T(\boldsymbol{x},(\boldsymbol{\theta}^*,\boldsymbol{R}^*))$.

This is repeated for many samples of $\boldsymbol{R}^*$ and $\boldsymbol{\theta}^*$, and we look to see if, overall, the statistics calculated from the observed data and the statistics calculated from the predictive distribution disagree in any significant way. For example, in the case that $T(\boldsymbol{x},(\boldsymbol{\theta},\boldsymbol{R}))$ is

a scalar, one can determine the proportion of times $T(\boldsymbol{x}, (\boldsymbol{\theta}^*, \boldsymbol{R}^*))$ exceeds $T(\boldsymbol{x}, (\boldsymbol{\theta}, \boldsymbol{R}))$ as a measure of how well the model and data agree.

PPCs are a very flexible tool as they can be applied to a wide range of statistics T. PPCs are able to notify us where a model might obviously be failing, however, like any statistical process, they cannot tell us if our model is definitely correct.

PPCs were proposed by Rubin in 1984, [93]; Chapter 6 of Gelman *et al.* [37], is also useful for this topic.

Some useful posterior predictive checks

Here we describe how we construct some posterior predictive checks used in our analysis. Note that these procedures are repeated for many samples from the posterior distribution, and suitability of a model is assessed using *all* these samples.

QQ plots To test the distributional assumptions of each regime, samples of $\boldsymbol{R}^*$ from the posterior distribution can be used to classify data into regimes and *quantile-quantile* (QQ) plots for each regime can be generated. QQ plots display empirical quantiles, calculated from observed data, versus theoretical quantiles, calculated as if the model were true. If the distributional assumption underlying the model is true, we expect to see the data in the QQ plot to follow a straight line, and deviations from this suggests that the model may not be correct. For i.i.d. regimes this is straightforward to implement once the data has been classified by $\boldsymbol{R}^*$, since these observations are i.i.d. and $\boldsymbol{\theta}^*$ defines their theoretical distribution. For AR(1) regimes, the *residuals* of the AR(1) regime are calculated using the sampled $\boldsymbol{R}^*$ and $\boldsymbol{\theta}^*$. Since our models assume Gaussian AR(1) processes, the theoretical distribution of the residuals is $\mathrm{N}(0, 1)$ and QQ plots can then be generated. More specifically, suppose Regime i is an AR(1) regime, then, for Type II models, the residuals are calculated as

$$r_t^* = \frac{x_t - \alpha_i^* \dfrac{1 - (\phi_i^*)^\ell}{1 - (\phi_i^*)} - (\phi_i^*)^\ell x_{t-\ell}}{\sigma_i^* \left(\dfrac{1 - (\phi_i^*)^{2\ell}}{1 - (\phi_i^*)^2} \right)^{1/2}}, \tag{4.6}$$

where x_t has been classified into Regime i by $\boldsymbol{R}^*$, and $N_{t,i}^* = \ell$. Equation (4.6) follows from Equation (3.18), where we show the mean and variance of x_t in Regime i, given

the last observed value from Regime i, $x_{t-\ell}$, are

$$\alpha_i^* \frac{1-(\phi_i^*)^\ell}{1-(\phi_i^*)} + (\phi_i^*)^\ell x_{t-\ell}, \quad \text{and} \quad (\sigma_i^*)^2 \left(\frac{1-(\phi_i^*)^{2\ell}}{1-(\phi_i^*)^2} \right),$$

respectively.

For Type III models,

$$r_t^* = \frac{x_t - \alpha_i^* - \phi_i^* x_{t-\ell}}{\sigma_i^*},$$

where x_t and $x_{t-\ell}$ are defined as before.

Since the main goal of our models is often to model the distribution of prices, this is one of the more important PPCs that we use. If the QQ plots deviate from what we expect in a systematic way, for a collection of samples from the posterior, this suggests there is an issue with the distributional assumptions.

Residual plots The other posterior predictive checks that we use plot, for each regime, residuals against time and residuals against lagged values. Since our MRS model assumes the variance within each regime is constant across time, plotting residuals against time can warn us if there is any significant time-heterogeneity present.

Similarly, for AR(1) regimes, our models assume constant variance with respect to lagged values; that is, for an AR(1) process, $\{Y_t\}$, we assume the variance of Y_t does not depend on Y_{t-1}. Plotting residuals against lagged values can warn us if this assumption is violated. Some analyses in the literature [58, 61] have found electricity prices can follow a constant elasticity of variance (CEV) process, $Y_t = \alpha + \phi Y_{t-1} + \sigma |Y_{t-1}|^\gamma \varepsilon_t$ for some non-zero γ. To assess the assumption of constant variance with respect to lagged values, and therefore reject the need for a CEV model, we use a *scale-location* plot, which plots $\sqrt{|r_t|}$ from AR(1) regimes against the magnitude of the last value before time t from the same regime, $|x_{t-\ell}|$.

4.4 Validation of methods

To validate our methods we simulated datasets from MRS models and used our Bayesian methodology to estimate parameters from the simulations. We also produced PPCs for each simulated dataset to observe their behaviour. In the following we present only simulations of MRS models of Type II, since the conclusions for models of Type III are exactly the same. The parameters used in the following simulations are chosen to

approximately match the parameters estimated from the South Australian electricity market data, as explored in Chapter 5.

4.4.1 When the model fitted to data is correct

We simulated twenty datasets of length $T = 2000$ from the following MRS model of Type II (with independent regimes which evolve at all time points):

$$X_t = \begin{cases} B_t, & R_t = 1, \\ S_t, & R_t = 2, \end{cases} \tag{4.7}$$

where

$$B_t = 0.55B_{t-1} + \sqrt{53}\varepsilon_t$$

is an AR(1) process, $\{\varepsilon_t\}$ is a sequence of i.i.d. N(0,1) random variables, $\{S_t\}$ is a sequence of shifted-log-normal random variables, i.e. $\log(S_t - 12) \sim$ i.i.d. N(3.5,1), and $\{R_t\}$ is a Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}.$$

We used our Bayesian methodology to fit the correct model to the data by estimating the posterior distributions, and produced PPCs. The posterior means, medians, and univariate marginal modes are summarised in boxplots in Figure 4.1. It appears that the point estimates of ϕ and σ_1^2 are biased. In particular, posterior point estimates of ϕ are biased towards 0, and the posterior point estimates of σ_1^2 are biased upwards. Other simulations have also shown this behaviour. Our hypothesis is that the prior distribution is affecting the posterior inferences. The prior distribution on σ_1^2 is the improper uniform distribution on $(0, \infty)$, and assigns equal weight to every value on the positive half-line, no matter how large, and we hypothesise this biases point estimates of σ_1^2 upwards. Similarly, the prior distribution for ϕ is uniform on $(-1, 1)$, which has mean zero, and we hypothesise this has a shrinkage affect on these point estimates of ϕ, pulling them closer to 0.

To investigate this further, we simulated datasets of length $T = 4000$ from the same model, applied our Bayesian methods and produced the same boxplot summary of posterior point estimates, shown in Figure 4.2. Notice that the bias in these point estimates is smaller than before. This suggests that these point estimates are at least consistent.

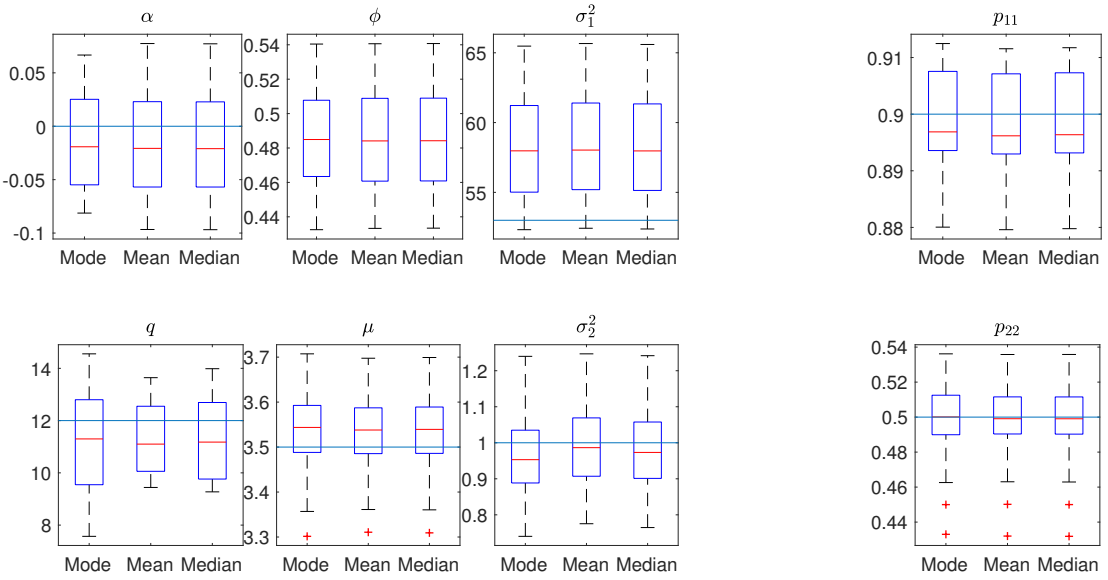


FIGURE 4.1: Boxplots summarising Bayesian posterior point estimates of the parameters of the model in Equation (4.7) for twenty simulated datasets of length $T = 2000$, when the correct model is fitted to the data. The true parameters are $\alpha = 0$, $\phi = 0.55$, $\sigma_1^2 = 53$, $q = 12$, $\mu = 3.5$, $\sigma_2^2 = 1$, $p_{11} = 0.9$ and $p_{22} = 0.5$. The mode is the univariate *marginal* mode. Note the apparent bias in the point estimates of ϕ and σ_1^2.

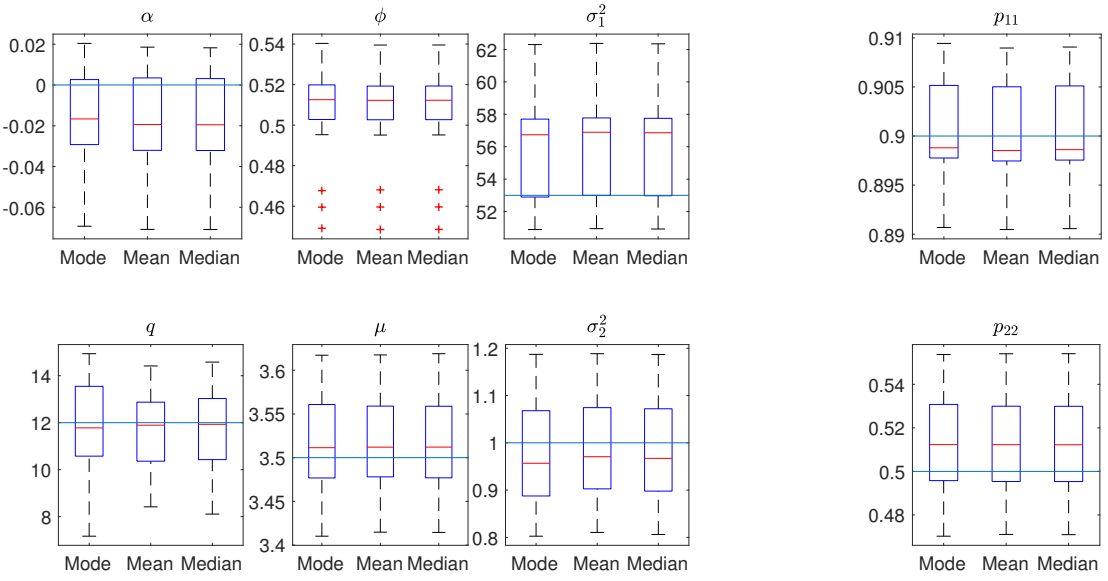


FIGURE 4.2: Boxplots summarising Bayesian posterior point estimates of the parameters of the model in Equation (4.7) for twenty simulated datasets of length $T = 4000$, when the correct model is fitted to the data. The true parameters are $\alpha = 0$, $\phi = 0.55$, $\sigma_1^2 = 53$, $q = 12$, $\mu = 3.5$, $\sigma_2^2 = 1$, $p_{11} = 0.9$ and $p_{22} = 0.5$. The mode is the univariate *marginal* mode. Notice the bias in the parameters ϕ and σ_1^2 is smaller here than in Figure 4.1.

In Figures 4.3-4.5 the univariate marginal posterior distributions are plotted for the simulations of length $T = 2000$. We see the marginal distributions for all parameters, except q, are approximately symmetric.

In Figure 4.6 our QQ plot-PPCs are shown for the simulations of length $T = 2000$. We use these to assess within-regime distributional assumptions. The model fitted to the data is correct (it has one AR(1) regime and one i.i.d. shifted-log-normal regime, which is the same as the model that generated the data), so we expect to see our PPCs reflect this. Since the points on the QQ plots lie relatively close to a straight line, this PPC suggests that the within-regime distributional assumptions are reasonable. Note that in Figure 4.6 each pair of plots is generated by an independently simulated dataset, and a single sample of $\boldsymbol{\theta}$ and $\boldsymbol{R}$ from the posterior. This means there are two sources of variability in this figure, variability in the dataset, and variability in the sampled parameters from the posterior. Since these QQ plot PPCs are our main model checking tool, we also investigate the variability of these plots due to the posterior only, that is, for a fixed dataset and many samples from the posterior. In Figure 4.7 ten QQ plot PPCs are show for a fixed dataset.

Figure 4.8 shows the residuals-versus-time plots produced as part of our PPCs. The fitted model assumes the variance within each regime does not vary over time. Since there is no obvious pattern in these residuals plots, this PPC suggests that this assumption is reasonable.

Figure 4.9 shows the scale-location PPCs. The fitted model assumes the variance does not depend on the magnitude of the lagged realisations from each regime. Since Figure 4.9 shows no obvious increase or decrease in the *spread* of the residuals as the magnitude of lagged values increases or more generally any significant shape, this PPC suggests that constant variance is appropriate.

4.4.2 Fitting a model with an incorrect spike regime

To see if our methods have any power to reject a model with an incorrect spike distribution, we simulated datasets of length $T = 2000$ from the model in Equation (4.7) and used our Bayesian methodology to fit a MRS model of Type II with one auto-regressive regime and one i.i.d. shifted-Gamma regime.

In Figure 4.10 ten pairs of the QQ plot-PPCs are shown for these simulations. Each pair of plots was produced from a different simulated dataset. Since the model fitted to the data has an i.i.d. shifted-Gamma regime, instead of the i.i.d. shifted-log-normal regime that was simulated, we expect this PPC to suggest the spike distribution is not

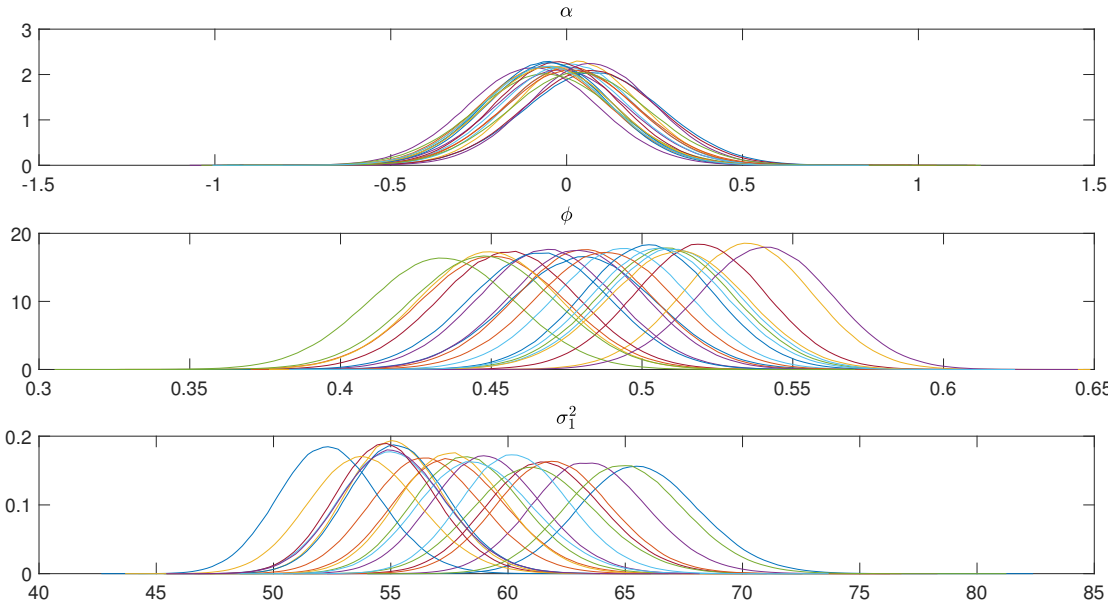


FIGURE 4.3: Univariate marginal posterior distributions for the parameters of Regime 1 (the AR(1) regime) constructed from twenty simulated datasets of length 2000 of the model in Equation (4.7), when the correct model is fitted to the data. There is one marginal posterior density curve for each simulated dataset. The true parameter values are $\alpha = 0$, $\phi = 0.55$, and $\sigma_1^2 = 53$.

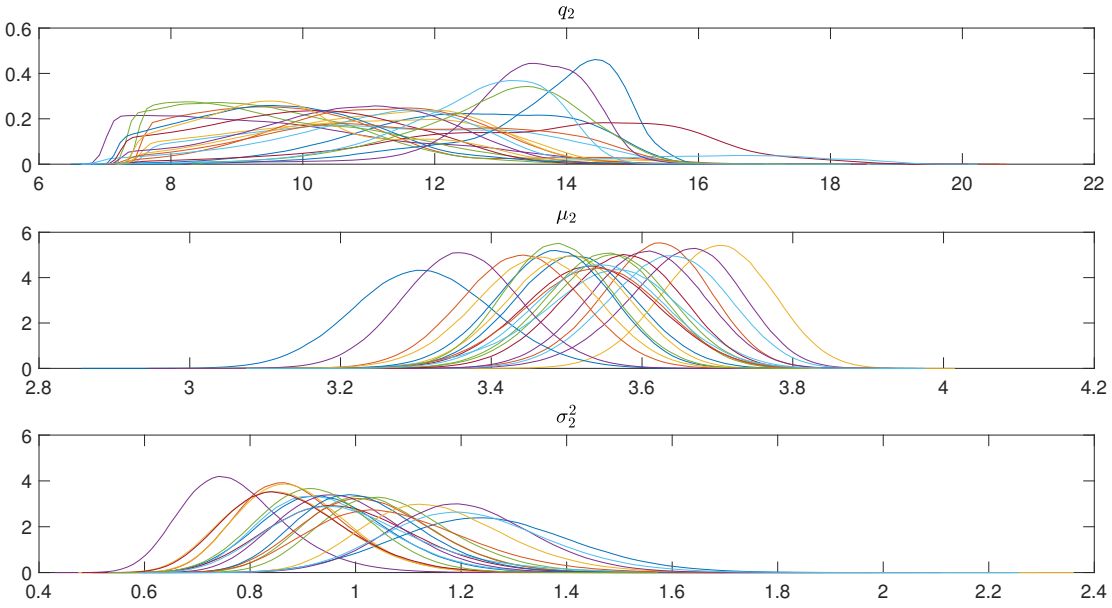


FIGURE 4.4: Univariate marginal posterior distributions for the parameters of Regime 2 (the shifted-log-normal regime) constructed from twenty simulated datasets of length 2000 of the model in Equation (4.7), when the correct model is fitted to the data. There is one marginal posterior density curve for each simulated dataset. The true parameter values are $q_2 = 12$, $\mu_2 = 3.5$, and $\sigma_2^2 = 1$. Notice that the marginal posterior distribution for q_2 is not symmetric, and the marginal posterior distributions for μ_2 and σ_2^2 are approximately symmetric.

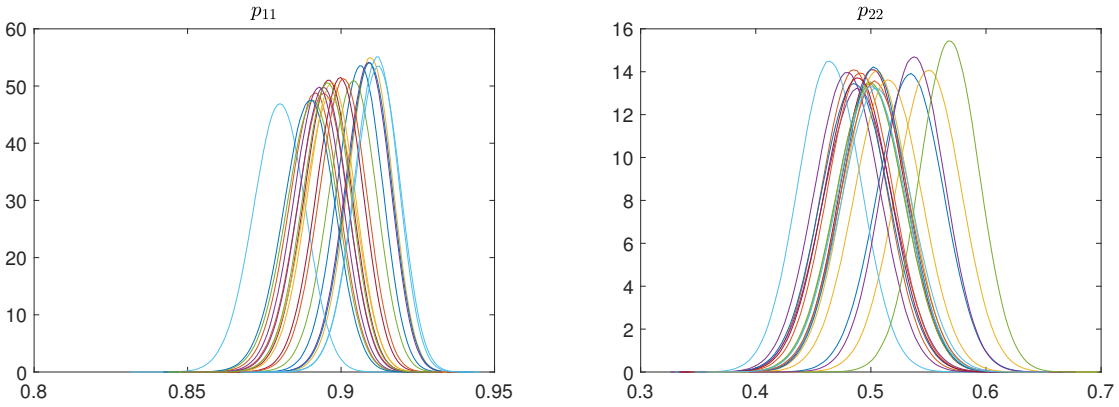


FIGURE 4.5: Univariate marginal posterior distributions for the parameters of the transition matrix constructed from twenty simulated datasets of the model in Equation (4.7), when the correct model is fitted to the data. There is one marginal posterior density curve for each simulated dataset. The true parameter values are $p_{11} = 0.9$ and $p_{22} = 0.5$.

suitable. Notice that points in the QQ plots from the shifted-Gamma regime do not follow a straight line, which suggests the Gamma regime is indeed unsuitable in this model. This in turn implies that the QQ plot-PPC is able to distinguish between the shifted-Gamma distribution and the shifted-log-normal distribution.

4.4.3 Fitting a constant variance model to data with non-constant variance

To see if our methods have any power to reject a model with incorrect homoscedasticity assumptions, we simulated datasets of length $T = 2000$ from the following CEV model

$$X_t = \begin{cases} B_t, & R_t = 1, \\ S_t, & R_t = 2, \end{cases} \tag{4.8}$$

where

$$B_t = 0.55B_{t-1} + \sqrt{53}|B_{t-1}|^{\gamma}\varepsilon_t$$

is an AR(1) process, and $\{S_t\}$ is a sequence of shifted-log-normal random variables, i.e. $\log(S_t - 12)$ i.i.d. N(3.5,1), and $\{R_t\}$ is a Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix},$$

for $\gamma = -0.5, -0.25, 0.25$ and 0.5. We then used our Bayesian methodology to fit the MRS model of Type II in Equation (4.7) (a constant variance model) to each dataset, and produced the scale-location PPCs. We expected to see evidence in the scale-location plots that the fitted model is incorrect.

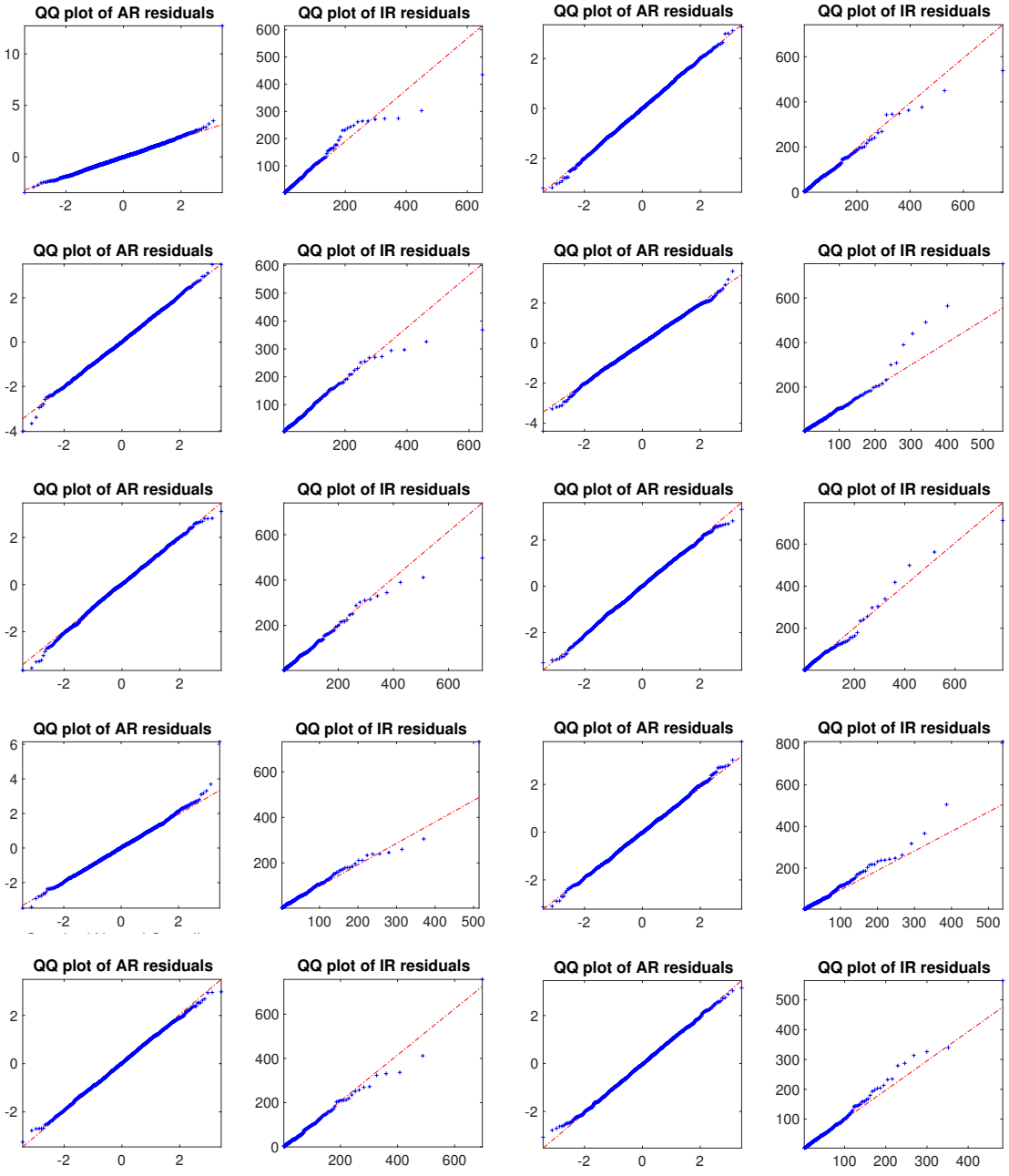


FIGURE 4.6: Ten pairs of QQ plots generated as part of our PPCs for simulations of the model in Equation (4.7), when the correct model is fitted to the data. Each pair of plots was generated by an independently simulated dataset. These PPCs are used to assess within-regime distributional assumptions. In each pair, the plot on the left is the QQ plot of the residuals of the AR(1) regime, and the plot on the right is for the shifted-log-normal regime. Since the points on each QQ plot are relatively close to the reference line, this PPC does not reject the fitted model.

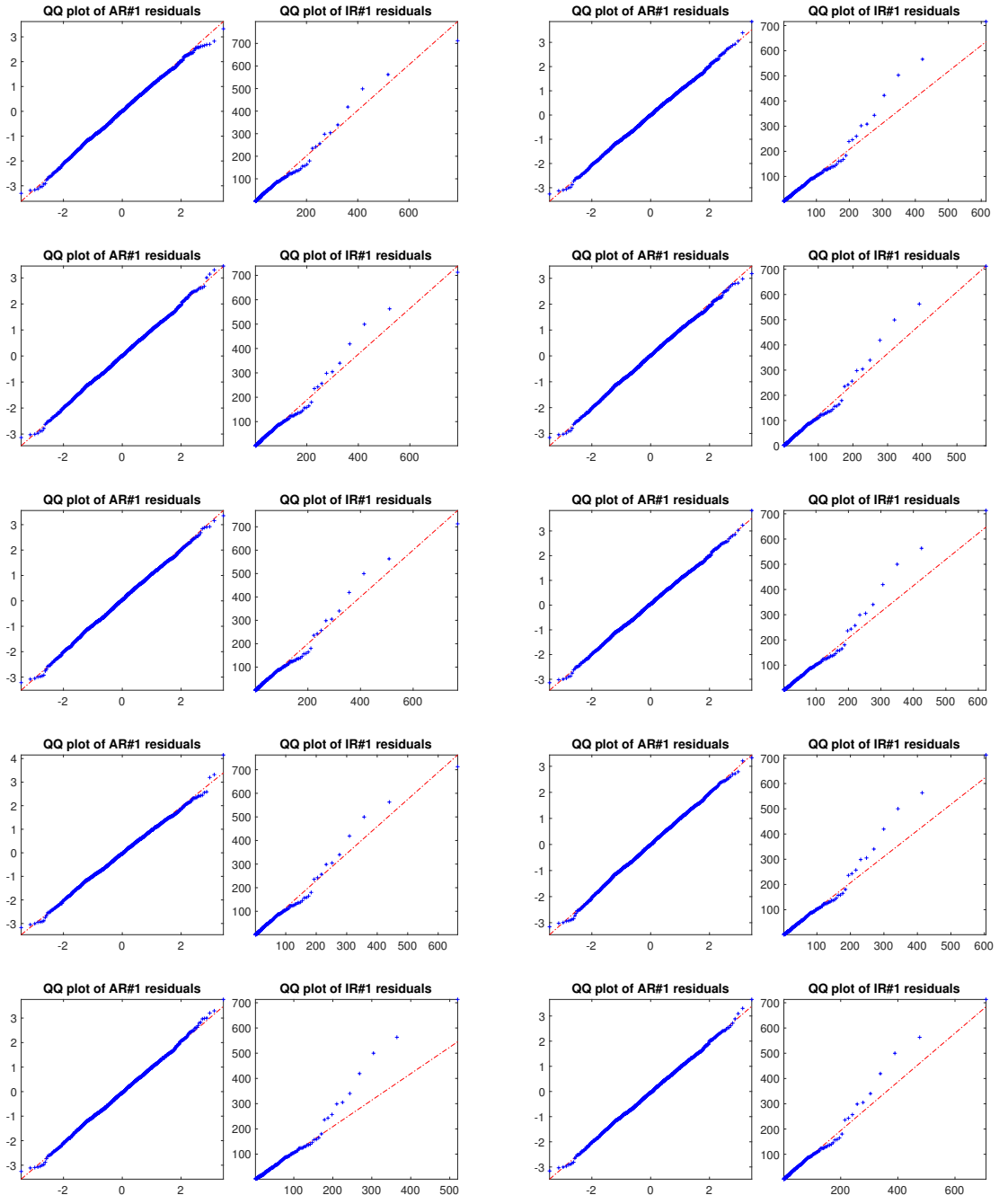


FIGURE 4.7: Ten pairs of QQ plots generated as part of our PPCs for a *single* simulation of the model in Equation (4.7), when the correct model is fitted to the data. Each pair of plots was generated from an independent draw of $\boldsymbol{\theta}$ and $\boldsymbol{R}$ from the posterior, but from the same dataset. These PPCs are used to assess within-regime distributional assumptions. In each pair, the plot on the left is the QQ plot of the residuals of the AR(1) regime, and the plot on the right is for the shifted-log-normal regime.

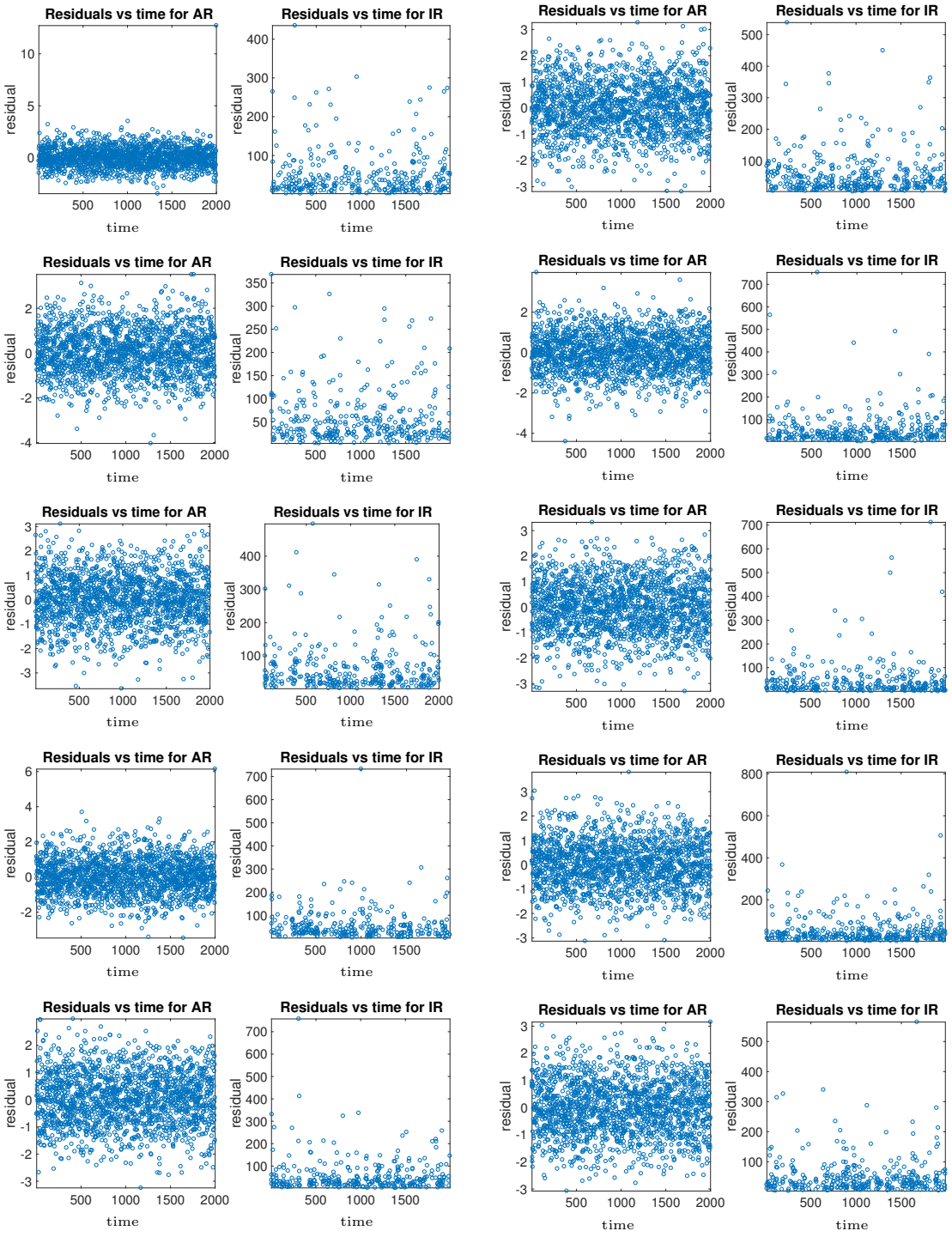


FIGURE 4.8: Ten pairs of residual-versus-time plots generated as part of our PPCs for simulations of the model in Equation (4.7), when the correct model is fitted to the data. Each pair of plots was generated by an independently simulated dataset. These PPCs are used to assess within-regime time-homoscedasticity assumptions. In each pair, on the left is the residuals-versus-time plot for the AR(1) regime, and on the right is the residuals-versus-time plot for the shifted-log-normal regime. Since these plots show no obvious pattern, the assumption of homoscedasticity seems reasonable.

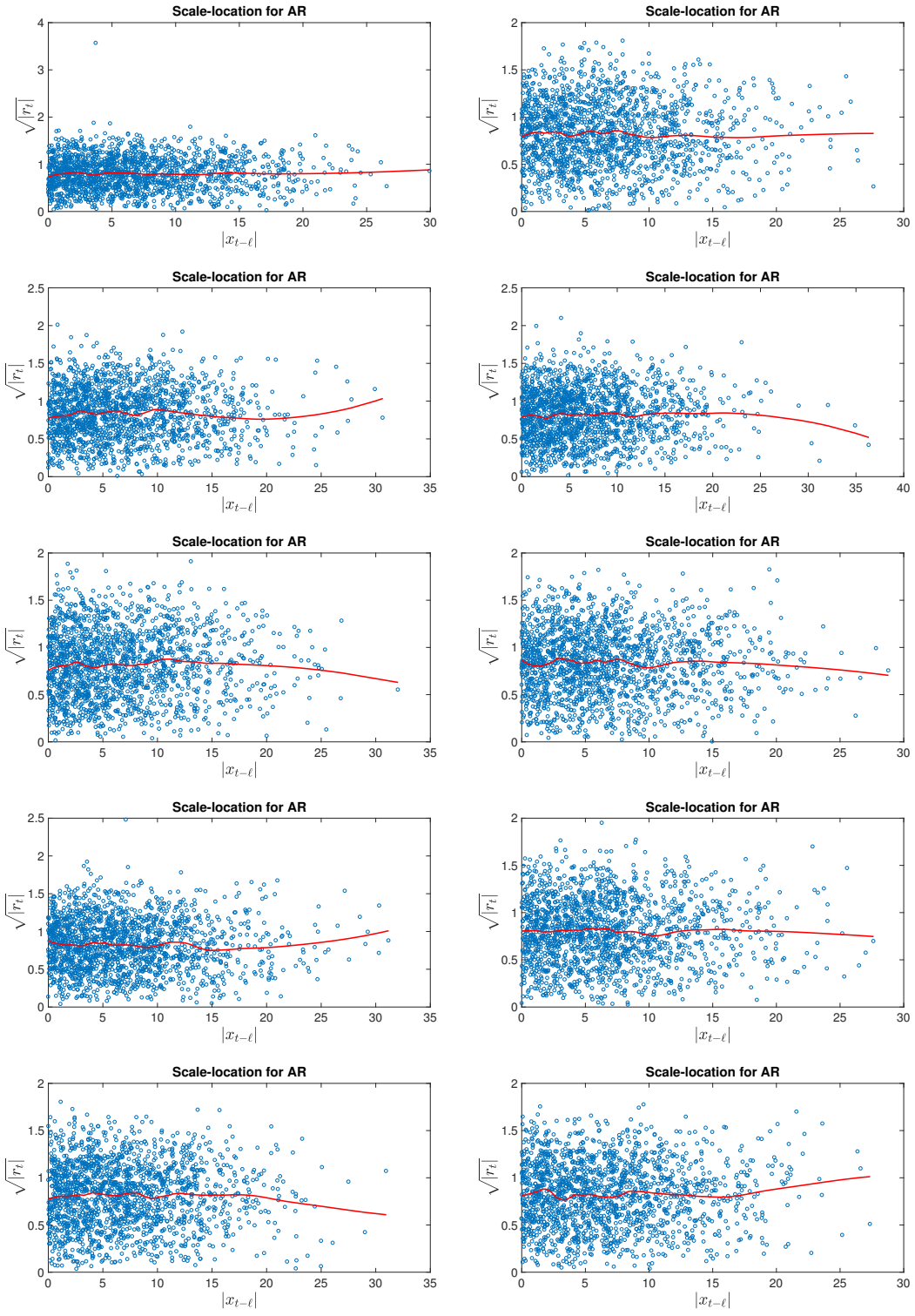


FIGURE 4.9: Ten scale-location plots for AR(1) regime residuals, where $\sqrt{|r_t|}$ is plotted against the absolute value of lagged values, $|x_{t-\ell}|$, generated as part of our PPCs for the simulations of the model in Equation (4.7), when the correct model is fitted to the data. Each plot was generated by an independently simulated dataset. This PPC assesses whether variance depends on the last observed value from each regime. A smoothed regression line is also included in these plots to help us spot any trend in the mean of the residuals. Since these plots show no obvious increase/decrease in spread as a function of lagged values or shape in the residuals, the assumption of constant variance with respect to lagged values seems reasonable.

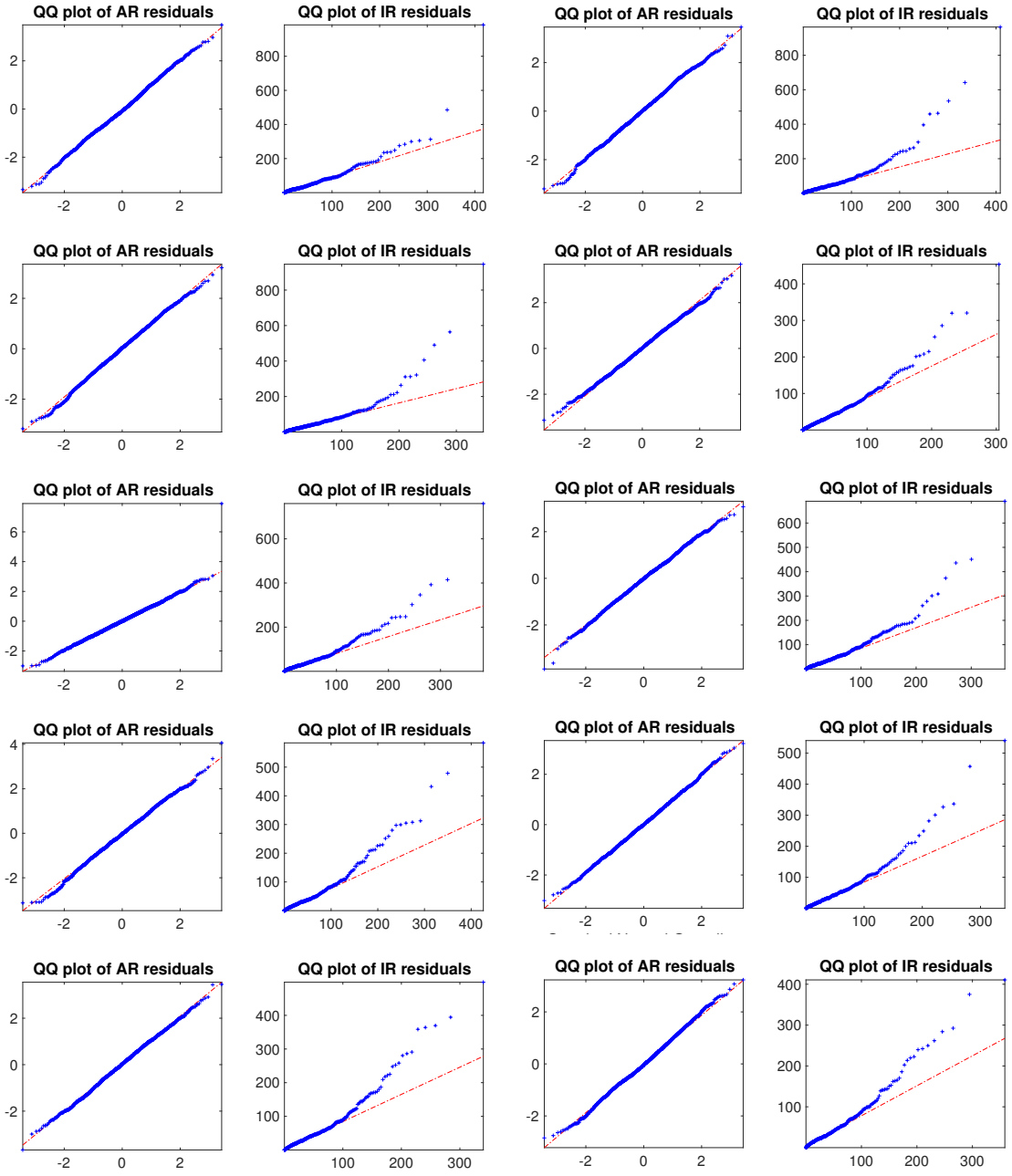


FIGURE 4.10: Ten pairs of QQ plots generated as part of our PPCs for when a model with Gamma distributed spikes is fitted to data generated from the model in Equation (4.7). Each pair of plots was generated by an independently simulated dataset. This PPC assesses within-regime distributional assumptions. In each pair, the plot on the left is the QQ plot of the residuals of the AR(1) regime, and the plot on the right is a QQ plot for the shifted-Gamma regime. Notice, in the QQ plots of the i.i.d. shifted-Gamma regime, the points stray well away from the straight line. In particular, all the points lie above the line. This suggests that the Gamma regime is unsuitable: it does not have enough mass in its tail.

In Figure 4.11 five scale-location PPCs are shown for $\gamma = -0.5$ (left), and $\gamma = -0.25$ (right), and in Figure 4.12 five scale-location PPCs are shown for $\gamma = 0.25$ (left), and $\gamma = 0.5$ (right). Each plot was produced from a different simulated dataset. The scale-location-PPC is used to assess constant variance assumptions: if the assumption of constant variance is correct, there should be no increase or decrease in the variance of the residuals as we move along the x-axis.

In all the plots in Figure 4.11, the spread of the residuals clearly narrows as the magnitude of the lagged value gets larger for both $\gamma = -0.25$ and -0.5, suggesting the constant variance model is inappropriate for both simulated datasets. In addition, notice the scale of the x-axis is much larger in these plots compared to the case when the model fits well (in Figure 4.9), which indicates that some of the extreme values from the shifted-log-normal regime are being captured by the AR(1) process.

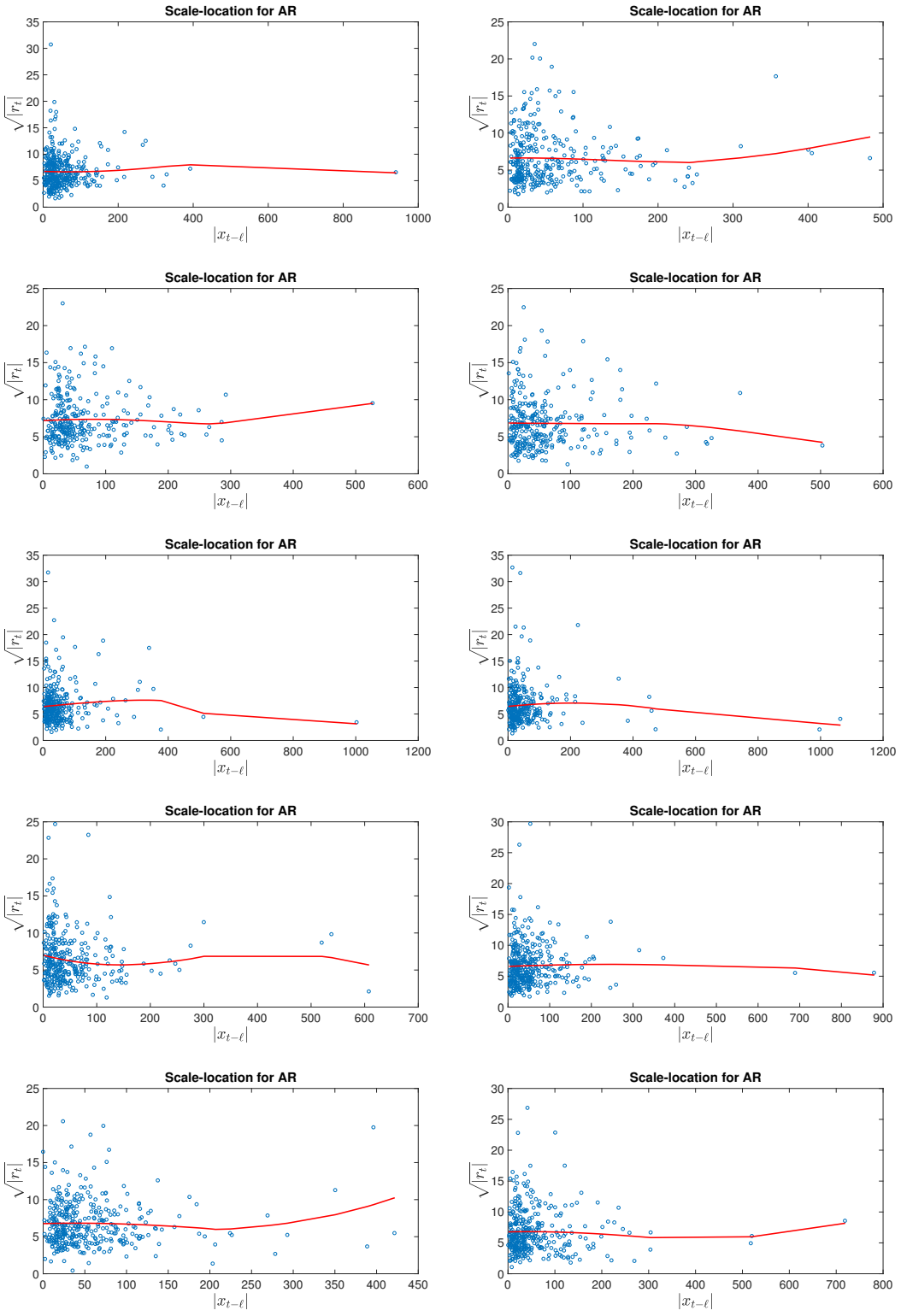
In Figure 4.12, when $\gamma = 0.25$ (left) there are limited significant indications that the variance may be non-constant; compared to Figure 4.9, there are only very subtle differences between the scale-location PPCs apart for the larger lagged values. When $\gamma = 0.5$ (right) there generally appears to be an increase in the residuals as a function of the magnitude of lagged values, $|x_{t-\ell}|$, compared to Figure 4.9 where there is no such trend, particularly for the more reasonable lagged values. This is an interesting observation. When $\gamma > 0$ we expected to see the spread of the residual increase as a function of $|x_{t-\ell}|$, but what we see instead is an increase in the magnitude of the residuals as a function of $|x_{t-\ell}|$, and an upward trend line. We suspect this is due to the shifted-log-normal regime capturing large observations generated by the CEV regime.

These observations indicate that the scale-location-PPC has some power to reject a constant variance model when data follows CEV dynamics, with negative values of γ. When γ is positive, this PPC has less power to differentiate between a constant variance and CEV model, although an upward trend in the residuals as a function of $|x_{t-\ell}|$, can indicate that a model may be inappropriate.

The poor model fit also shows up in our QQ plot-PPCs for the AR(1) regime (Figure 4.13) since the fitting process has had to compromise in the regime allocation process. Notice the residuals of the AR(1) regime deviate from a straight line in the tails. This is perhaps the stronger signal that the model is not appropriate.

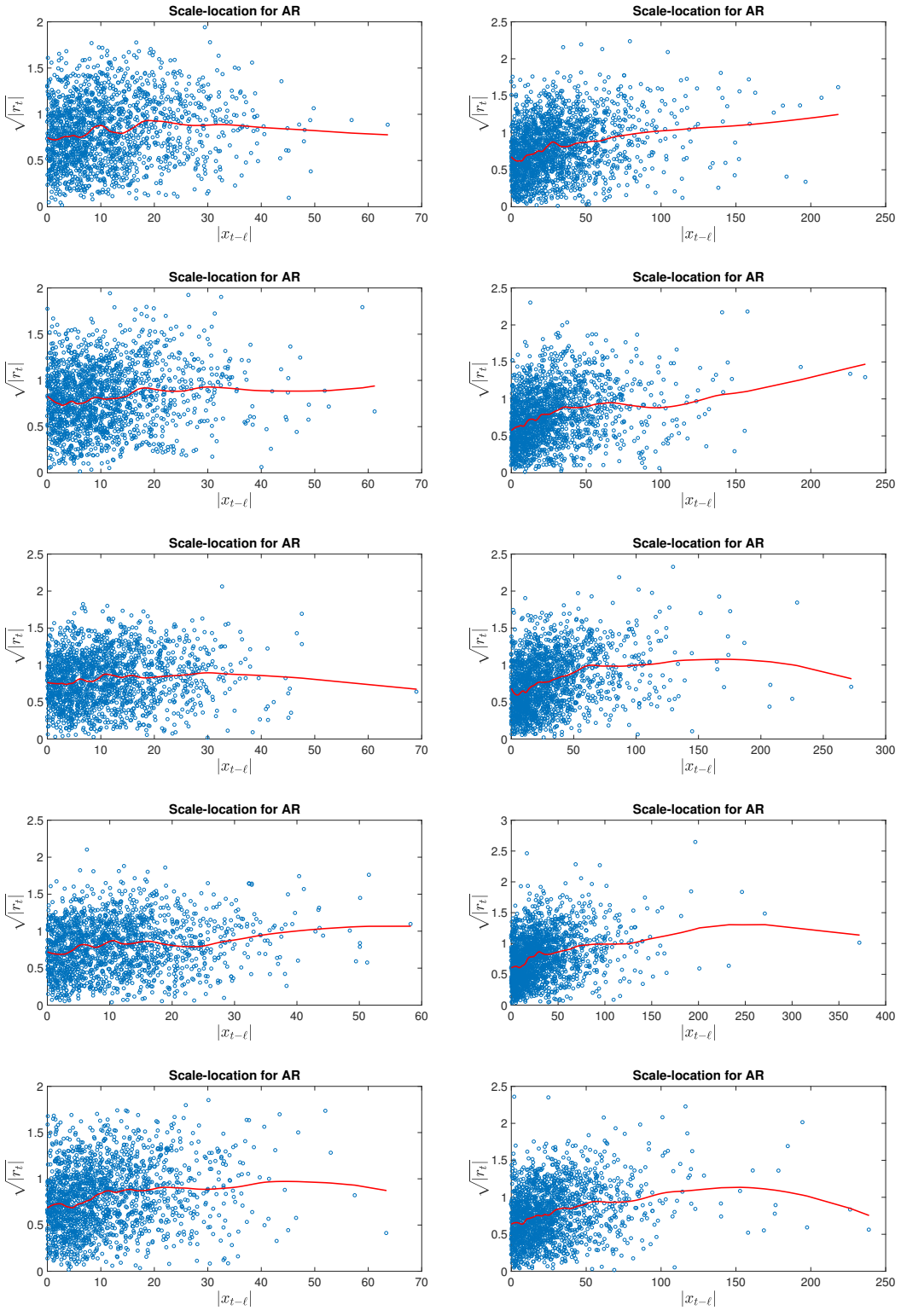
4.4.4 Determining when more regimes are needed

We also want our methods to inform us if more regimes are needed to model the data. First, we investigate the case when there should be more than one i.i.d. regime.



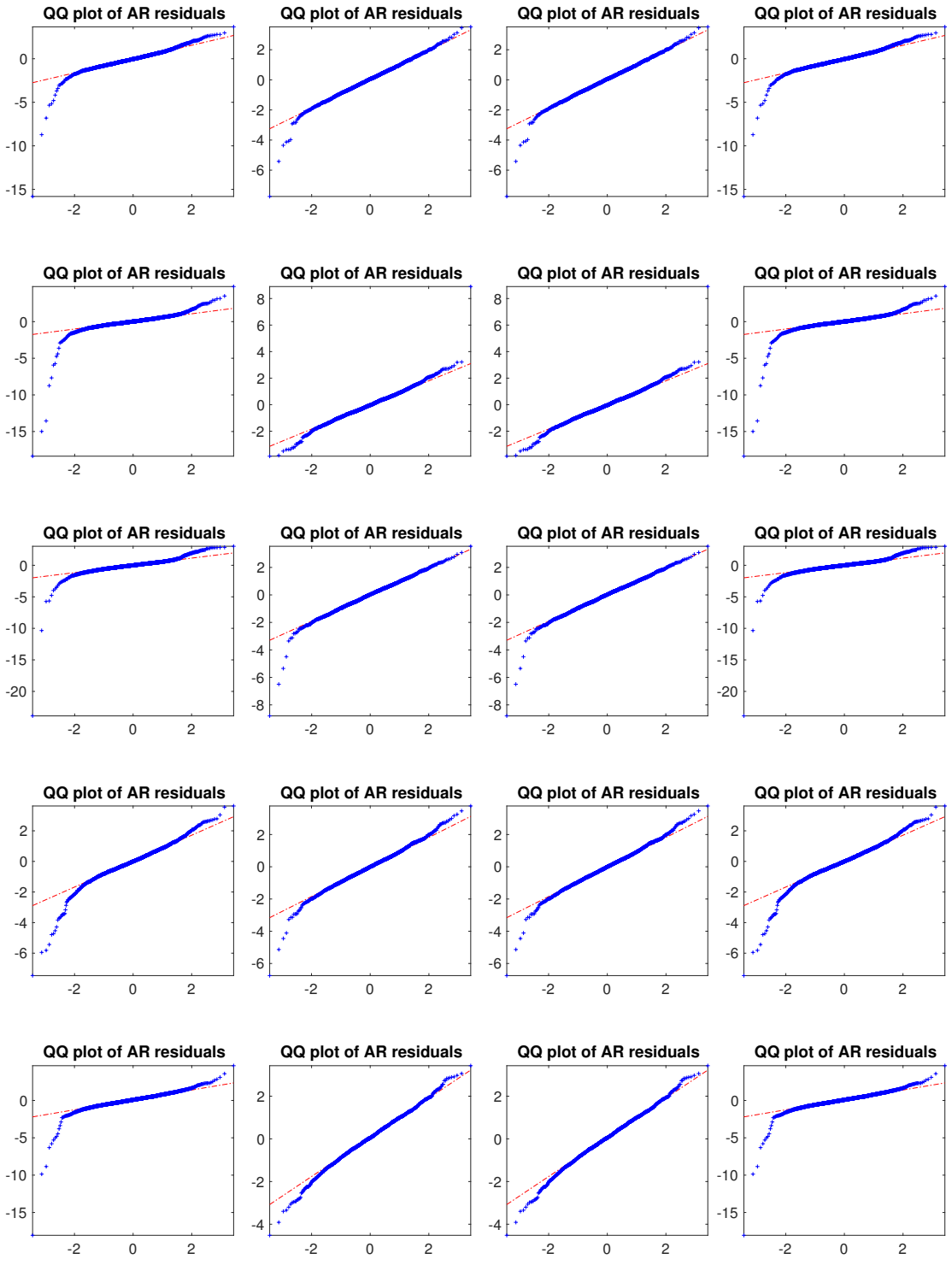
(A) $\gamma = -0.5$ (B) $\gamma = -0.25$

FIGURE 4.11: Scale-location-PPC plots generated as part of our PPCs when fitting the model in Equation (4.7) to simulated data generated from the model in Equation (4.8) for $\gamma = -0.5$ (left) and $\gamma = -0.25$ (right). Each plot is generated by an independent simulation. The red line is a smoothed regression line to help spot possible trends in the points. For both $\gamma = -0.25$ (left) and $\gamma = -0.5$ (right) it is clear the variance decreases as we move from left to right.



(A) $\gamma = 0.25$ (B) $\gamma = 0.5$

FIGURE 4.12: Scale-location-PPC plots generated as part of our PPCs when fitting the model in Equation (4.7) to simulated data generated from the model in Equation (4.8) for $\gamma = 0.25$ (left) and $\gamma = 0.5$ (right). Each plot is generated by an independent simulation. The red line is a smoothed regression line to help spot possible trends in the points. Comparing the plots produced when $\gamma = 0.25$ (left) to the plots in Figure 4.9, (where $\gamma = 0$, and the fitted model is correct) we see only slight differences. When $\gamma = 0.5$ (right) we see the residuals increase as we move from left to right.



(A) $\gamma = -0.5$ (B) $\gamma = -0.25$ (C) $\gamma = 0.25$ (D) $\gamma = 0.5$

FIGURE 4.13: QQ plots generated as part of our PPCs when fitting the model in Equation (4.7) to simulated data generated from the models in Equation (4.8) for $\gamma = -0.5$, (left), $\gamma = -0.25$ (centre-left), $\gamma = 0.25$ (centre-right), and $\gamma = 0.5$ (right). Each plot was generated by an independently simulated dataset. Notice there is some deviation from the reference line in these QQ plots, particularly for $\gamma = \pm 0.5$, suggesting either the AR(1) model in Equation (4.7) may not be appropriate for the data, or that the fitting process has poorly allocated points to this regime.

We simulated 20 independent realisations of length $T = 2000$ from the following MRS model of Type II:

$$X_t = \begin{cases} B_t, & R_t = 1, \\ S_t^{(2)}, & R_t = 2, \\ S_t^{(3)}, & R_t = 3, \end{cases} \tag{4.9}$$

where

$$B_t = 0.55B_{t-1} + \sqrt{53}\varepsilon_t$$

is an AR(1) process, $\{\varepsilon_t\}$ is a sequence of i.i.d.N(0,1) random variables, $\{S_t^{(2)}\}$ is a sequence of shifted-log-normal random variables with $\log(S_t^{(2)} - 12) \sim$ i.i.d. N(3.5,1), $\{S_t^{(3)}\}$ is also a sequence of shifted-log-normal random variables with $\log(S_t^{(3)} - 185) \sim$ i.i.d. N(4.7,0.3), and $\{R_t\}$ is a Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.82 & 0.13 & 0.05 \\ 0.36 & 0.62 & 0.02 \\ 0.38 & 0.13 & 0.49 \end{bmatrix}.$$

Then we used our Bayesian methodology to fit a two-regime MRS model of Type II, with one AR(1) regime and one shifted-log-normal regime. Some QQ plot-PPCs from this are shown in Figure 4.14. Clearly this two-regime model is unable to capture the extreme values as shown by the QQ plots for the shifted-log-normal regime.

We also simulated 20 independent realisations of length $T = 2000$ from the following MRS model of Type II:

$$X_t = \begin{cases} B_t^{(1)}, & R_t = 1, \\ B_t^{(2)}, & R_t = 2, \\ S_t, & R_t = 3, \end{cases} \tag{4.10}$$

where

$$B_t^{(1)} = 0.55B_{t-1}^{(1)} + \sqrt{53}\varepsilon_t^{(1)},$$

and

$$B_t^{(2)} = 0.55B_{t-1}^{(2)} + \sqrt{530}\varepsilon_t^{(2)},$$

are AR(1) processes and $\{\varepsilon_t^{(i)}\}$ $i = 1, 2$, are independent sequences of i.i.d. N(0,1) random variables, $\{S_t\}$ is a sequence of shifted-log-normal random variables, with $\log(S_t - 12) \sim$ i.i.d. N(3.5,1) and $\{R_t\}$ is a Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.82 & 0.13 & 0.05 \\ 0.36 & 0.62 & 0.02 \\ 0.38 & 0.13 & 0.49 \end{bmatrix}.$$

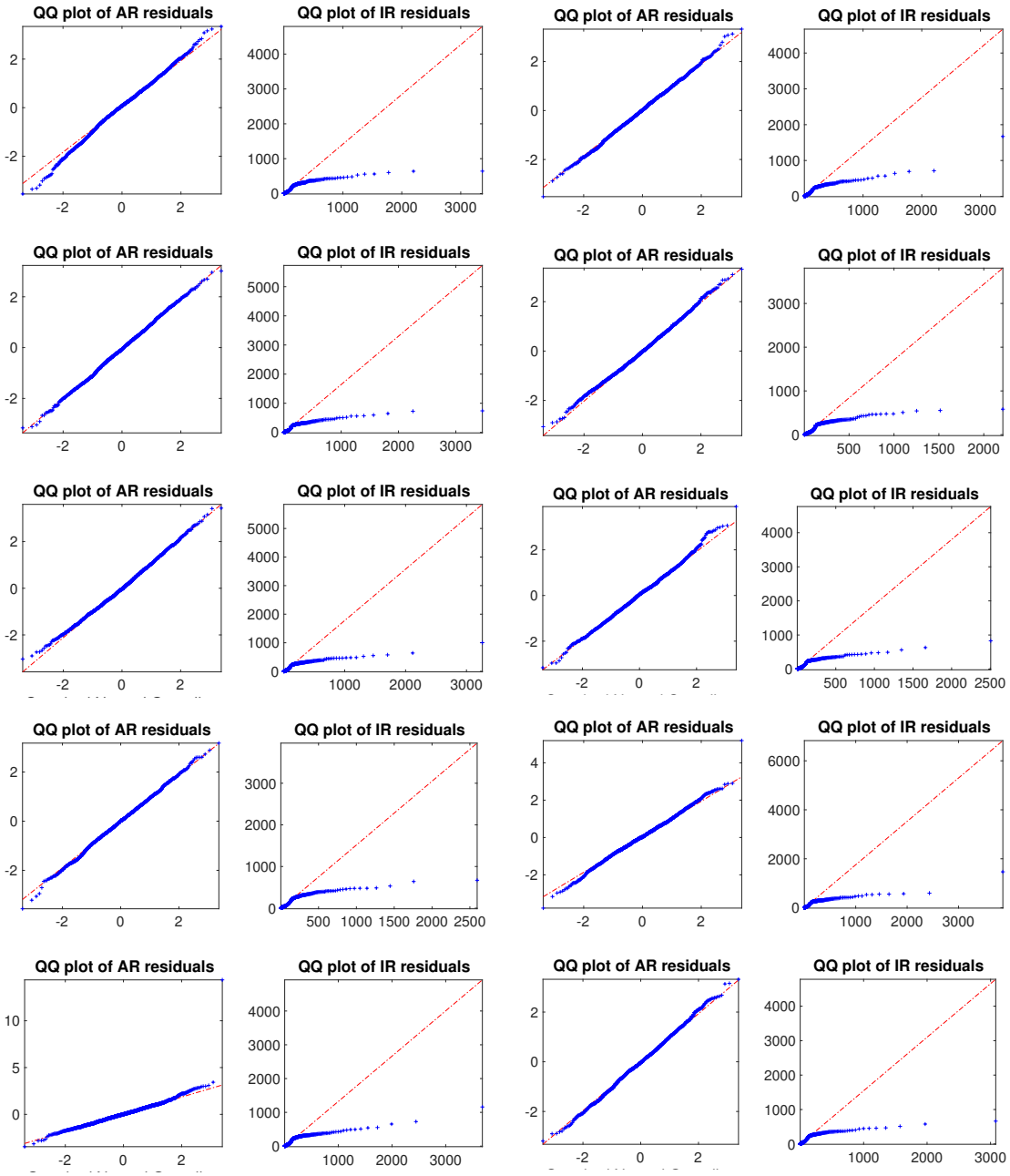


FIGURE 4.14: Ten pairs of QQ plots generated as part of our PPCs for simulations of the model in Equation (4.9), when a two-regime model, with one AR(1) regime and one shifted-log-normal regime, is fitted to data generated by a model with three regimes, one AR(1) regime and two shifted-log-normal regimes. Each pair of plots was generated by an independently simulated dataset. These PPCs are used to assess within-regime distributional assumptions. The plots on the left are QQ plots of the residuals of the AR(1) regime, and the plots on the right are for the shifted-log-normal regime. In the QQ plots for the shifted-log-normal regime, the points clearly deviate very badly from the reference line.

Then we used our Bayesian methodology to fit a two-regime MRS model of Type II, with one AR(1) regime and one shifted-log-normal regime. Some QQ plot-PPCs from this are shown in Figure 4.15. Clearly this two-regime model is unable to capture the the data generated by the model in Equation 4.10, as shown by the QQ plots for the AR(1) regime.

Summary

In this chapter we introduce our Bayesian framework for model estimation and selection/checking. For all models under consideration, we specify uniform prior distributions as a form of objective prior. To sample from posterior distributions of MRS models, we develop a data-augmented MCMC algorithm. Data-augmentation is advantageous as it permits an $\mathcal{O}(T)$ implementation of our MCMC algorithm. Our algorithm is also flexible: it can handle many types of model specifications without the need for manual tuning, thanks to the adaptive procedure that we implement.

For model checking in our Bayesian framework, we utilise PPCs, which we implement in a similar manner to residual diagnostics in a traditional ordinary regression setting. To validate our methods, we simulate MRS models, then use our data-augmented MCMC algorithm to sample from the posterior distributions given by the simulated datasets, and create PPCs. Our simulations confirm that our PPCs are able to distinguish between different models, and have some power to tell us when models are incorrect, although, of course, they cannot tell us if our model is correct.

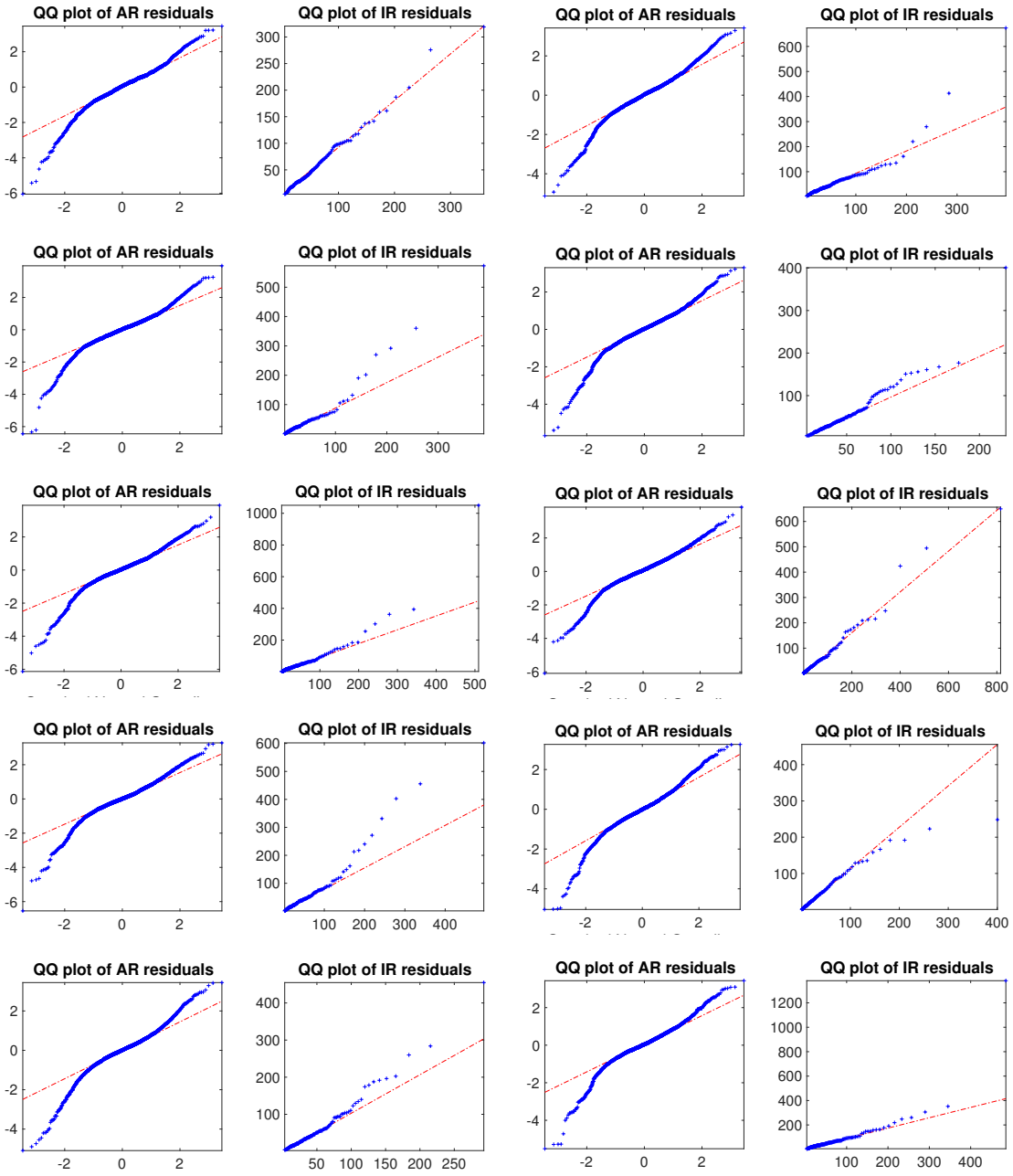


FIGURE 4.15: Ten pairs of QQ plots generated as part of our PPCs for simulations of the model in Equation (4.10), when a two-regime model, with one AR(1) regime and one shifted-log-normal regime, is fitted to data generated by a model with three regimes, two AR(1) regimes and one shifted-log-normal regime. Each pair of plots was generated by an independently simulated dataset. These PPCs are used to assess within-regime distributional assumptions. The plots on the left are QQ plots of the residuals of the AR(1) regime, and the plots on the right are for the shifted-log-normal regime. In the QQ plots of the AR(1) regime, the points clearly deviate significantly from the reference line.

Chapter 5

Applications to South Australian electricity prices

In this chapter we apply our likelihood and Bayesian inference methods to estimate the parameters of, and assess goodness-of-fit for, MRS models for the SA electricity market. Our MRS models for electricity prices are built out of two pieces: a deterministic trend component, T_t, and a stochastic component, X_t, and we model prices as the sum $P_t = T_t + X_t$. In Section 5.1 we present a novel technique to estimate the trend component of MRS models for electricity prices. There are an unlimited number of models we could consider for electricity prices, so in Section 5.2 we narrow the search and describe candidate models. In Sections 5.3 and 5.4 we discuss applications of our Bayesian and likelihood methods, respectively, and make concluding remarks in Section 5.5.

The dataset The South Australian electricity market is a particularly interesting case study due to a number of factors including its relative isolation, occasional extremely hot weather, and generation mix – in 2016 SA had 39.2% of its total generation come from wind farms, 50.5% from gas and 9.2% from residential solar panels [7]. All these factors can contribute to a high and volatile electricity price. In this dataset, on the 1$^{\text{st}}$ of December 2016, we observed a spot price of \$13,767, compare this to the average price for 2016 of \$80.59 per megawatt hour. Our dataset consists of 81,792 half-hourly spot prices from the South Australian electricity market (available at the AEMO website [8]) for the period 00:00 hours, 1$^{\text{st}}$ of January 2013 to 23:30 hours 31$^{\text{st}}$ of September 2017.

Note that this dataset contains a period of 14 days over which the market was suspended. The suspension was due to a market-wide blackout which occurred at 4:20pm on September 28, 2016. Although the majority of the state had power restored by

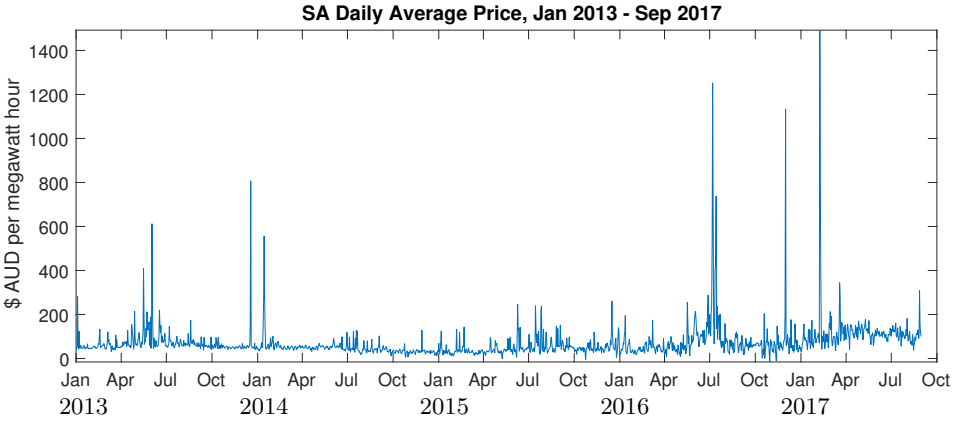


FIGURE 5.1: The daily average wholesale electricity spot price for South Australia for the period from the 1st of January 2013 to the 31st of September 2017, quoted in $AUD per megawatt hour.

the night of September 28, the market operator, AEMO, suspended the market from 4:00pm, on the 28th of September until 10:30pm on the 11th of October. During this period prices were set by AEMO. Prices for this period were calculated as the average price in SA in the 'same trading interval' over the last four weeks. For this calculation the 'same trading interval' means different things for weekdays and weekends. For a given 30-minute trading interval on a weekday, the price was calculated as the average price at the same time only on weekdays over the last four weeks. For a given 30-minute trading interval on a weekend, the price was calculated as the average price at the same time on weekend-days only. During the market suspension, market participants continued to submit price bids in the usual way, and AEMO used these to dispatch generators in an economic merit order, but the bids did not affect prices. In our modelling we do not take this market suspension into account, and model the data 'as is'.

We follow a common practice in the literature and model daily average prices, since the daily average price is sometimes used in derivative valuation. Thus we have a dataset of 1,704 daily average price observations to which we fit our model. The data that we model is plotted in Figure 5.1. Notice that there appears to be an increase in price volatility since about April 2016, which roughly corresponds to the closure of SA's last coal generator.

5.1 Estimation of the trend component

The trend model Electricity spot prices exhibit seasonality on daily, weekly, and longer scales. To capture this multi-scale seasonality, the trend component consists of two parts: a short-term component, g_t, and a long-term component, h_t. We model the

long-term component, h_t, using wavelet filtering since it has been shown to perform well for this application [57]. Among the many available wavelet families, we use D24 wavelets and filter out the long-term seasonality by applying the wavelet filter recursively six times [57]. We use the short-term component, g_t, to capture the mean price for different days of the week and indicator functions to model this:

$$g_t = \beta_{\text{Mon}}\mathbb{I}(t \in \text{Mon}) + \beta_{\text{Tue}}\mathbb{I}(t \in \text{Tue}) + ... + \beta_{\text{Sun}}\mathbb{I}(t \in \text{Sun}),$$

where β_{Mon}, β_{Tue}, ... , β_{Sun}, are the mean deviations from the long-term trend price on Monday, Tuesday, ... , Sunday, respectively. This model is very common for our application.

Estimation of the trend component Extreme prices in electricity markets can bias estimates of trend components [57]. A solution proposed by Janczura and Weron [57] is to first identify, remove and replace extreme prices with more reasonable values, and then estimate the trend component on this altered dataset, before ultimately estimating the stochastic model. We take this one step further, and iterate between identifying extreme prices, replacing them, and estimating the trend component.

The spike identification method that we choose uses the MRS model, which is one of the methods suggested in [57]. Janczura and Weron [57] conclude this classification technique can work well when the goal is to estimate parameters of an MRS model. We define extreme observations as observations that were not generated by an AR(1) (base) regime. An MRS model can be used to identify extreme observations using the posterior probabilities $\mathbb{P}(R_t = i|\boldsymbol{x}_{0:T})$, produced as a byproduct of the fitting process (in both the Bayesian and EM methods). We obtain a hard classification of prices as extreme if $\sum_{i\in\mathcal{S}_{AR}} \mathbb{P}(R_t = i|\boldsymbol{x}_{0:T}) < 0.5$, where $\mathcal{S}_{AR}$ is the set of regimes corresponding to AR(1) processes.

After classifying observations as extreme, they are removed and replaced by 'more reasonable' values. Some different options for these 'more reasonable' values are explored in [96], but they do not come to a conclusion about what the best option is. We replace extreme values with the value of the trend component at the last iteration of our estimation procedure.

To summarise, our trend estimation procedure is as follows:

Step 0. Estimate the trend components from the raw data.

Step 1. Remove the trend component from the raw data, and then estimate the stochastic component and classify observations into regimes.

Step 2. Replace prices not classified as base prices by their trend values from the last iteration of this process.

Step 3. Re-estimate the trend components.

Step 4. Iterate Steps 1-3.

In practice, we have found that four or five iterations are usually sufficient for satisfactory results for our purposes, since the difference between successive estimates of the trend is small compared to the magnitude of prices.

To estimate the parameters of the trend component, we first estimate h_t using wavelet filtering (see Section 2.2.7) and remove this from the current representation of the prices. The short-term component is then estimated using averaging:

$$\widehat{g}_t = \widehat{\beta}_{\text{Mon}}\mathbb{I}(t \in \text{Mon}) + \widehat{\beta}_{\text{Tue}}\mathbb{I}(t \in \text{Tue}) + ... + \widehat{\beta}_{\text{Sun}}\mathbb{I}(t \in \text{Sun}),$$

where

$$\widehat{\beta}_d = \frac{\sum_{t=1}^{T}(\widehat{P}_t - h_t)\mathbb{I}(t \in d)}{\sum_{t=1}^{T}\mathbb{I}(t \in d)},$$

for $d =$ Mon, Tue,..., Sun, where $\widehat{P}_t$, $t = 0, 1, ..., T$, is the current representation of prices without spikes.

5.2 Models under consideration

To simplify our exploration, we restrict attention to a specific subset of candidate models. We consider models with up to five regimes, with either one or two AR(1) regimes, and either one, two or three independent and identically distributed (i.i.d.) regimes. So the biggest model we consider is the five-regime model

$$X_t = \begin{cases} B_t^{(1)} & \text{if } R_t = 1, \\ B_t^{(2)} & \text{if } R_t = 2, \\ Y_t^{(3)} & \text{if } R_t = 3, \\ Y_t^{(4)} & \text{if } R_t = 4, \\ Y_t^{(5)} & \text{if } R_t = 5, \end{cases}$$

where $\left\{B_t^{(i)}\right\}$ for $i = 1, 2$ are AR(1) processes, and $\left\{Y_t^{(j)}\right\}$ for $j = 3, 4, 5$ are i.i.d. processes. We only ever specify AR(1) processes of the form

$$B_t^{(i)} = \alpha_i + \phi_i B_{t-1}^{(i)} + \sigma_i \varepsilon_t^{(i)},$$

where $\left\{\varepsilon_t^{(i)}\right\}$ is a sequence of i.i.d. N(0,1) random variables. We label AR(1) regimes as *base* regimes, since they are included in the model to capture prices under normal operating conditions.

The i.i.d. components either capture price spikes, or price drops, depending on their specification. Following [58] we specify shifted i.i.d. distributions (with shifting parameter q), as these can more accurately separate spikes and drops from base prices. We explore the following distributions for the i.i.d. processes, to attempt to find the distribution that fits the data best:

- $Y_t^{(j)} - q_j \sim$ Gamma(μ_j, σ_j^2), to capture spikes.
- $Y_t^{(j)} - q_j \sim$ Log-normal(μ_j, σ_j^2), to capture spikes.
- $q_j - Y_t^{(j)} \sim$ Log-normal(μ_j, σ_j^2), to capture drops.

We only ever specify models with one drop regime, and allow up to two spike regimes.

In our Bayesian model estimation, we leave the shifting parameters as parameters to be inferred by the model but on a restricted domain, since, leaving them completely unrestricted can lead to erroneous results. When left unrestricted, the shifting parameter for the spike distribution (drop distributions) becomes negative (positive), and rather than capturing extreme events, the spike (drop) regime captures periods of high volatility in the base regime. For this reason we restrict the support of the posterior of q, using the prior distribution. For the drop regime the support of the posterior for q is below the $\frac{1}{3}$-quantile of the detrended data, and for the spike regimes the support of the posterior for q is above the $\frac{2}{3}$-quantile for the first spike regime, or above the 98^{th} percentile for the second spike regime.

As discussed in Chapter 3 there are issues in estimating the shifting parameter q when using maximum likelihood. When fitting shifted log-normal distributions using maximum likelihood, we fix the value of q based on our Bayesian analysis. We leave the shifting parameter q for the shifted-Gamma distribution to be estimated using maximum likelihood, but recall from Section 3.5.3 that we restrict the shape parameter, μ, to be greater than 2.5 so that estimation of q is more stable. We also restrict these shifting parameters, exactly as in the Bayesian analysis.

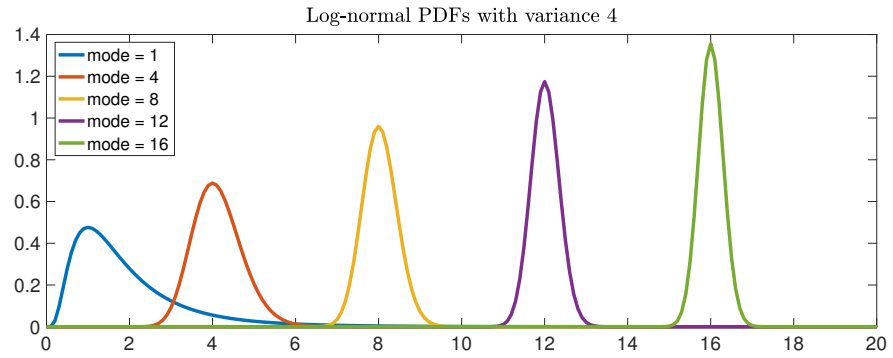
We restrict the shape parameter, μ, of the shifted-Gamma distribution in our Bayesian analysis to be greater than 2.5 also. Restricting the shape parameter in this way in the Bayesian case is a modelling choice. With a shape parameter $\mu > 1$ the Gamma density function is continuous at 0, whereas when $\mu \leq 1$ it is not, and the mode of the distribution is at 0. The mode of the shifted-Gamma distribution is $q + (\mu - 1)\sigma^2$, therefore, when we specify $\mu > 2.5$, we are requiring the majority of the mass in the Gamma distribution to be away from the boundary at q. We believe this makes for a more sensible model since there should not be a mass of spike-regime points at the boundary q. Rather, the majority of spikes should be above q, but there should be the possibility of having low spikes (near q), where prices are not high relative to the whole dataset, but are higher than, and/or not highly correlated with, surrounding prices so they are not well-modelled by the base process(es).

The reader may have noticed that, in the previous paragraph, we imply that the location of the mode of the shifted-Gamma distribution can be shifted via either the parameter q or μ and σ^2. Similarly, for the shifted-log-normal distribution the location of the mode of the distribution can be shifted by the parameters μ and σ^2 or by the shifting parameter q (the mode is given by $q + \exp(\mu - \sigma^2)$). We want to make clear that these are not equivalent. The difference between shifting the mass of these distributions using the parameters μ and σ^2 rather than q is that when the mode of these distributions gets further from q, they must become more and more symmetric. Figure 5.2 shows this behaviour. In Figure 5.2 five log-normal and five Gamma density functions with different modes but constant variance (a variance of 4 and $q = 0$) are plotted. Notice that as the locations of the modes increase the distributions become more symmetric. Thus, shifting the mass of these distributions via q is different from shifting the mass of these distributions via μ or σ^2.

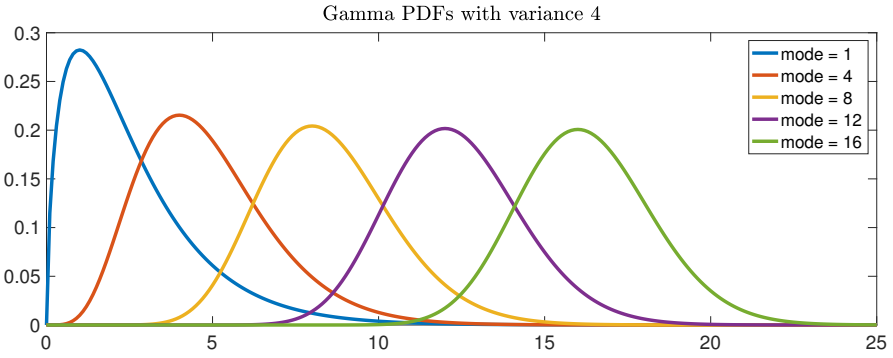
When two AR(1) regimes are included in the model we specify $\sigma_1^2 < \sigma_2^2$, so the second AR(1) regime has a higher variance than the first. This is to stop the *aliasing* of the regimes. We do not consider models with an AR(1) base regime with CEV dynamics, i.e. of the form

$$B_t = \alpha + \phi B_{t-1} + \sigma |B_{t-1}|^{\gamma} \varepsilon_t,$$

and leave this as an area for future research.



(A)



(B)

FIGURE 5.2: (a) Log-normal density functions with different modes, all with variance 4 and $q = 0$. (b) Gamma density functions with different modes, all with variance 4 and $q = 0$. Notice that as the location of the modes of these distributions increases they become more symmetric.

5.3 Bayesian estimation and selection

Here we apply our Bayesian methodology to the South Australian electricity prices. We first consider MRS models of Type II, followed by models of Type III. There are many similarities between the two analyses, and so some of the details for Type III models are shown in Appendix B.1.

Type II MRS models

In Table 5.1 we summarise our Bayesian model selection for MRS models of Type II.

	Model 1	Model 2	Model 4
	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \\ Y_t^{(4)}, & \text{if } R_t = 4, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{Gamma}(\mu_3, \sigma_3^2)$, $Y_t^{(4)} - q_4 \sim \text{Gamma}(\mu_4, \sigma_4^2)$.
QQ plots	Figure 5.3: distributional assumptions violated for Regimes 1 and 3.	Figure 5.6: distributional assumptions are suitable for Regimes 1 and 2, but questionable for Regime 3.	Figure 5.8: distributional assumptions are suitable for all regimes.
Residuals vs time	Figure 5.4: variance is non-constant over time.	Figure 5.9: there are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance can be attributed to fewer observations in Regime 2 at earlier times. We conclude that the time-homoscedasticity assumptions are reasonable.	Figure 5.9: there are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance can be attributed to fewer observations in Regime 2 at earlier times. We conclude that the time-homoscedasticity assumptions are reasonable.
Scale-location	Not shown.	Figure 5.10: self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.	Figure 5.11 self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.
Comments	Not a good model.	Regime 2 is included to capture time-heteroscedasticity. QQ plots in Figure 5.6 suggest shifted-log-normal spikes are more suitable than shifted-Gamma spikes (not shown). When two AR(1) regimes are included in a model, no drop regime is necessary.	Regime 4, a second spike regime, is included to capture the very largest spikes. QQ plots in Figures 5.7 and 5.8 show slight differences between sifted-log-normal spikes (Model 3) and shifted-Gamma spikes, but suggest the latter are more suitable.

TABLE 5.1: A summary of our Bayesian model selection process for Type II MRS models for the SA dataset. We also considered a range of other models, such as the models in this table with an added drop regime or alternative spike distribution specifications, but, compared to these models, they either did not fit well, or the drop regime was not necessary, and so discussing them in this thesis is not necessary.

We first considered the following two-regime model.

Model 1 of Type II

$$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases} \qquad \text{(Model 1 of Type II)}$$

where $Y_t^{(3)} - q_3$ follows a log-normal distribution.

A sample of five QQ plot PPCs for each regime in Model 1 of Type II are shown in Figure 5.3 where it is clear that distributional assumptions for both regimes are inappropriate. A sample of five residuals versus time PPCs for each regime in Model 1 of Type II are show in Figure 5.4, where it is obvious that the the residuals of the AR(1) regime are not constant over time. We also fitted Model 1 of Type II with shifted-Gamma spikes instead, but observed even worse violations of distributional assumptions in our QQ plots (results not shown).

The time heteroscedasticity observed for Model 1 of Type II suggests that we might need two AR(1) base regimes.

Model 2 of Type II with two base regimes $B_t^{(i)}$ $i = 1, 2$, and one spike regime, $Y_t^{(3)}$:

$$X_t = \begin{cases} B_t^{(1)} & \text{if } R_t = 1, \\ B_t^{(2)} & \text{if } R_t = 2, \\ Y_t^{(3)} & \text{if } R_t = 3, \end{cases} \qquad \text{(Model 2 of Type II)}$$

where $Y_t^{(3)} - q_3$ follows a log-normal distribution.

This model was fitted to the data and PPCs produced. To visualise which prices each regime in this model are capturing, we classify prices using the posterior probabilities $\mathbb{P}(R_t = i|\boldsymbol{x}_{0:T})$. We highlight a data point in red if $\mathbb{P}(R_t = i|\boldsymbol{x}_{0:T}) > 0.5$; this is shown in Figure 5.5. Here we see the second base regime, $B_t^{(2)}$, capture a significant jump in volatility around April 2016, which roughly coincides with the closure of South Australia's last coal generation facility, and therefore a change in market structure.

We also fitted Model 2 of Type II with a drop regime (results not shown), however, we found this was not needed and all drops are preferably modelled by the AR(1) processes, as evidenced by the fact that our Bayesian methodology assigned no mass to the drop regime at all. Moreover, for *any* model with two base regimes and at least one spike

regime, we found any drops in prices are best modelled by the AR(1) regimes, and not a drop regime.

A sample of five QQ plot PPCs for the residuals of each regime of Model 2 of Type II are shown in Figure 5.6. These QQ plot PPCs suggest the assumptions for the AR(1) regimes are reasonable. However, there is some evidence that the single shifted-log-normal spike regime is unable to capture extreme spikes. We also fitted a three-regime model like Model 2 of Type II, except with shifted-Gamma spikes, and the QQ plot PPCs for this model suggested that shifted-log-normal spikes are more appropriate (results not shown). The addition of the second AR(1) regime also removed the time-heteroscedasticity, as discussed below.

To investigate if a second spike regime is needed to capture the largest spikes, we fit the following models.

Model 3 of Type II which has two base regimes $B_t^{(i)}$, $i = 1, 2$, two spike regimes, one for 'typical' spikes, $Y_t^{(3)}$, and another for extreme spikes, $Y_t^{(4)}$, and no drop regime:

$$X_t = \begin{cases} B_t^{(1)} & \text{if } R_t = 1, \\ B_t^{(2)} & \text{if } R_t = 2, \\ Y_t^{(3)} & \text{if } R_t = 3, \\ Y_t^{(4)} & \text{if } R_t = 4, \end{cases} \qquad \text{(Model 3 of Type II)}$$

where $Y_t^{(3)} - q_3$ and $Y_t^{(4)} - q_4$ follow log-normal distributions.

Model 4 of Type II which is the same as Model 3 of Type II except the spike distributions follow shifted-Gamma distributions.

A sample of QQ plot PPCs for the spike regimes in Models 3 and 4 of Type II are shown in Figures 5.7 and 5.8, respectively. Observing Figures 5.7 and 5.8 we see the performance of Models 3 and 4, as measured by these PPCs, is similar, however, it appears as if Model 4 captures extreme observations more accurately.

Our QQ plot PPCs suggest both Models 2 and 4 of Type II are reasonable, and now we investigate other assumptions of these models using our other PPCs. In Figure 5.9 a sample of five residuals versus time PPC plots are shown for each AR(1) regime in Models 2 and 4 of Type II. For Regime 1, since there is no obvious fanning of the residuals as a function of time, or shape in the residuals, we conclude that the assumption of time-homoscedasticity is reasonable for both models. For Regime 2 there is a small amount of evidence that the variance of residuals increases over time for both models,

but this evidence is not strong. Furthermore, some of the apparent change in variation can be attributed to fewer observations in Regime 2 at earlier times, rather than time-heteroscedasticity, and we conclude that the assumption of time-homoscedasticity is reasonable for Regime 2 for both models. In Figures 5.10 and 5.11 a sample of five scale-location PPCs are shown for each AR(1) regime in Models 2 and 4 of Type II, respectively. There is little evidence in Figures 5.10 and 5.11 that self-dependent-homoscedasticity assumptions are violated for either regime in either model since there is no obvious increase or decrease in the magnitude or variance of residuals as a function of lagged values, $|x_{t-\ell}|$.

To summarise, the QQ plots for Model 4 of Type II are a slight improvement on the QQ plots for Model 2 of Type II; the residuals versus time PPCs show no serious violation of the time-homoscedasticity assumptions for either Model 2 or 4 of Type II; and the scale-location PPCs suggests self-dependent-homoscedasticity assumptions are not violated for either model. From these PPCs alone we could conclude Model 4 of Type II is best. However, we should note that Model 4 of Type II has nine more parameters than Model 2 of Type II and may be subject to overfitting. More work is needed here such as out-of-sample model assessment.

The posterior means for the parameters of Models 2 and 4 of Type II are shown in Table 5.2, while some of the non-trivial correlation structures in the posterior distributions are shown in the scatter-plots of Figures 5.13-5.17. Surprisingly, there is no obvious correlation structure between the parameters α_i, ϕ_i and σ_i^2 within the AR(1) regimes. The estimated trend components for Models 2 and 4 of Type II are shown in Figure 5.12.

Type II MRS models

See Appendix B.1.

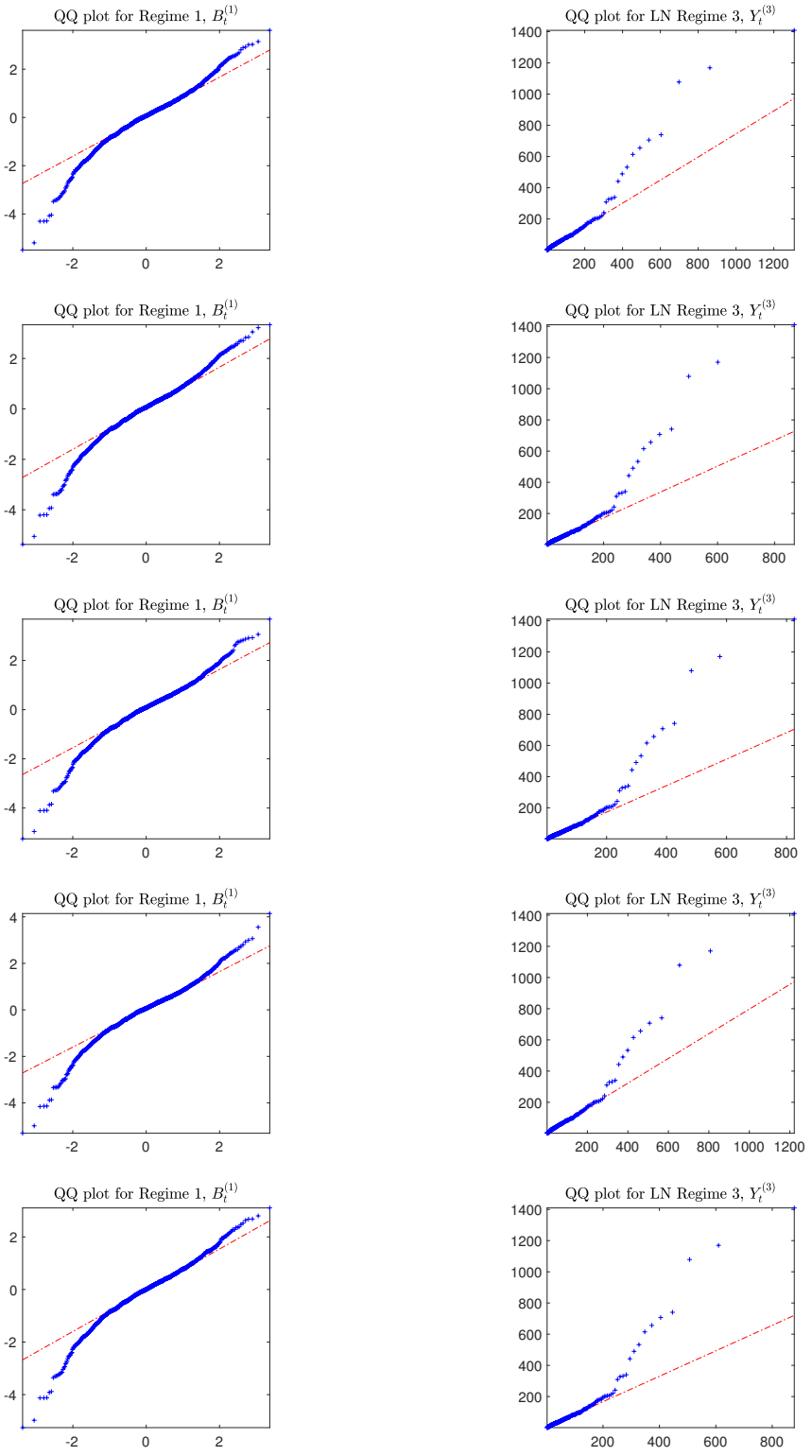


FIGURE 5.3: A sample of five QQ plot PPCs for each regime in Model 1 of Type II. (Left) QQ plot PPCs for Regime 1, the AR(1) regime. (Right) QQ plot PPCs for Regime 3, the shifted-log-normal spike regime. The points in the QQ plots for both regimes clearly do not lie on a straight line, suggesting the model does not capture the data well.

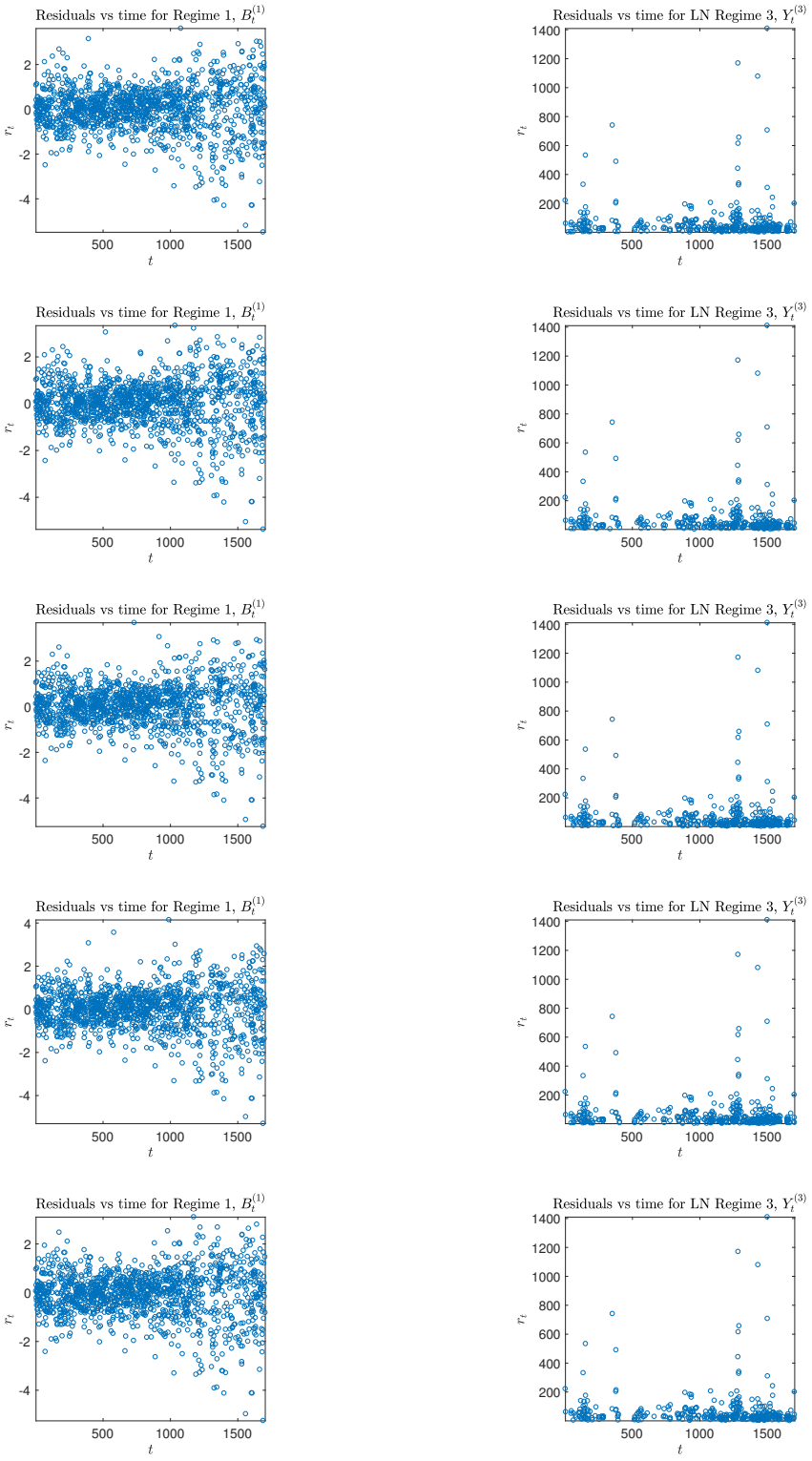
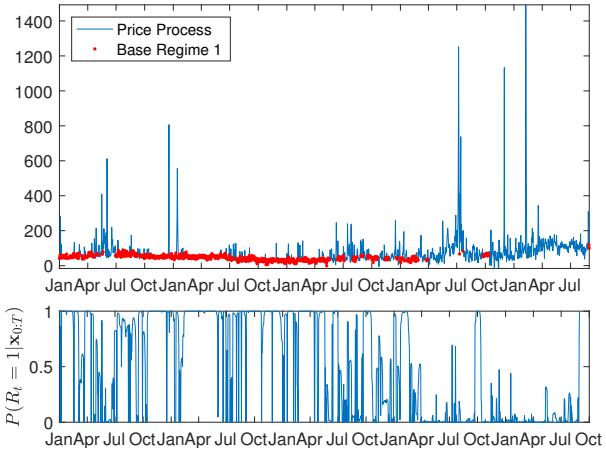
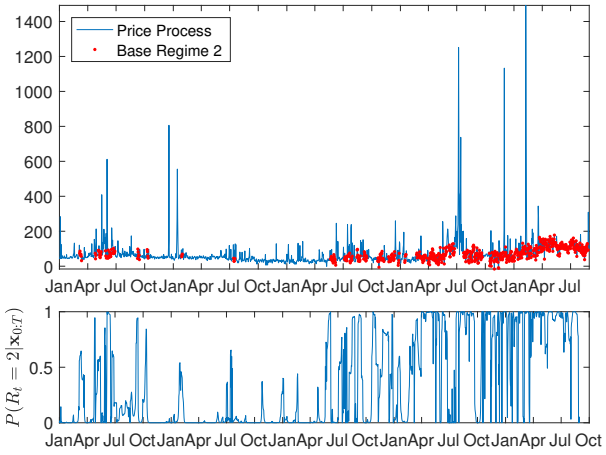


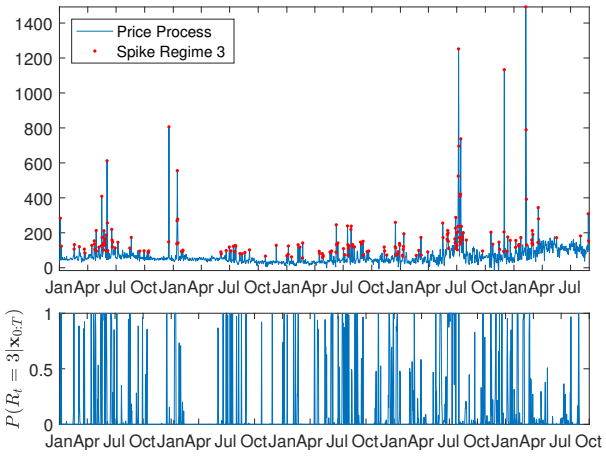
FIGURE 5.4: A sample of five residuals versus time plots for each regime in Model 1 of Type II. (Left) Residuals-versus-time PPC plots for Regime 1, the AR(1) regime. (Right) Residuals-versus-time PPC plots for Regime 3, the shifted-log-normal spike regime. The residuals of Regime 1 clearly increase over time, which suggests our assumptions of time-homoscedasticity is violated.



(A)



(B)



(C)

FIGURE 5.5: Prices classified into regimes according to their posterior probabilities using the rule $\mathbb{P}(R_t = i|\boldsymbol{x}_{0:T}) > 0.5$ for Model 2 of Type II. The plots at the top show the prices series with data highlighted red when it is classified into a regime. The plots at the bottom show the posterior probabilities, $\mathbb{P}(R_t = i|\boldsymbol{x}_{0:T})$. (a) Data allocated into base regime 1 (Regime 1). (b) Data allocated into base regime 2 (Regime 2). (c) Data allocated into the spike regime (Regime 3).

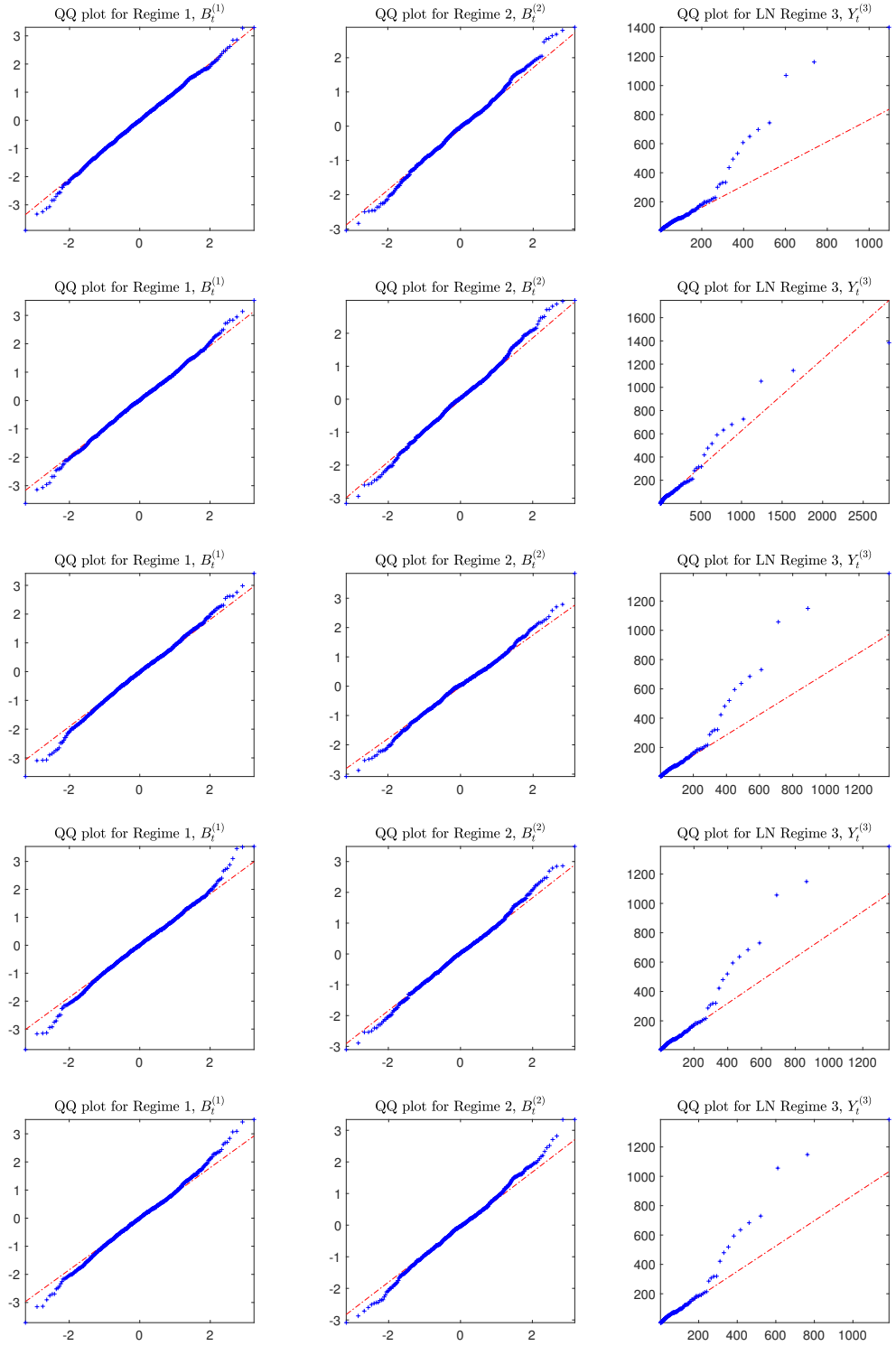


FIGURE 5.6: A sample of five QQ plot PPCs for the residuals of Regimes 1, 2 and 3 in Model 2 of Type II. (Left) QQ plot PPCs for the first AR(1) base regime, $B_t^{(1)}$. (Middle) QQ plot PPCs for the second AR(1) base regime, $B_t^{(2)}$. (Right) QQ plot PPCs for the first shifted-log-normal spike regime, $Y_t^{(3)}$. The points in the QQ plots for the spike regime (right) do not lie on a straight line, suggesting the single shifted-log-normal distribution is unable to capture extreme observations. However, this violation may not be too significant in practice, and more work is needed to determine this. The QQ plots for Regimes 1 and 2 suggest the assumptions about the AR(1) regimes are reasonable.

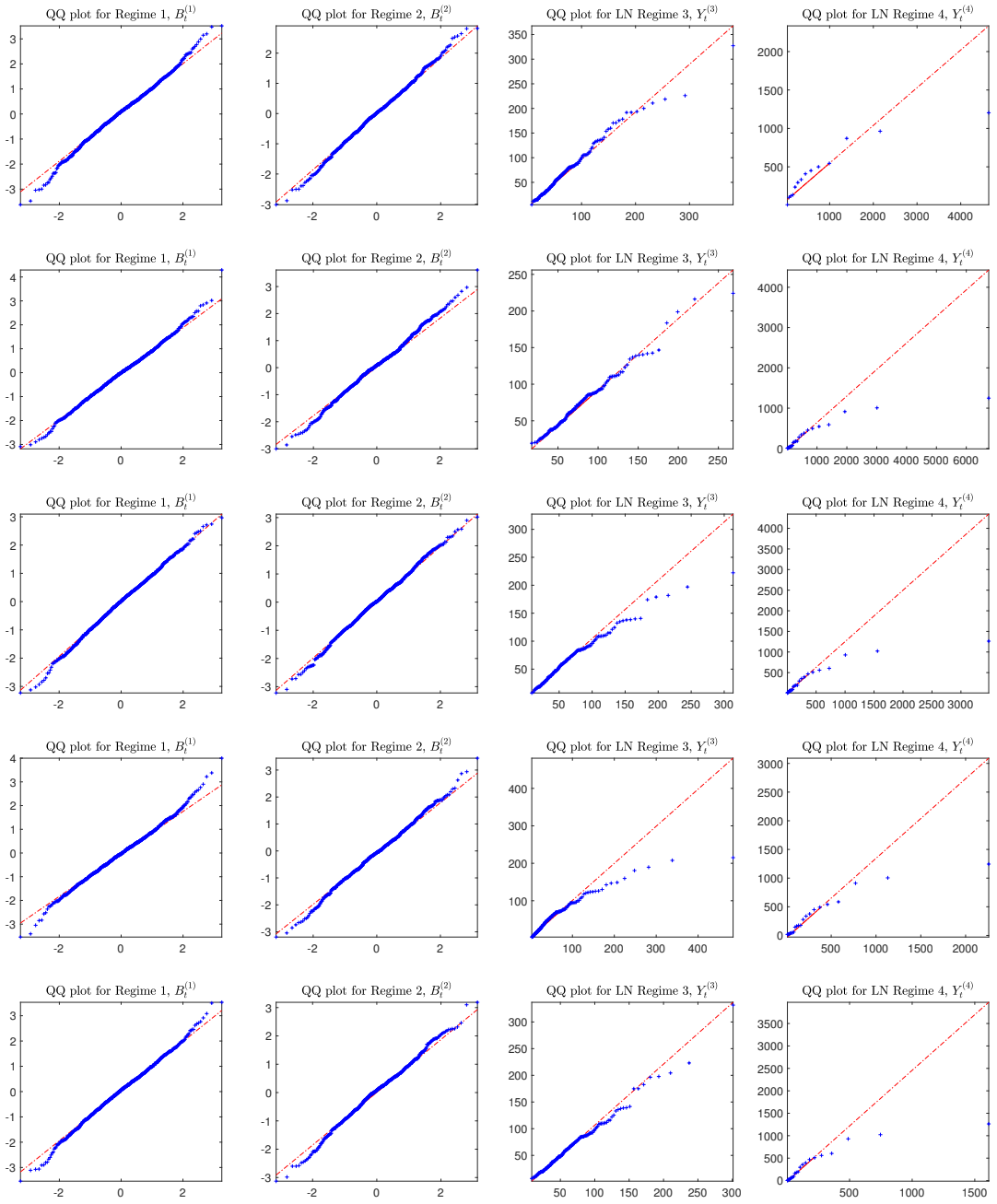


FIGURE 5.7: A sample of five QQ plot PPCs for each regime in Model 3 of Type II. (Left) QQ plot PPCs for base regime 1, $B_t^{(1)}$, an AR(1) regime. (Center-left) QQ plot PPCs for base Regime 2, $B_t^{(2)}$, another AR(1) regime. (Centre-right) QQ plot PPCs for Regime 3, $Y_t^{(3)}$, a shifted-log-normal spike regime. (Right) QQ plot PPCs for Regime 4, $Y_t^{(4)}$, a second shifted-log-normal spike regime for extreme spikes. The points in the QQ plots for the spike regimes (Regimes 3 and 4) stray from the reference line, suggesting the log-normal assumption may not be appropriate.

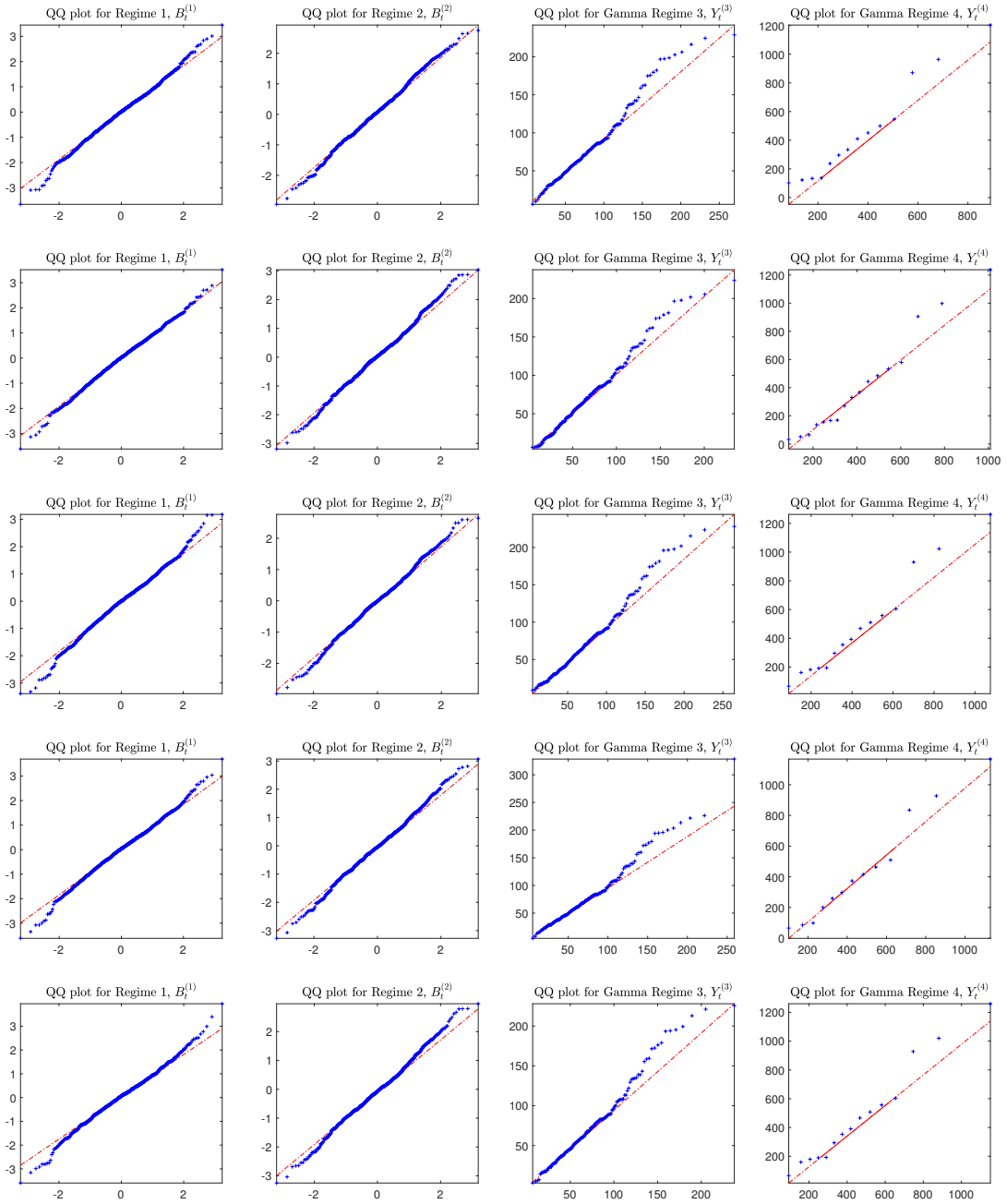
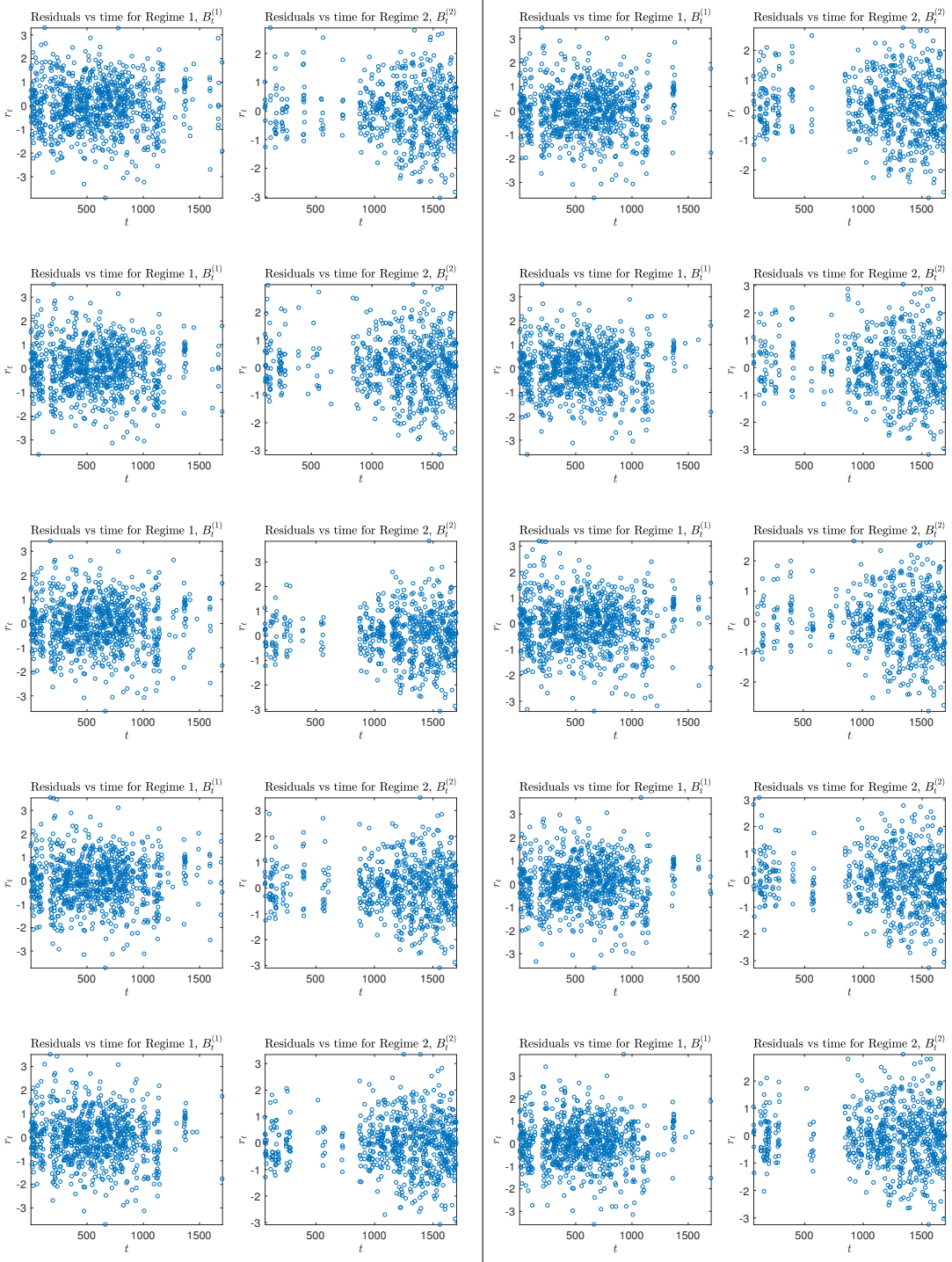


FIGURE 5.8: A sample of five QQ plot PPCs for each regime in Model 4 of Type II. (Left) QQ plot PPCs for base regime 1, $B_t^{(1)}$, an AR(1) regime. (Center-left) QQ plot PPCs for base Regime 2, $B_t^{(2)}$, another AR(1) regime. (Centre-right) QQ plot PPCs for Regime 3, $Y_t^{(3)}$, a shifted-Gamma spike regime. (Right) QQ plot PPCs for Regime 4, $Y_t^{(4)}$, a second shifted-Gamma spike regime for extreme spikes. The QQ plots for Regimes 1, 2 and 4 suggest the distributional assumptions are good for these regimes. However, there is some evidence to suggest Regime 3 is not well-modelled by a Gamma distribution, but this evidence is very not strong.



(A) Model 2 of Type II (B) Model 4 of Type II

FIGURE 5.9: A sample of five residuals versus time PPCs for each AR(1) regime in Models 2 (Left) and 4 (Right) of Type II. These PPCs show no obvious signs that the time-homoscedasticity assumption is violated for Regime 1. There are slight indications that the variance of the residuals from Regime 2 increase over time for both models. However, it is not clear how much change in variation is due to time-heteroscedasticity, or due to less observations at earlier times. We conclude that the time-heteroscedasticity assumption is reasonable for both models.

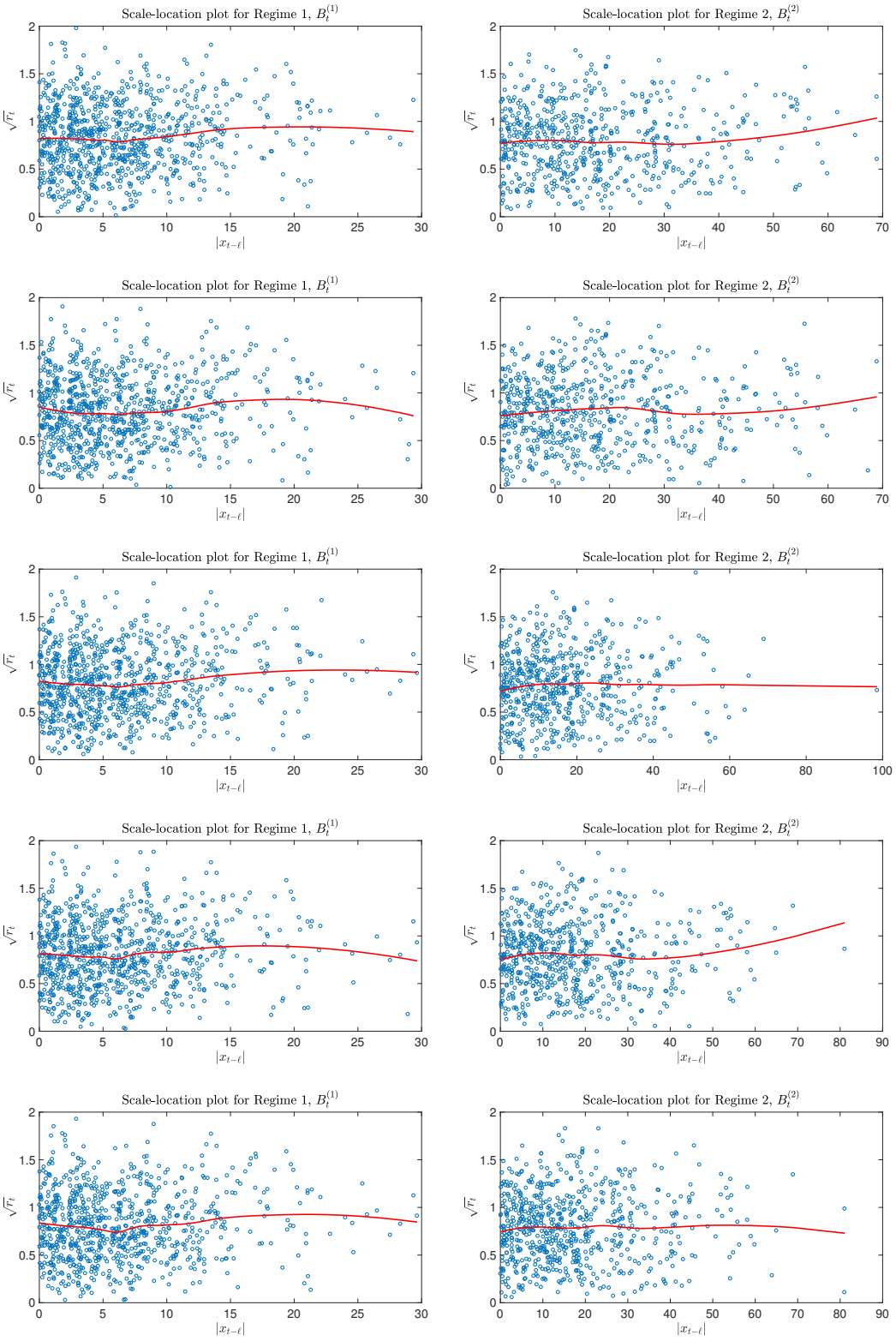


FIGURE 5.10: A sample of five scale-location PPCs for each AR(1) regime in Model 2 of Type II. These PPCs show no obvious signs that self-dependent-homoscedasticity assumptions are violated for either regime.

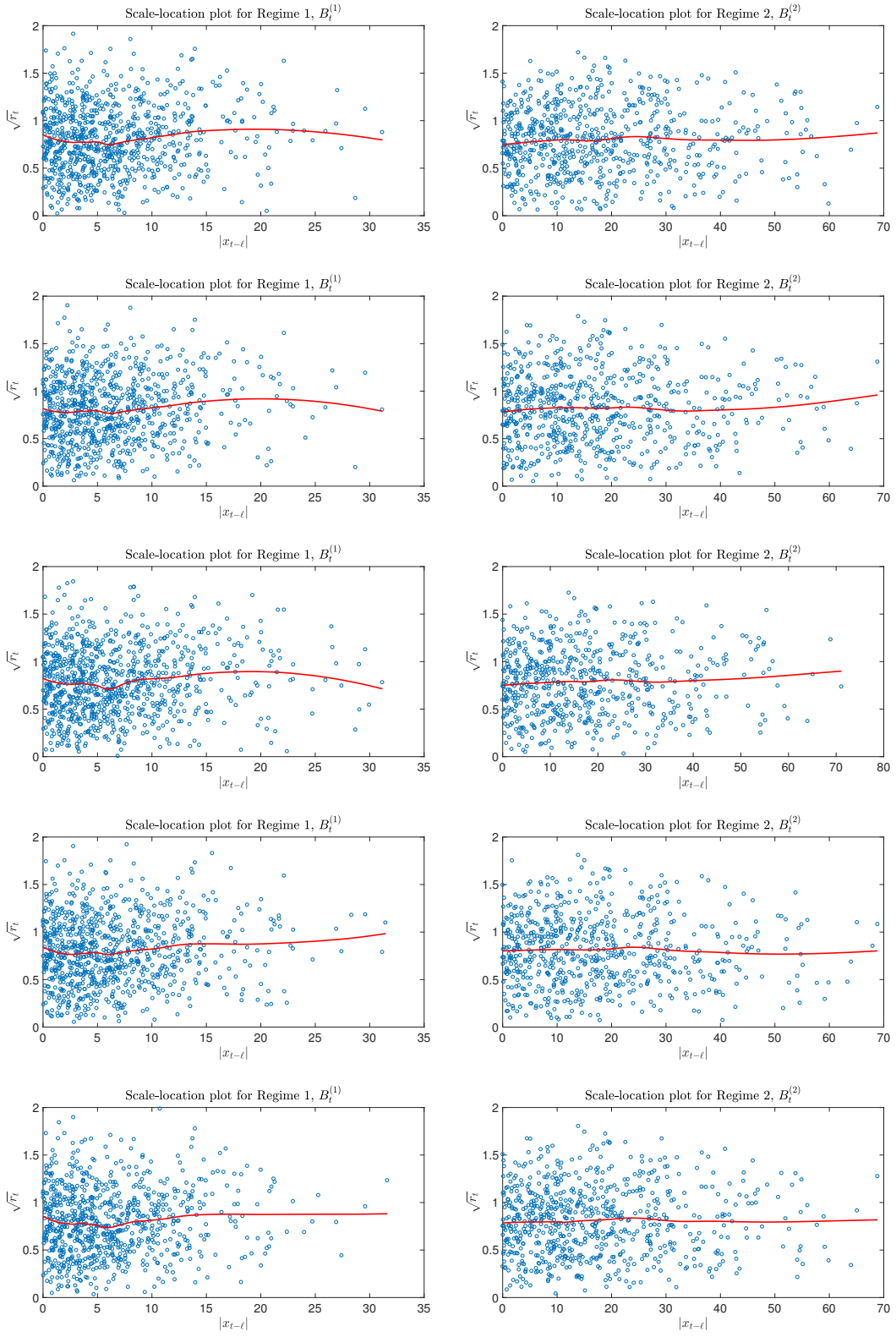
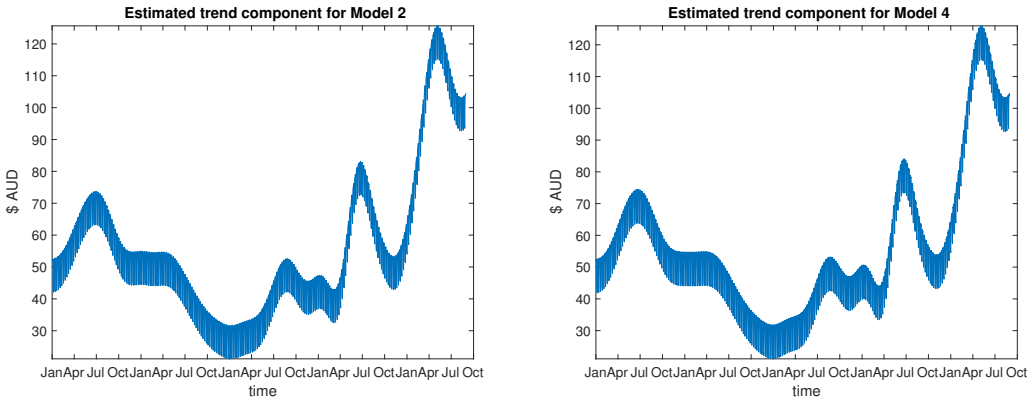


FIGURE 5.11: A sample of five scale-location PPCs for each AR(1) regime in Model 4 of Type II. These PPCs show no obvious signs that self-dependent-homoscedasticity assumptions are violated for either regime.



(A) Model 2 of Type II (B) Model 4 of Type II

FIGURE 5.12: Estimated trend components for Models 2 (Left) and 4 (Right) of Type II

Parameter	Model 2 of Type II	Model 4 of Type II
α_1	-0.0881	-0.218
ϕ_1	0.536	0.559
σ_1^2	50.2	49.5
α_2	0.283	0.601
ϕ_2	0.412	0.416
σ_2^2	413	409
q_3	17.6	9.711
μ_3	3.89	2.69
σ_3^2	1.34	26.5
q_4	-	167
μ_4	-	2.88
σ_4^2	-	165
Transition matrix	$\begin{pmatrix} 0.923 & 0.014 & 0.063 \\ 0.011 & 0.905 & 0.084 \\ 0.297 & 0.225 & 0.479 \end{pmatrix}$	$\begin{pmatrix} 0.923 & 0.019 & 0.056 & 0.001 \\ 0.016 & 0.900 & 0.080 & 0.003 \\ 0.304 & 0.279 & 0.370 & 0.047 \\ 0.116 & 0.094 & 0.378 & 0.411 \end{pmatrix}$

TABLE 5.2: Posterior mean estimates for the parameter of Models 2 and 4 of Type II.

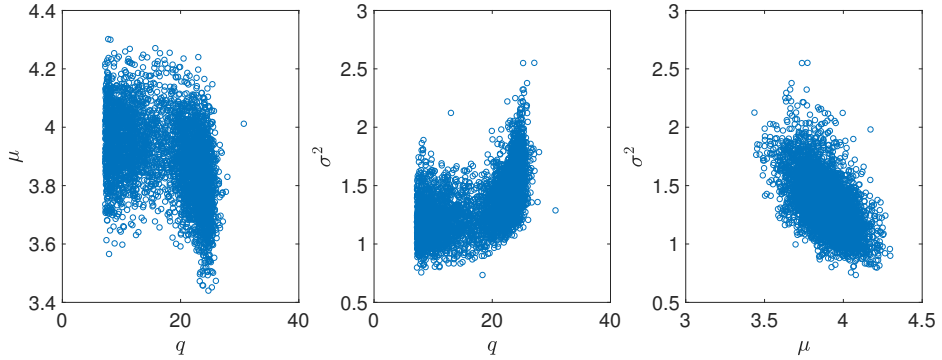


FIGURE 5.13: Bivariate scatter-plots of samples from the posterior distribution of parameters from Regime 3, $Y_t^{(3)}$, for Model 2 of Type II.

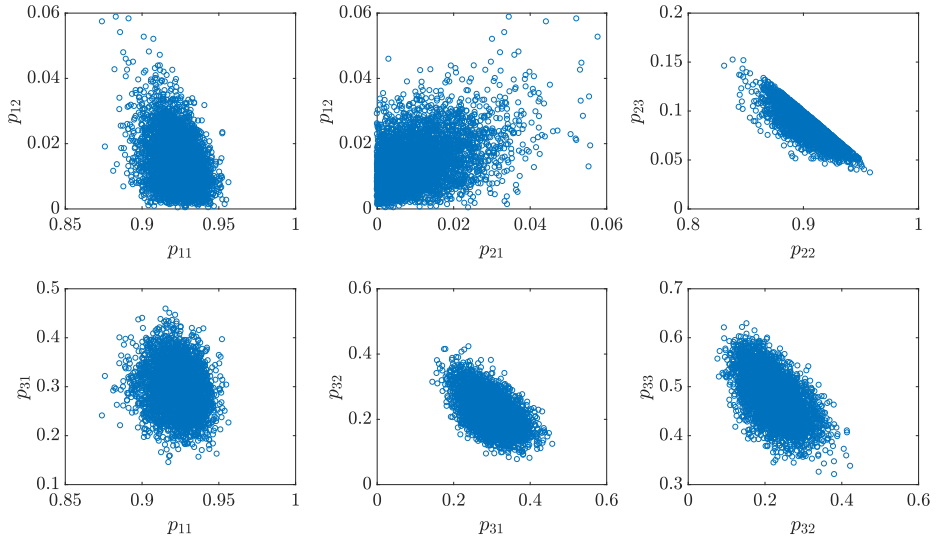


FIGURE 5.14: Bivariate scatter-plots of samples from the posterior distribution of the transition matrix P, for Model 2 of Type II.

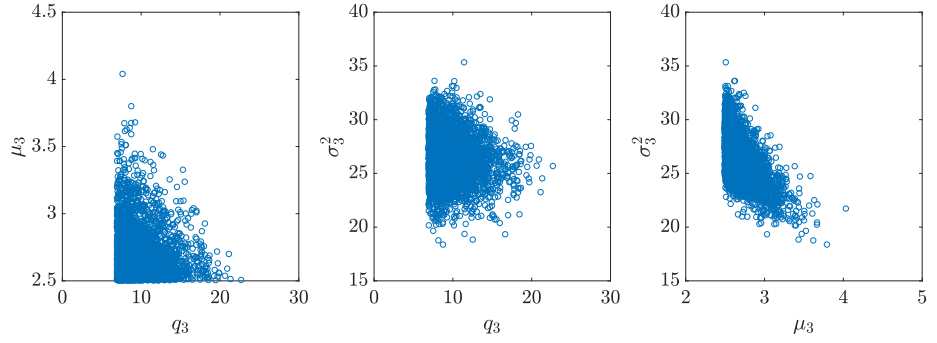


FIGURE 5.15: Bivariate scatter-plots of samples from the posterior distribution of parameters from Regime 3, $Y_t^{(3)}$, for Model 4 of Type II.

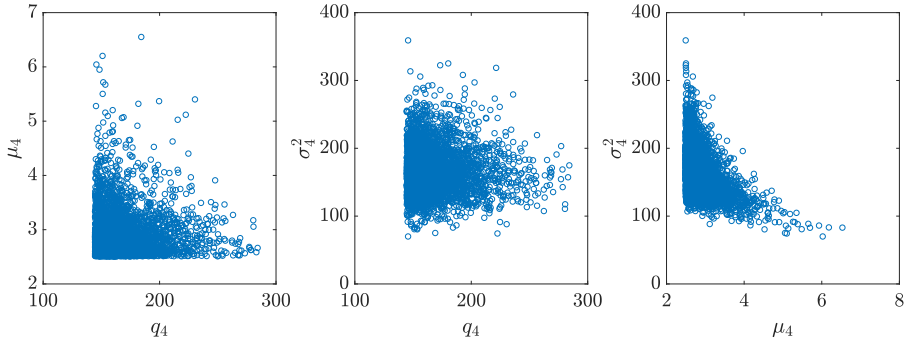


FIGURE 5.16: Bivariate scatter-plots of samples from the posterior distribution of parameters from Regime 4, $Y_t^{(4)}$, for Model 4 of Type II.

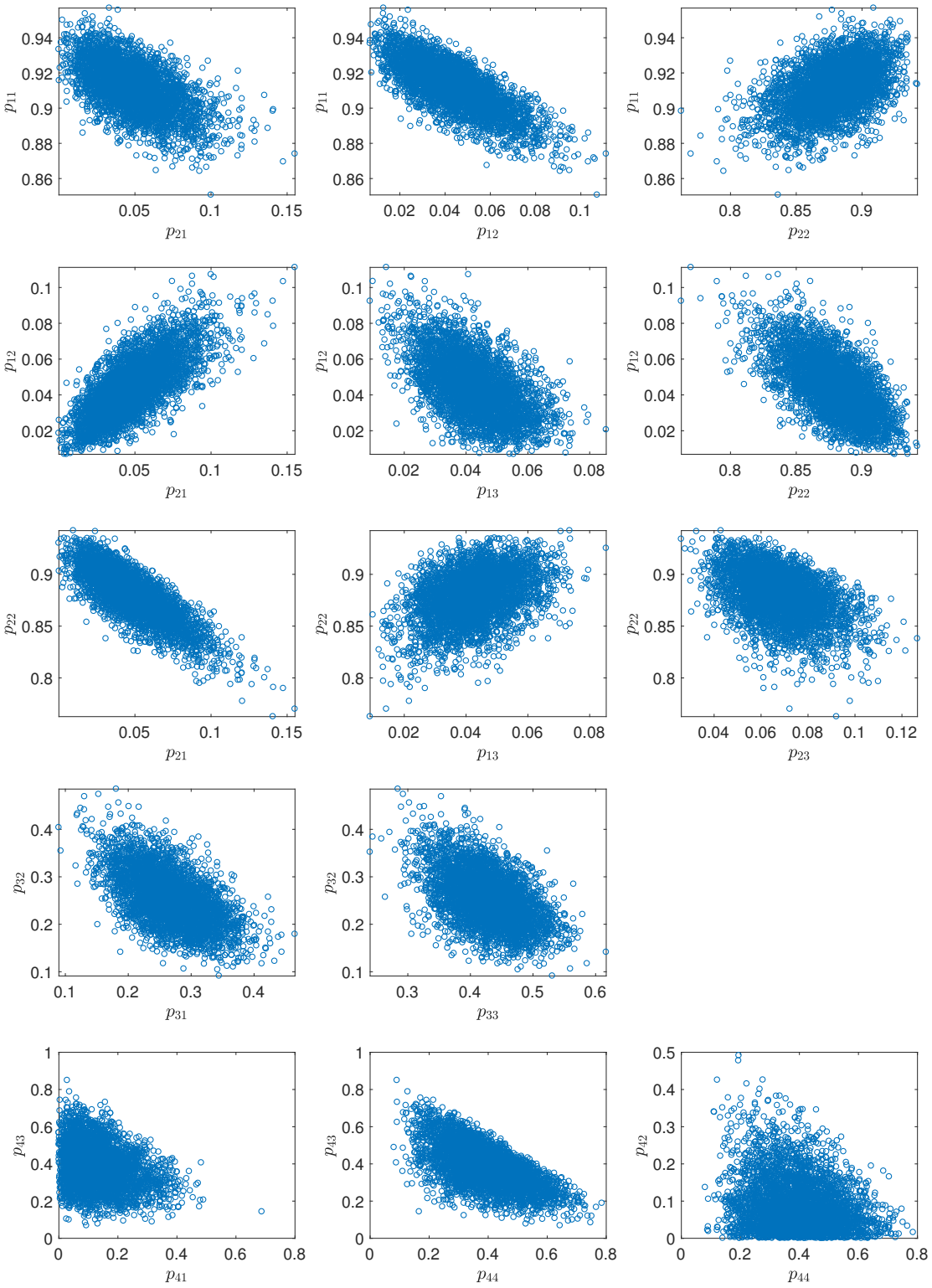


FIGURE 5.17: Bivariate scatter-plots of samples from the posterior distribution of the transition matrix P, for Model 4 of Type II.

5.4 Maximum likelihood model estimation

Here we encountered one practical limitation of our EM methodology, for models with two AR(1) regimes, the run time and memory requirements of the algorithm are impractically large (for the more complex models, we would have had to use the high memory nodes of the University's high-performance computer, and run the algorithm for twenty days). For this reason we simplify *all* models in this section by limiting the maximum memory of each within-regime process to 56 days (eight weeks), and after this, the within-regime process is assumed to be stationary. That is, we limit the value of $N_{t,i} \leq 56$, so at time t, given $R_t = i$, x_t can only possibly depend on x_{t-1} or x_{t-2} or ... or x_{t-56}, else x_t comes from the stationary distribution in Regime i. For MRS models of Type II we expect this simplification of the model to have minimal impact on the likelihood function and MLEs, since for MRS models of Type II the conditional distributions, $f^{\boldsymbol{\theta}}(x_t|R_t = i, N_{t,i} = n, \boldsymbol{x}_{0:t-1})$, decay exponentially to stationary as n gets large. For MRS models of Type III within-regime processes cease between visits, and therefore we do not see the same decay to stationary. However, we think it is reasonable to truncate the memory of within-regime processes for these models, since it is a reasonable approximation to assume this type of long-range dependence in prices is not a significant feature of the market.

Additionally, recall from Chapter 3 that maximum likelihood estimation of MRS models with shifted-log-normal regimes is not possible. For this reason, when shifted-log-normal distributions are included in our MRS models, we fix the shifting parameter at a value informed by the mode of the marginal posterior distribution of the parameter q in the corresponding Bayesian analysis. Furthermore, we also saw in Chapter 3 that many local maximisers can exist, particularly when trying to estimate shifting parameters. For this reason, we initialise our EM-algorithm-based model estimation procedure at 20 random starting points, and pick the terminating point with the highest likelihood to increase our chances of finding the MLE.

Our EM algorithm and trend estimation technique were used to iteratively filter the data as discussed in Section 5.1, and ultimately estimate the parameters of the models. In existing literature for electricity prices, once models are fitted to data, some information theoretic model selection criterion, such as AIC or BIC, is typically used to rank models on how well they fit the data. However, we believe this to be erroneous in this setting due to the model-dependent nature of the trend estimation technique used. Since the stochastic model is used to obtain a classification of data into regimes, and the trend is estimated from data classified into base regime(s) only, then the specification of the stochastic model affects the estimate of the trend. Therefore the stochastic component of each model is ultimately fitted to a slightly different dataset depending on the model

	Model 2	Model 4
	$X_t = \begin{cases} B_t^{(1)} & \text{if } R_t = 1, \\ B_t^{(2)} & \text{if } R_t = 2, \\ Y_t^{(3)} & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)} & \text{if } R_t = 1, \\ B_t^{(2)} & \text{if } R_t = 2, \\ Y_t^{(3)} & \text{if } R_t = 3, \\ Y_t^{(4)} & \text{if } R_t = 4, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{Gamma}(\mu_3, \sigma_3^2)$, $Y_t^{(4)} - q_4 \sim \text{Gamma}(\mu_4, \sigma_4^2)$.
QQ plots	Figure 5.18: distributional assumptions are suitable for Regimes 1 and 2, but questionable for Regime 3.	Figure 5.20: distributional assumptions are reasonable for all Regimes.
Residuals vs time	Figure 5.19 (A): the time-homoscedasticity assumption is suitable for Regime 1. Figure 5.19 (B): Regime 2 shows only slight indications that the variance of residuals increases over time, and part of the apparent increase in variation can be attributed to fewer observations at earlier times. We conclude that the time-homoscedasticity assumptions are reasonable.	Figure 5.21 (A): the time-homoscedasticity assumption is suitable for Regime 1. Figure 5.21 (B): Regime 2 shows only slight indications that the variance increases over time, and part of the apparent increase in variation can be attributed to fewer observations at earlier times. We conclude that the time-homoscedasticity assumptions are reasonable.
Scale-location	Figure 5.19 (C) and (D): self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.	Figure 5.21 (C) and (D): self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.

TABLE 5.3: A summary of our likelihood-based analysis for Type II MRS models for the SA dataset.

specification, and accordingly the likelihoods for each model are therefore incomparable [84]. In other words, estimation of the trend component varies between models, and cannot be directly accounted for in the value of the likelihood. Therefore, the traditional information theoretic measures of model fit are not applicable.

So instead we resort to common sense checks. We use the soft classification of data from the EM algorithm to classify data into regimes. In particular, we classify the data point x_t into Regime $j = \arg\max_{i \in \mathcal{S}} \mathbb{P}^{\widehat{\boldsymbol{\theta}}}(R_t = i | \boldsymbol{x}_{0:T})$, where $\widehat{\boldsymbol{\theta}}$ is the MLE. Using this classification we construct QQ plots, residuals versus time plots, and scale-location plots, exactly as for the PPCs in Section 5.3.

Maximum likelihood estimation for Type II models

Our likelihood-based analysis of Type II MRS models is summarised in Table 5.3.

First, consider Model 2 of Type II, which has two AR(1) base regimes, and one shifted-log-normal spike regime. We fit this model to the SA dataset using our EM algorithm methodology. Classifying data into regimes, we produced the QQ plots in Figure 5.18 to check within-regime distributional assumptions. These QQ plots suggest the distributional assumptions on the AR(1) regimes are suitable but the shifted-log-normal distribution is not capturing the very largest observations. We also produced residuals-versus-time plots, and scale-location plots for the residuals of the AR(1) regimes, to assess homoscedasticity assumptions; these are shown in Figure 5.19. Figure 5.19 (A) suggests the time-homoscedasticity assumption is suitable for Regime 1. Figure 5.19 (B) shows slight indications that the variance of Regime 2 may increase with time, although part of the apparent increase in variation can be attributed to the fact that there are fewer observations from Regime 2 at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2 of this model. The scale-location plots in Figures 5.19 (C) and (D) show no obvious evidence that the variance of Regimes 1 and 2 increase as a function of the magnitude of the last observed value from the same regime.

Now consider Model 4 of Type II, which has two AR(1) base regimes, and two shifted-Gamma spike regimes. Using our EM algorithm methodology, we fit this model also. Then, classifying data into regimes, we produced the QQ plots in Figure 5.20 to check within regime distributional assumptions. These QQ plots suggest the distributional assumptions of all four regimes are suitable. We also produced residuals versus time plots, and scale-location plots for the residuals of the AR(1) regimes, to assess homoscedasticity assumptions; these are shown in Figure 5.21. Figure 5.21 (A), suggests that the time-homoscedasticity assumption is reasonable for Regime 1. Figure 5.21 (B), shows slight indications that the variance of Regime 2 may increase over time, although, as is the case for Model 2 above, part of the apparent increase in the variance can be attributed to the fact that there are fewer observations in Regime 2 at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2 of this model. The scale-location plots in Figures 5.21 (C) and (D) show no obvious evidence that the variance of either regime increases as a function of the magnitude of lagged values.

The MLEs for Models 2 and 4 are shown in Table 5.4.

(A) (B) (C)

FIGURE 5.18: QQ plots of residuals from each regime for Model 2 of Type II, estimated by maximum likelihood. The QQ plots for the base regimes, (A) and (B), suggest the distributional assumptions of these regime are reasonable. The QQ plot of the shifted-log-normal spike regime, (C), suggests that the log-normal assumption may not be appropriate, at least in the tail of the distribution.

Parameter	Model 2	Model 4
α_1	-0.0420	-0.0658
ϕ_1	0.512	0.532
σ_1^2	45.3	50.6
α_2	0.127	0.280
ϕ_2	0.415	0.415
σ_2^2	431	382
q_3	14	11.9
μ_3	3.87	2.50
σ_3^2	1.16	21.9
q_4	-	168
μ_4	-	2.50
σ_4^2	-	104.6
Transition matrix	$\begin{pmatrix} 0.917 & 0.005 & 0.078 \\ 0.000 & 0.923 & 0.077 \\ 0.256 & 0.134 & 0.610 \end{pmatrix}$	$\begin{pmatrix} 0.929 & 0.008 & 0.062 & 0.000 \\ 0.000 & 0.906 & 0.092 & 0.002 \\ 0.313 & 0.260 & 0.377 & 0.050 \\ 0.062 & 0.048 & 0.456 & 0.433 \end{pmatrix}$

TABLE 5.4: MLEs of the parameter of Type II Models 2 and 4 for the SA dataset.

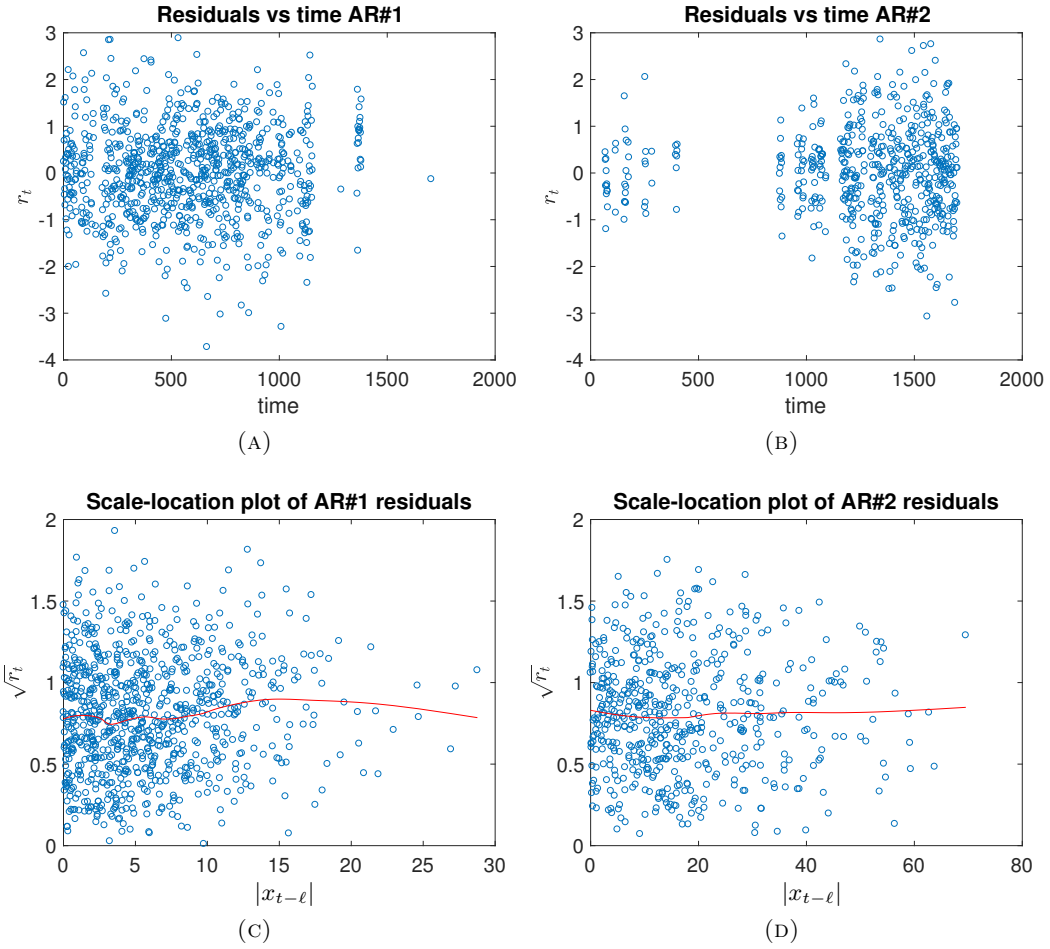
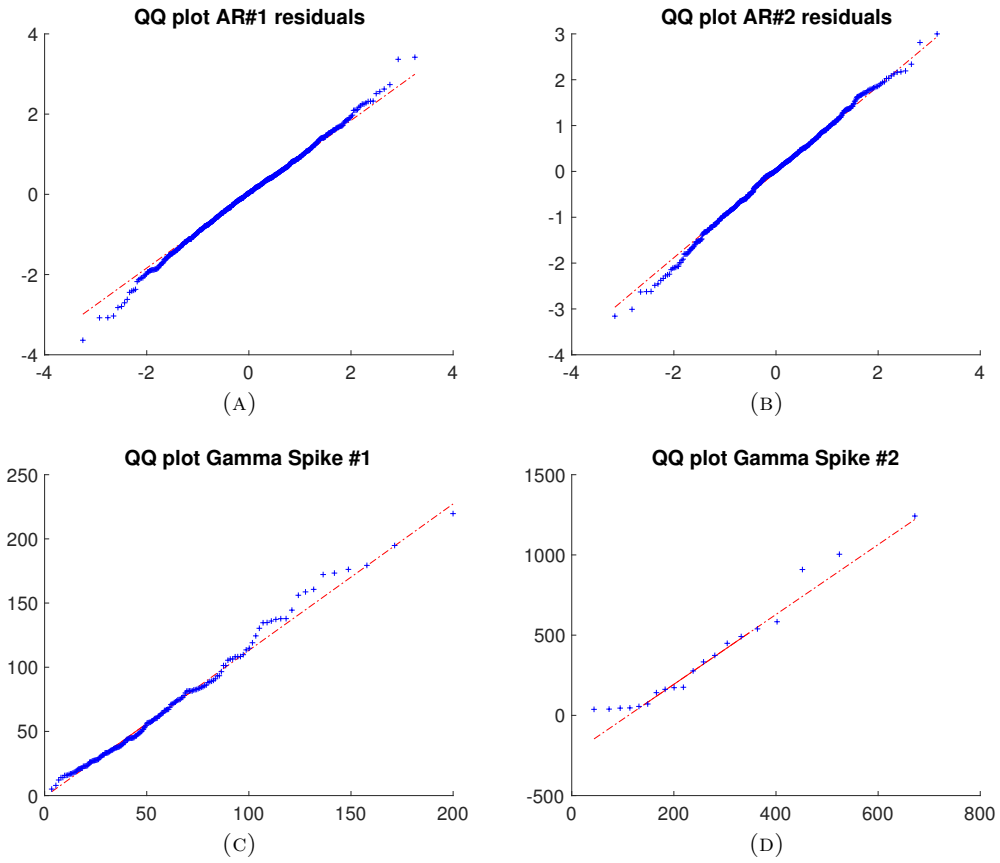


FIGURE 5.19: Residuals plots for AR(1) regimes of Model 2 of Type II, estimated by maximum likelihood. Figures (A) and (B) plot the raw residual against time for Regimes 1 and 2 respectively. Figures (C) and (D) plot $\sqrt{|r_t|}$ against the absolute value of the last observed value from the same regime, before time t, $|x_{t-\ell}|$. Figure (A) suggests there is no issue with the time-homoscedasticity assumption for Regime 1. Figure (B) shows slight evidence that the variance of Regime 2 may increase over time, although it is not clear how much of the apparent change in variation is due to time-heteroscedasticity, or due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2. Figures (C) and (D) suggest the variance of the residuals does not vary with the magnitude of lagged values.



(A) (B) (C) (D)

FIGURE 5.20: QQ plots of residuals from each regime for Model 4 of Type II, estimated by maximum likelihood. In all four plots, the points lie in a relatively straight line, suggesting the distributional assumptions are reasonable.

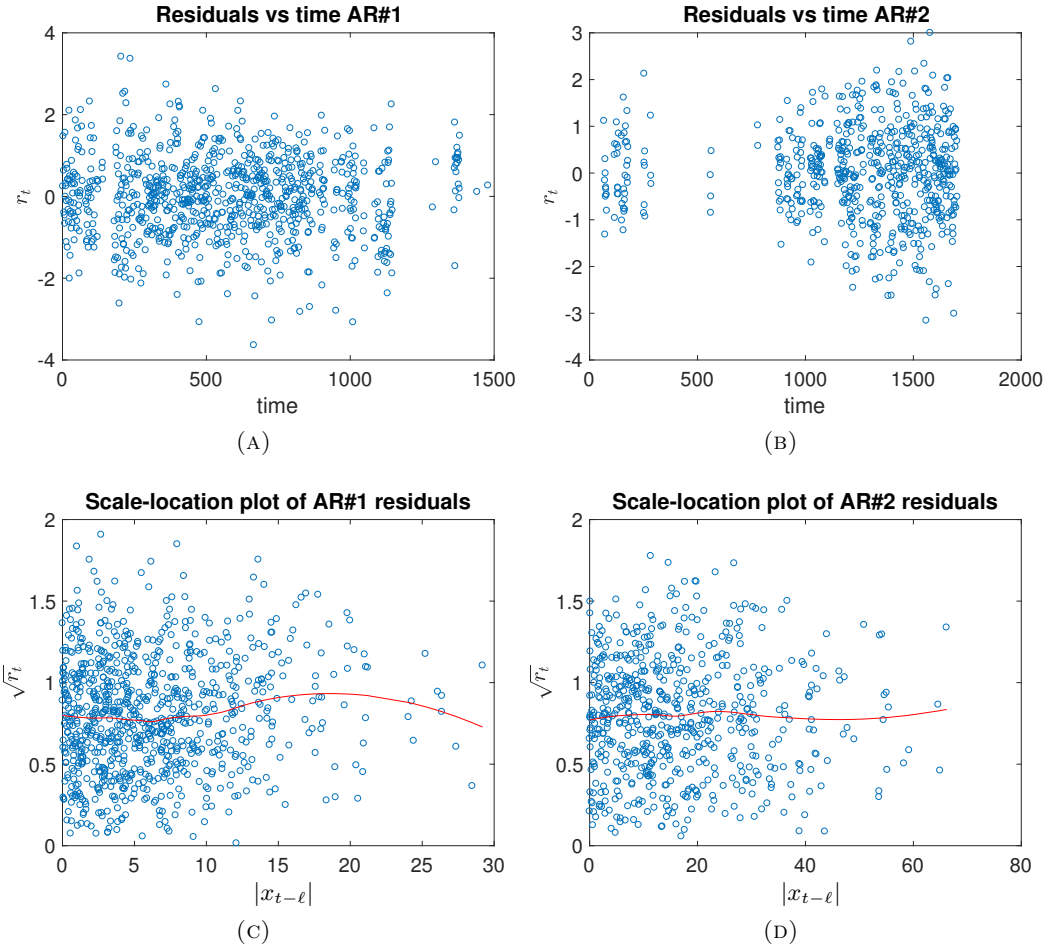


FIGURE 5.21: Residuals plots for AR(1) regimes of Model 4 of Type II, estimated by maximum likelihood. Figures (A) and (B) plot the raw residual against time for Regimes 1 and 2 respectively. Figures (C) and (D) plot $\sqrt{|r_t|}$ against the absolute value of the last observed value from the same regime, before time t. Figure (A) suggest there is no issue with the time-homoscedasticity assumption for Regime 1. Figure (B) shows only slight evidence that the variance of Regime 2 may increase over time, although it is not clear how much change in variation is due to time-heteroscedasticity, or due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2. Figures (C) and (D) show no obvious signs that the variance of Regimes 1 or 2 vary as a function of the magnitude of the last observed value from the same regime, $|x_{t-\ell}|$.

Maximum likelihood estimation for Type III models

In this likelihood setting, our analysis of Type III models has many similarities than the analysis of Type II models above, and so the details for Type III model fitting are reserved for Appendix B.2.

5.5 Discussion

In this chapter we apply techniques we developed in the previous chapters, to estimate the parameters of independent regime MRS models for the South Australian dataset. This dataset is far from friendly for two main reasons. First, the dataset contains a two-week market suspension, from the 28$^{\text{th}}$ of September 2016 to the 11$^{\text{th}}$ of October 2016. During this period prices were determined based on average prices of the past four weeks. We do not address this issue at all in our modelling and use the dataset 'as is'. Second, around April 2016 there is a significant jump in market volatility, which roughly coincides with the closure of South Australia's last remaining coal-fired generation facility. Thus, at this time there is a significant change in the market structure. We do not account for this directly, rather we specify two possible base regimes, one with a higher variance than the other. Upon fitting our models to the data, in either the Bayesian or likelihood framework, we see our models automatically pick out this change in market structure.

Another issue with our modelling is that the trend estimation technique that we employ has ramifications when it comes to model comparison. Our trend estimation technique relies on iterating[1] between estimating the stochastic model to classify data into regimes, and then, using the classified data, estimating the trend. This technique has been employed in previous literature, and shown to produce good estimates of the parameters of MRS models for electricity prices [57]. However, this technique can cause problems with model comparison due to the fact that the estimate of the trend component is dependent on the specification of the stochastic component model, and this is not directly accounted for in the value of the likelihood. Another way to think about this is, since the trend component estimated for each dataset is different, the stochastic model is fitted to a slightly different set of data for each different model specification. Thus, likelihood-based comparison techniques such as AIC, BIC or likelihood ratio, are not appropriate. Further work is needed to either investigate model-independent trend estimation techniques, or to include the trend directly in the stochastic component of the model, in which cases the use of likelihood-based model comparisons would then be permissable.

[1] Note that this issue also applies under the original implementation of this trend estimation method where this process is executed once, since the chosen model still affects the trend and hence the data used for fitting the stochastic component.

As a consequence, we rely on 'common-sense' model checking. In the Bayesian setting this comes in the form of our PPCs, which we use to assess within-regime distributional assumptions. Using our PPCs we narrow our search for a good model down to two candidates: Model 2, with two AR(1) base regimes and one shifted-log-normal spike regime, and Model 4 with two AR(1) base regimes and two shifted-Gamma-spike regimes. Since we cannot use AIC or BIC, we have no simple way to recommend either model and more work is needed, nor can we give preference to either Type II or Type III models.

We investigate Models 2 and 4 in our maximum-likelihood setting also. Our common-sense checks in this setting rely on classifying data into regimes based on the smoothed inferences obtained as part of the EM algorithm. We then produce QQ plots and residuals plots to assess within-regime distributional assumptions. One issue with this checking procedure is that the Markovian nature of the hidden regime sequence is not fully accounted for in this classification, and more work is needed here. One possible solution would be to develop a type of Viterbi algorithm for these independent regime MRS models, which would give a hard classification of the data while respecting the Markovian nature of the hidden regime process. Another possible solution would be to sample hidden sequences from the distribution $\mathbb{P}^{\widehat{\boldsymbol{\theta}}}(\boldsymbol{R}|\boldsymbol{x}_{0:T})$, which could then be used in a PPC-type framework, but with parameters fixed at $\widehat{\boldsymbol{\theta}}$.

Our modelling is further complicated by considering two types of independent regime MRS models, Types II and III. Type III models are simpler than Type II models since there are no unobserved values of within-regime processes in Type II models. However, our modelling does not give any indication about which type of model is preferable, and this is yet another area for further research.

Estimation of our models is complicated by the inclusion of shifted regimes. In the Bayesian setting this issue is not as troublesome, since the MCMC sampling technique is not attracted into areas of the likelihood that are infinity. Furthermore, for this dataset, we found that we had to restrict the support of the shifting parameters so that they remain in some reasonable range. We found that, if left unrestricted, the shifting parameter for the spike distributions would become negative and the spike distribution would capture base prices. In the maximum-likelihood setting, we are forced to fix the value of the shifting parameter for shifted-log-normal distributions, and we must restrict the shape parameters of the shifted-Gamma distribution to ease optimisation issues. However, this does not eliminate all convergence issues, since, as we saw in Chapter 3, the likelihood can have many local maximisers, especially when shifting parameters are to be estimated by maximum likelihood. For this reason, we suggest that Bayesian estimation of parameters of MRS models be used when shifting parameters are involved.

Lastly, we understand that this is a preliminary analysis of the application of MRS models to this dataset, and is by no means exhaustive. For example, one simple refinement of our work would be to investigate models with two base regimes where only the variance of the processes is allowed to differ. We have also not challenged the time-homogeneous Markovian assumptions in the model. It would be interesting to investigate models for which the hidden regime process is time varying, such as in [80], dependent on the time since the last spike, a feature sometimes used in *spike-only models*, or semi-Markovian, and we believe our methods can be extended for estimation of all of these cases. Another feature of electricity prices that we have not considered in our modelling is dependence on exogenous factors such as weather (as in, for example, [80] or [77]), and this is another possibility for future research.

Chapter 6

Conclusion

In this thesis we have developed novel forward, backward and EM algorithms to evaluate the likelihood for, and find MLE of, independent-regime MRS models, and investigated issues related to these methods. We have also developed a Bayesian framework for inference of independent-regime MRS models for electricity prices, which has not been done in this detail to date. Furthermore, we have provided an initial analysis of the South Australian electricity market using our methods. Here we recap our findings of each chapter, discuss lessons learned, and describe some future work.

Chapter 3: Likelihood methods for MRS models with independent regimes

Findings In this chapter we first discussed the EM-like algorithm, which, until this thesis, was the current method of choice for MRS models with independent regimes. We showed that the EM-like algorithm has some theoretical failings, and provided examples where it failed to recover parameters from simulated datasets.

We then developed novel and computationally feasible, forward, backward and EM algorithms to evaluate the likelihood for, and find maximum likelihood estimates of, independent-regime MRS models. We followed this by a discussion of issues related to these methods: a comparison of our EM algorithm to 'black-box' optimisation, bias and consistency, difficulties of shifted distributions, and extensions of our work.

Of note, we found there can be numerous local maximisers of the likelihood, and this issue is accentuated when shifted distributions are included in the model. Restricting the parameter space when searching for maximisers can increase the chances of finding the true global maximiser, but may not eliminate erroneous behaviours. Via simulations, we showed that the maximum likelihood estimator appears to be a consistent estimator,

and minimal bias is present for reasonably-sized datasets. We described some of the known difficulties when estimating the parameters of shifted-log-normal and shifted-Gamma distributions via maximum likelihood, and then discussed how these issues are even more apparent when included in an MRS model. We concluded that maximum likelihood estimation of the shift parameter for shifted-log-normal distributions in an MRS context is not possible, and it is best to restrict the scale parameter of the shifted-Gamma distribution to reduce erroneous behaviour when searching the likelihood surface for the MLE.

Lessons learned and future work Limitations of our algorithms are their time and memory complexity, which can make exact computations impractical for models with just two AR(1) regimes. As such, one area of future work is to investigate computational techniques to reduce time and memory requirements. In practice our algorithm can produce many quantities that are zero, and a smart implementation of our algorithms would be able to take advantage of this to reduce complexity.

In Chapter 5 we resorted to truncating the memory of the counting processes in our models to reduce computational demands. So, another area for future research would be to investigate the effects of this approximation. Our algorithm could be used for more complex models, where transition probabilities of the hidden regime sequence are allowed to depend on exogenous variables. Lastly, it would be useful to prove consistency results for our algorithm.

Chapter 4: Bayesian inference methods for independent-regime MRS models

Findings In this chapter we explored a Bayesian framework for parameter inference of MRS models with independent regimes; something which has not been done in this detail before. We implemented a data-augmented MCMC algorithm which enabled efficient sampling of the posterior distribution. One advantage of a data-augmented framework is that it enables $\mathcal{O}(T)$ computations, whereas our likelihood methods are $\mathcal{O}(T^{k+1})$ where k is the number of AR(1) components in the model.

We described some practical issues faced when implementing our algorithms, and some expressions to simplify computations. We also implemented an adaptive scheme to automatically tune our MCMC algorithm, which made our algorithm practical for a wide range of models and datasets. Our main tools for model checking in this Bayesian environment are PPCs, where samples from the posterior are used to calculate statistics that can then be compared to model assumptions. We provided simulations to show

that our Bayesian methods are valid and our PPCs have some power to reject models when they are false.

Lessons learned and future work In this chapter we implemented three PPCs which target distributional assumptions of within-regime processes only. We investigated some other PPCs based on *the periodogram*, *autocorrelation function* and *partial autocorrelation function*, but our analysis of these was not thorough, and is omitted from the thesis. One model assumption that we did not assess is the time-homogeneous Markovian assumption of the hidden regime process. Thus, it would be interesting to investigate PPCs to assess these assumptions. Another assumption that we did not address was the independence within the spike regimes, and PPCs could be developed to assess this assumption also.

Chapter 5: Applications to South Australian electricity prices

Findings We started by introducing the South Australian dataset which consists of prices from 1st of January 2013, to the 31st of September 2017. Interesting features of this dataset are the significant jump in volatility during 2016 – which roughly coincides with the closure of SA's only coal generation facility, and therefore a significant change in market structure – and the magnitude of price spikes, and a period of 14 days during which the market was suspended. We then detailed our trend estimation method. Since extreme observations can bias the estimate of trend components, we used an iterative method to remove and replace extreme values, and then estimate the trend on this altered dataset. Our trend model was built out of two parts, a short-term periodic component to capture weekly seasonalities, and a long-term component estimated using wavelet filtering.

We used our Bayesian methodology to fit models to the dataset and assess their goodness-of-fit. Using our PPCs we concluded that prices are well-modelled by MRS processes that include two AR(1) base processes, due to the significant jump in price volatility in 2016. We observed that one AR(1) regime, with a lower volatility, predominantly modelled prices before April 2016, and the other predominantly model prices from this point onward. To capture spikes, our PPCs suggested either a single shifted-log-normal regime, or two shifted-Gamma distributions are suitable. The model with two shifted-Gamma spike regimes was motivated by the fact that the shifted-log-normal regime struggled to capture the most extreme observations. We used one shifted-Gamma distribution to capture typical spikes, and one to capture the most extreme spikes. We also found that price drops, if there are any, are preferably modelled by AR(1) processes than by a drop

regime, since our Bayesian analysis allocated no probability mass to a drop regime when it was included in models with two AR(1) regimes.

We then applied maximum likelihood methods to the two final models from our Bayesian analysis. Since one of these models included a shifted-log-normal regime, we fixed its shifting parameter at the mode of the marginal posterior distribution for the corresponding parameter from our Bayesian analysis. We noted that typical model comparison techniques, such as AIC or BIC, are not suitable in this setting. This is due to the way in which we estimated the trend component. The trend was estimated for each model by iteratively using the stochastic itself model to classify and remove extreme prices. As a result, the stochastic model is ultimately being fitted to a different dataset for each model, which means any likelihood-based measure of fit cannot be used to compare models. Hence we resorted to using 'common-sense' model checking, by classifying prices into regimes and producing QQ plots and residuals plots using this classification. Our QQ plots tend to agree with our Bayesian analyses about the appropriateness of the distributional assumptions.

We concluded Chapter 5 with a brief discussion of our methods, and some suggested improvements.

Lessons learned and future work This chapter sets the scene for more future work, which we had intentions to cover in this thesis, but instead we developed computationally feasible likelihood methods, which absorbed much of our time. First, since electricity prices are known to be affected by weather, business activities, day of the week, it would be interesting to investigate including exogenous predictors in our model. Exogenous factors could be included in any, or all of, the mean of within-regime processes, the overall trend component, the volatility component, or the regime-switching component. In fact, we extended our Bayesian methodology to be able to cope with models including exogenous factors in the regime-switching probabilities via a multinomial logistic regression, and in the mean of the AR(1) base regime (not presented in this thesis), but never fully investigated this approach since we shifted our focus to the development of the maximum likelihood methods.

Another related area of future research is to challenge the time-homogeneous Markovian assumption of the hidden regime process, an assumption that is likely violated by this dataset. One possibility would be to include a dependence on the time since the last observed spike in the model. This could be achieved by supposing the transition probabilities of $\{\boldsymbol{H}_t\}$ depend on the counting process $\{\boldsymbol{N}_t\}$, as well as the regime process $\{R_t\}$. Another interesting possibility would be to include a semi-Markovian structure in

the hidden regime process, and we believe our methods can be extended to accommodate this.

Yet another related area of future research is to include the trend model within the stochastic model, so that the likelihood accounts for the model-dependent trend. This would mean that AIC or BIC could be used for model comparison. Another advantage of including the trend model in the stochastic component would be that one could test significance of coefficients of the trend model using likelihood ratio techniques. Alternatively, model-independent trend estimation techniques could be used, which would also permit the use of AIC, BIC, or likelihood ratio tests for comparing components of the stochastic model only.

Finally, one more important aspect of modelling, which this thesis does not address, is the use of cross-validation techniques or out of sample testing, and we recommend this as an important area for future research.

Closing remarks

Here we have provided a brief analysis of South Australian electricity prices using MRS models with independent regimes, and we hope our methods will be used in future work investigating electricity markets. We would recommend the South Australian electricity market as a case study due to its many interesting features. We believe this thesis leaves such an analysis well-posed and informed, and we hope researchers learn from our lessons summarised above. More generally, we hope our methods find applications elsewhere, and make valuable contributions to those fields.

Appendix A

General-state-space Discrete-time Markov Chains

A Markov Chain is a sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ that have the *Markov property*. Define $\mathcal{S}$ as the *state space*, which is the set of possible values that X_t can take and define Σ as a *σ-algebra*, which, in a rough sense, corresponds to the set of all subsets of $\mathcal{S}$ that we could possibly be interested in. More specifically, a σ-algebra is a collection of subsets of $\mathcal{S}$ that contains the empty set, is closed under complement, and countable unions. The Markov property says that the probability of moving into a set $A \in \pm$ at time $t+1$, given the entire history of the process $X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0$ for $x_0, ..., x_t \in \mathcal{S}$, depends only on the current position of the chain, $X_t = x_t$. Thus

$$\mathbb{P}\left(X_{t+1} \in A | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0\right) = \mathbb{P}\left(X_{t+1} \in A | X_t = x_t\right),$$

assuming both conditional probabilities are well defined. A Markov chain is called *time homogeneous* if $\mathbb{P}(X_{t+1} \in A | X_t = i) = \mathbb{P}(X_1 \in A | X_0 = i)$ for all $t \in \mathbb{N}$, $i \in \mathcal{S}$, $A \in \pm$.

When $\mathcal{S}$ is countable the probabilities $p_{ij} := \mathbb{P}\left(X_{t+1} = j | X_t = x_t\right)$, $i, j \in \mathcal{S}$, are known as *transition probabilities*. Since, $\mathcal{S}$ is countable then there is a one-to-one mapping between some set $\{1, 2, ...\} =: N \subseteq \mathbb{N}$ and $\mathcal{S}$, so, without loss of generality, we may assume N is the state space which makes notation simpler. The transition probabilities and are often represented collectively as the (possibly infinite dimensional) *transition matrix*

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

When $\mathcal{S}$ is a more general space, we represent the movement of the chain using the *transition kernel*, $K(x, A) := \mathbb{P}(X_{t+1} \in A | X_t = x)$ for $A \in \Sigma$.

For example, an AR(1) process is a Markov chain on $\mathbb{R}$ with transition kernel

$$K(y_{t-1}, A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t - \alpha - \phi y_{t-1})^2} dy_t.$$

The n-step probabilities are defined as $\mathbb{P}(X_{t+n} \in A | X_t = x)$. It can be shown that the n-step transition probabilities are given by

$$P^{(n)} = P^n$$

in the countable-state-space case. For general state space processes we define the n-step transition kernel as

$$K^{(n)}(x, A) := \int_{\mathcal{S}} ... \int_{\mathcal{S}} K(x, dy_1)...K(x, dy_{n-1})K(y_{n-1}, A).$$

A discrete-state Markov chain is said to be *irreducible* if it is possible to get from any state to any other state, i.e. if for all $i, j \in \mathcal{S}$, there exists an $n \in \mathbb{N}$ such that

$$\mathbb{P}(X_n = i | X_0 = j) > 0.$$

For general state space chains the definition of irreducibility is slightly more complex since we have to take into account the size of sets in Σ. Suppose we use the measure $\psi(\cdot)$ to assign a notion of size to sets in Σ, so $\psi(A) \geq 0$ for all $A \in \Sigma$. We say that a Markov chain on a general space $\mathcal{S}$ is ψ-irreducible if, for every $A \in \Sigma$ with $\psi(A) > 0$, and for every $x \in \mathcal{S}$, $K^{(n)}(x, A) > 0$ for some $n \in \mathbb{N}$. In words, this means that from any point (x) in the state space, there is positive probability of reaching any 'sufficiently big' set (A such that $\psi(A) > 0$) after some number of transitions (n).

We define $\pi(\cdot)$ as the invariant measure of the transition kernel $K(x, A)$, if it satisfies

$$\pi(A) = \int_{\mathcal{S}} K(x, A) d\pi(x), \quad \forall A \in \Sigma$$

and note that it may not always be possible to find such a measure (in the discrete case the integral is replaced by the appropriate sum). If π is a finite measure ($\pi(\mathcal{S}) < \infty$), then we say that the associated Markov chain is *positive recurrent* and we may also call the measure $\frac{\pi(\cdot)}{\pi(\mathcal{S})}$ the stationary distribution; otherwise the associated Markov chain is *null recurrent* or *transient*. Recurrence implies that a Markov chain returns to every set

$A \in \Sigma$ with $\psi(A) > 0$ with positive probability, and positive recurrence implies that the expected return time between visits is finite.

In the discrete-state-space case, the period k, of a Markov chain is defined as

$$k := gcd\{n > 0 : \mathbb{P}(X_n = x | X_0 = x) > 0\},$$

where *gcd* means the greatest common divisor. An analogous definition of period exists for Markov chains on general spaces, but this requires us to define numerous objects that are beyond the scope of what is needed for this thesis. The interested reader should consult Meyn and Tweedie [75], on which this section is based. For us it suffices to know that if a Markov chain has *period* k, then transitions from a set C with $\psi(C) > 0$ to itself, can only occur with positive probability at multiples of k time steps. If a Markov chain has $k = 1$ then the chain is said to be *aperiodic*. A Markov chain is called *strongly aperiodic* if $K(x, x) > 0$ for all $x \in \mathcal{S}$.

Now we can describe an important result regarding convergence of Markov chains to their stationary distribution. Suppose that an irreducible Markov chain admits a finite invariant probability measure π, then π is unique, and

$$\sup_{A \in \Sigma} |K^{(n)}(x, A) - \pi(A)| \to 0,$$

as $n \to \infty$, for every $x \in \mathcal{S}$. That is, π is the limiting distribution of the Markov chain. The fact that π is unique follows from Theorems 10.0.1 and 10.1.2 of Meyn and Tweedie [75], while the fact that it is the limiting distribution follows from Theorem 13.0.1 of Meyn and Tweedie [75] also.

A transition kernel $K(x, A)$ is said to be *reversible* with respect to a measure π if

$$\int_{\mathcal{S}} \int_{\mathcal{S}} g(x, y) \pi(dx) K(x, dy) = \int_{\mathcal{S}} \int_{\mathcal{S}} g(y, x) \pi(dx) K(x, dy),$$

for any bounded function g. When this is the case, then π is invariant for the kernel $K(x, A)$. This is a key result used to construct Markov chains with specific stationary distributions, such as those used in Markov chain Monte Carlo (Section 2.2.5), and coupled with the convergence result above, ensures convergence to the stationary distribution of the Markov chain.

Appendix B

Model fitting and checking

B.1 Bayesian Model selection for Type III MRS models

In Table B.1 we summarise our Bayesian model selection for MRS models of Type III. Figures B.1-B.8 show our checking for Type III models. The estimated trend components for Models 2 and 4 of Type III are shown in Figure B.9, and the posterior means are shown in Table B.2.

	Model 1	Model 2	Model 4
	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \\ Y_t^{(4)}, & \text{if } R_t = 4, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{Gamma}(\mu_3, \sigma_3^2)$, $Y_t^{(4)} - q_4 \sim \text{Gamma}(\mu_4, \sigma_4^2)$.
QQ-plots	Figure B.1: distributional assumptions violated for Regimes 1 and 3.	Figure B.3: distributional assumptions are suitable for Regimes 1 and 2, but questionable for Regime 3.	Figure B.5: distributional assumptions are suitable for all regimes.
Residuals vs time	Figure B.2: variance is non-constant over time.	Figure B.6: there are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance is due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for both AR(1) regimes.	Figure B.6: there are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance is due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for both AR(1) regimes.
Scale-location	Not shown.	Figure B.7: self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.	Figure B.8 self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.
Comments	Model not suitable.	Regime 2 is included to capture time-heteroscedasticity. QQ plots in Figure B.3 suggest shifted-log-normal spikes are more suitable than shifted-Gamma spikes (not shown). When two AR(1) regimes are included in a model, no drop regime is necessary.	Regime 4, a second spike regime, is included to capture the very largest spikes. QQ plots in Figures B.4 and B.5 show slight differences between sifted-log-normal spikes (Model 3) and shifted-Gamma spikes, but suggest shifted-Gamma spike are more suitable.

TABLE B.1: A summary of our Bayesian model selection process for Type III MRS models for the SA dataset.

Parameter	Model 2	Model 4
α_1	-0.239	-0.411
ϕ_1	0.523	0.548
σ_1^2	49.8	48.8
α_2	0.323	0.596
ϕ_2	0.418	0.420
σ_2^2	415	413
q_3	18.9	10.2
μ_3	3.88	2.71
σ_3	1.38	26.5
q_4	-	168
μ_4	-	2.88
σ_4	-	165
Transition matrix	$\begin{pmatrix} 0.921 & 0.020 & 0.059 \\ 0.016 & 0.895 & 0.089 \\ 0.292 & 0.239 & 0.469 \end{pmatrix}$	$\begin{pmatrix} 0.920 & 0.026 & 0.053 & 0.001 \\ 0.024 & 0.890 & 0.083 & 0.004 \\ 0.297 & 0.294 & 0.361 & 0.048 \\ 0.128 & 0.082 & 0.380 & 0.410 \end{pmatrix}$

TABLE B.2: Posterior mean estimates for the parameter of Type III Models 2 and 4 for the SA dataset.

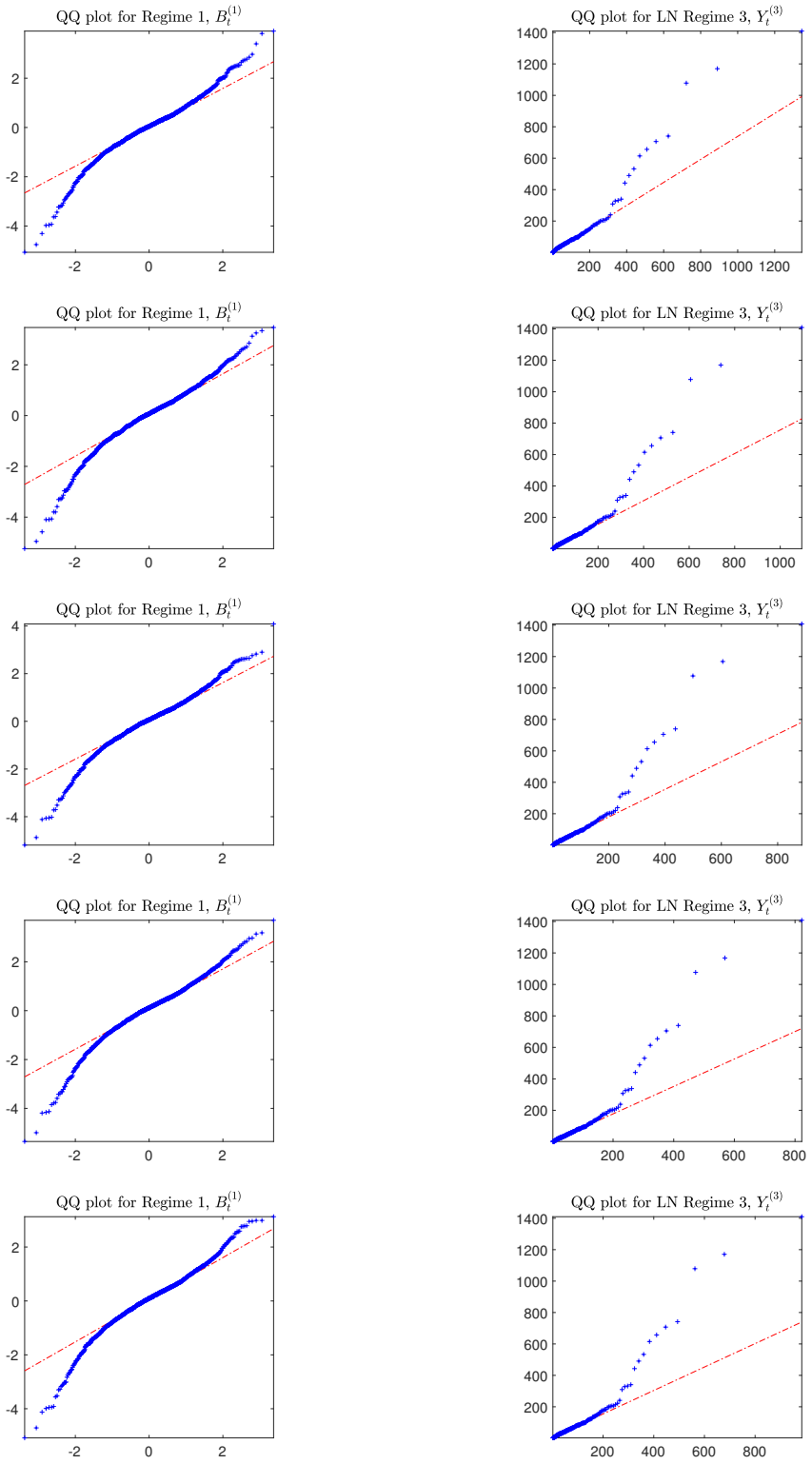


FIGURE B.1: A sample of five QQ-plot PPCs for each regime in Model 1 of Type III. (Left) QQ-plot PPCs for Regime 1, $B_t^{(1)}$, the AR(1) regime. (Right) QQ-plot PPCs for Regime 3, $Y_t^{(3)}$, the shifted-log-normal spike regime. The points in the QQ plots for both regimes clearly do not lie on a straight line, suggesting this model does not capture the data well.

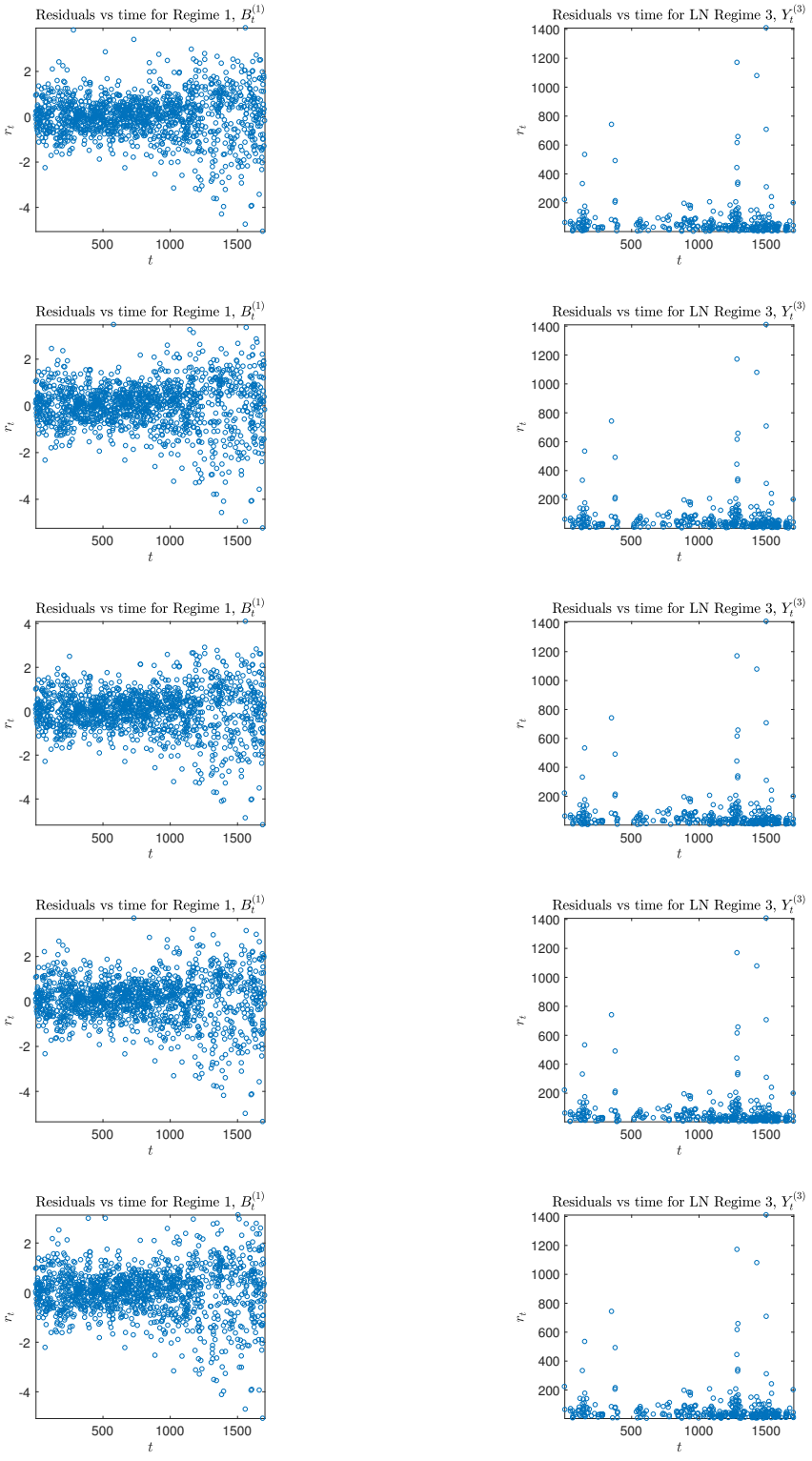


FIGURE B.2: A sample of five residuals versus time plots for each regime in Model 1 of Type III. (Left) Residuals versus time PPC plots for Regime 1, $B_t^{(1)}$, the AR(1) regime. (Right) Residuals versus time PPC plots for Regime 3, $Y_t^{(3)}$, the shifted-log-normal spike regime. The residuals of Regime 1 clearly increase over time which suggests our assumptions of time-homoscedasticity is violated.

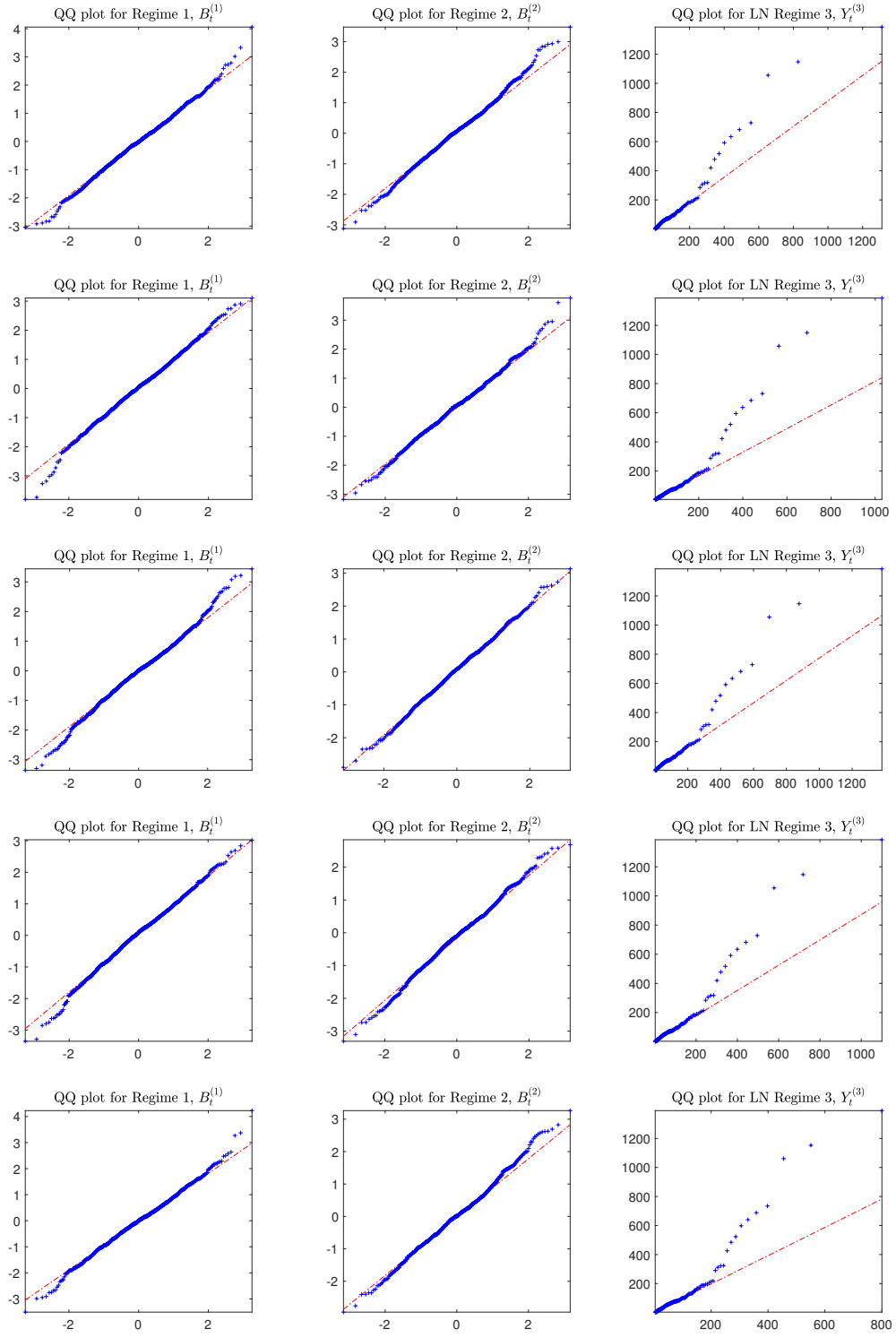


FIGURE B.3: A sample of five QQ-plot PPCs for the residuals of Regimes 1, 2 and 3 in Model 2 of Type III. (Left) QQ-plot PPCs for the first AR(1) base regime, $B_t^{(1)}$. (Middle) QQ-plot PPCs for the second AR(1) base regime, $B_t^{(2)}$. (Right) QQ-plot PPCs for the first shifted-log-normal spike regime, $Y_t^{(3)}$. The points in the QQ plots for the spike regime (right) do not lie on a straight line, suggesting the single shifted-log-normal distribution is unable to capture extreme observations. However, this violation may not be too significant in practice, and more work is needed to determine this. The QQ-plots for Regimes 1 and 2 suggest the assumptions about the AR(1) regimes are reasonable.

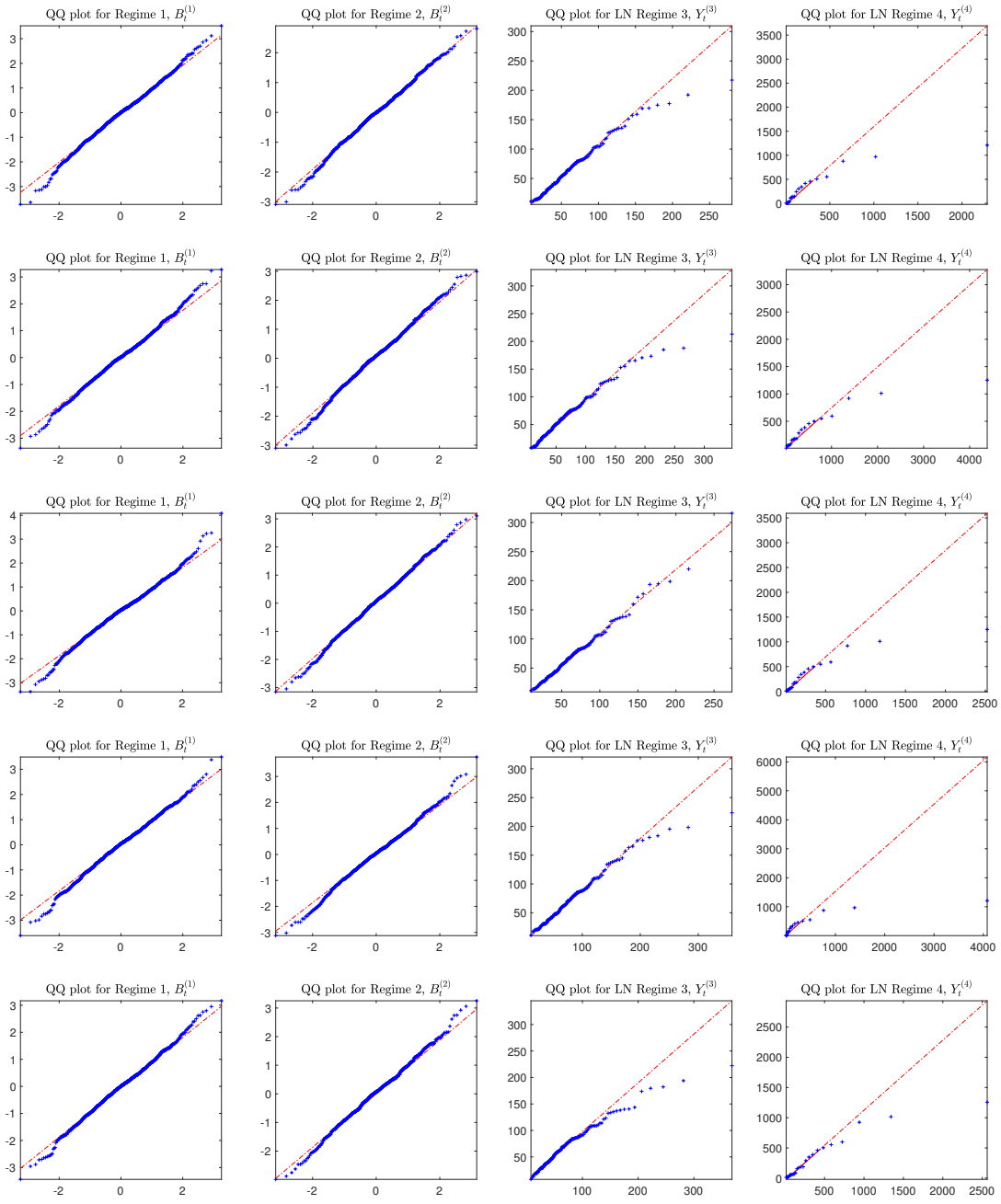


FIGURE B.4: A sample of five QQ-plot PPCs for each regime in Model 3 of Type III. (Left) QQ-plot PPCs for base regime 1, $B_t^{(1)}$, an AR(1) regime. (Center-left) QQ-plot PPCs for base Regime 2, $B_t^{(2)}$, another AR(1) regime. (Centre-right) QQ-plot PPCs for Regime 3, $Y_t^{(3)}$, a shifted-log-normal spike regime. (Right) QQ-plot PPCs for Regime 4, $Y_t^{(4)}$, a second shifted-log-normal spike regime for extreme spikes. The points in the QQ plots for the spike regimes (Regimes 3 and 4) stray slightly from a straight line, suggesting the shifted-log-normal distributions may not be suitable.

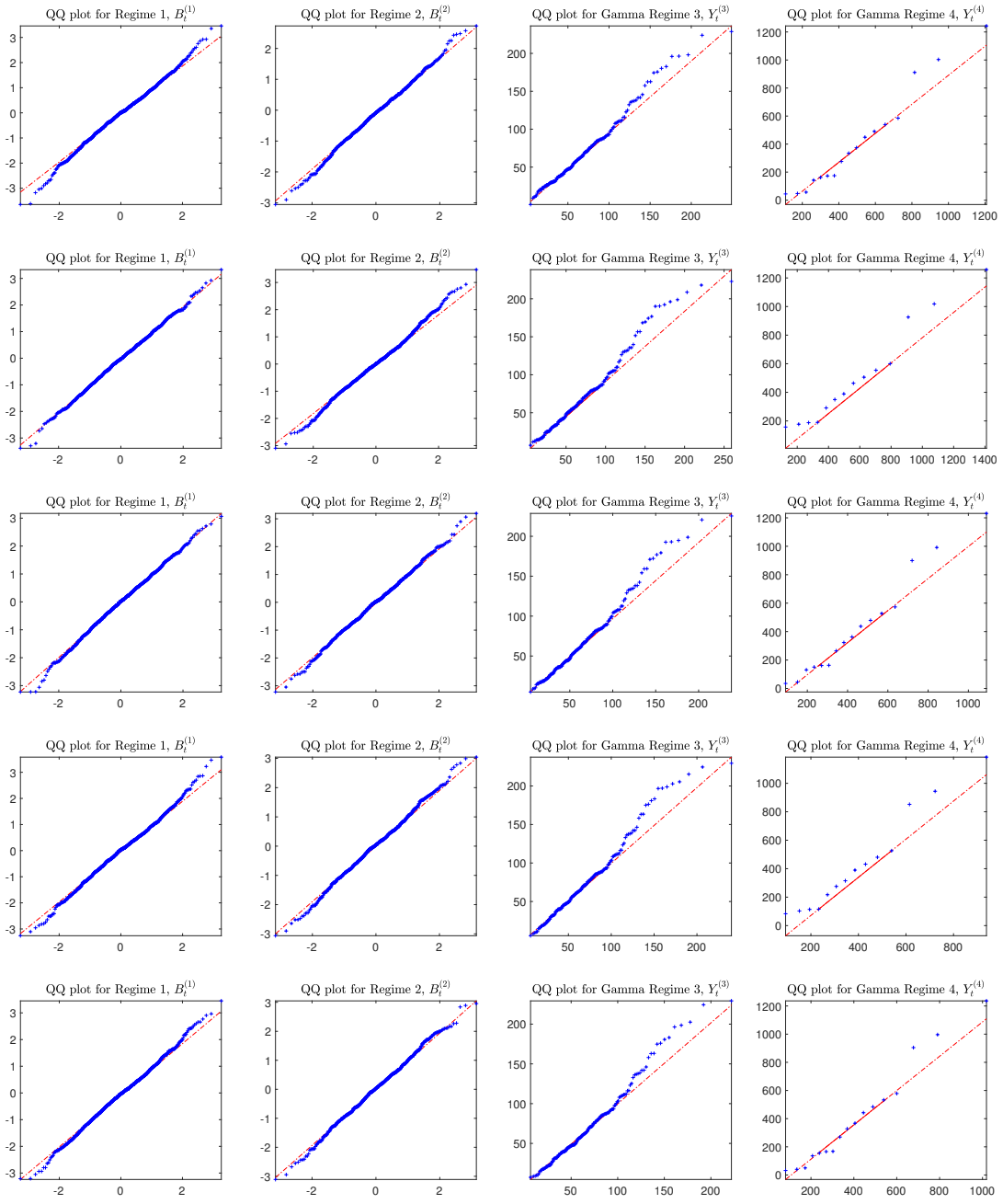
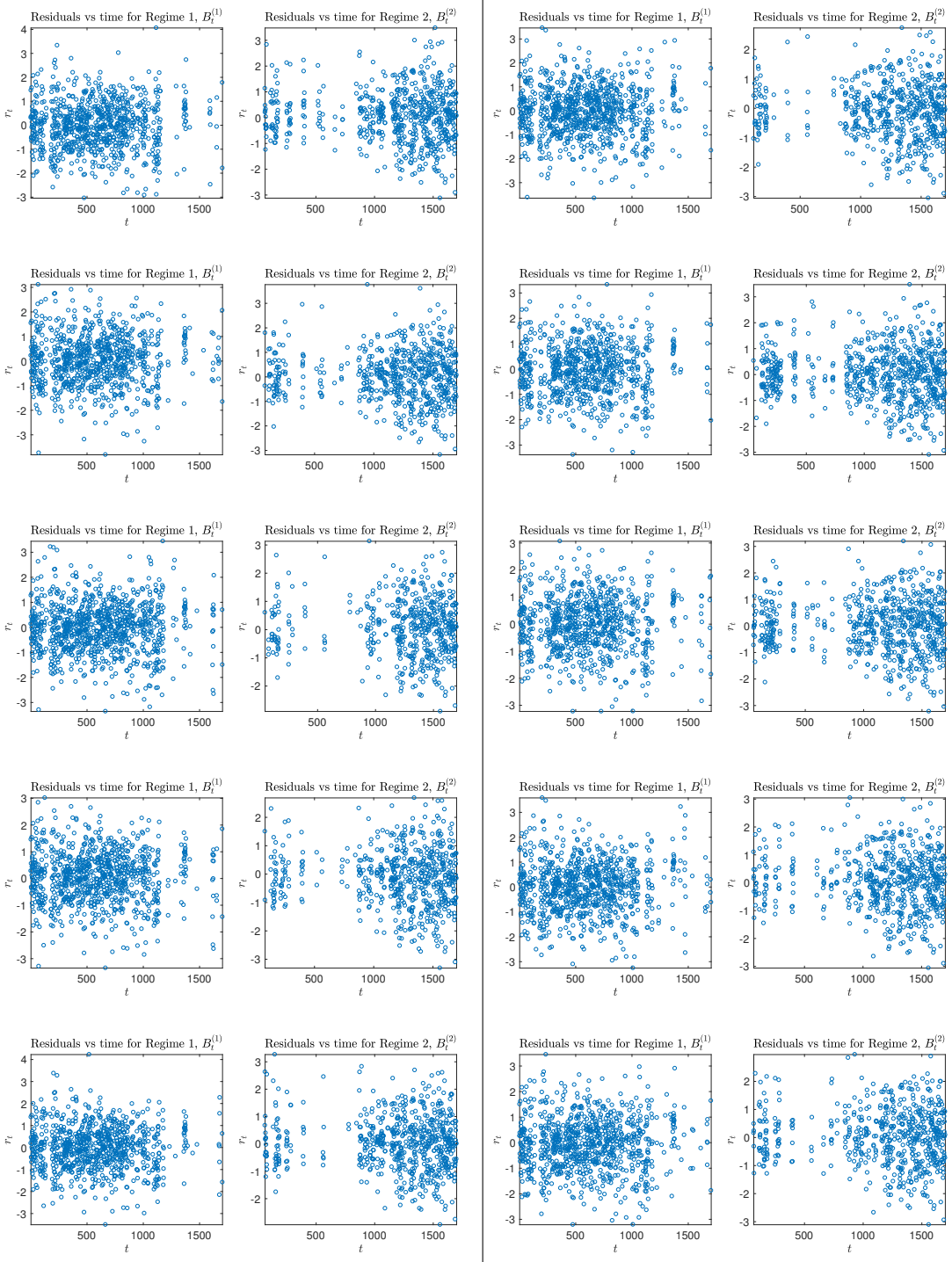


FIGURE B.5: A sample of five QQ-plot PPCs for each regime in Model 4 of Type III. (Left) QQ-plot PPCs for base regime 1, $B_t^{(1)}$, an AR(1) regime. (Center-left) QQ-plot PPCs for base regime 2, $B_t^{(2)}$, another AR(1) regime. (Centre-right) QQ-plot PPCs for Regime 3, $Y_t^{(3)}$, a shifted-Gamma spike regime. (Right) QQ-plot PPCs for Regime 4, $Y_t^{(4)}$, a second shifted-Gamma spike regime for extreme spikes. The QQ plots for Regimes 1, 2 and 4 suggest the distributional assumptions for these regimes are suitable. The QQ plots for Regime 3 suggests the Gamma distribution may not be suitable for this regime since the points stray from the reference line, however this evidence is not strong.



(A) Model 2 of Type III

(B) Model 4 of Type III

FIGURE B.6: A sample of five residuals versus time PPCs for each AR(1) regime in Models 2 (Left) and 4 (Right) of Type III. These PPCs show no obvious signs that the time-homoscedasticity assumption is violated for Regime 1. For both models, there are slight indications that the variance of Regime 2 may increase over time, although it is not clear how much of the apparent increase in variance is due to time-heteroscedasticity, or due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for the AR(1) regimes of both models.

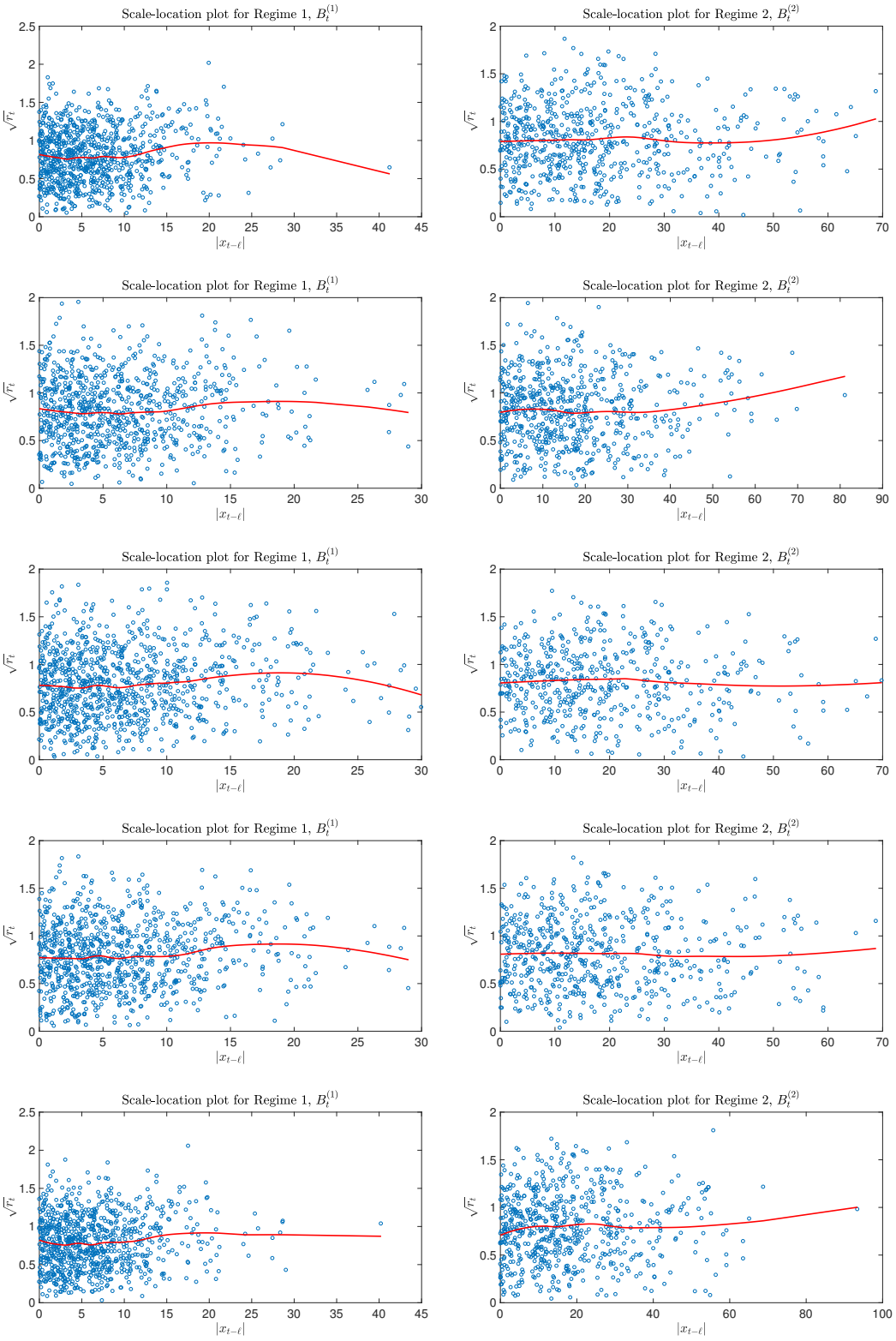


FIGURE B.7: A sample of five scale-location PPCs for each AR(1) regime in Model 2 of Type III. These PPCs show no obvious signs that self-dependent-homoscedasticity assumptions are violated for either regime.

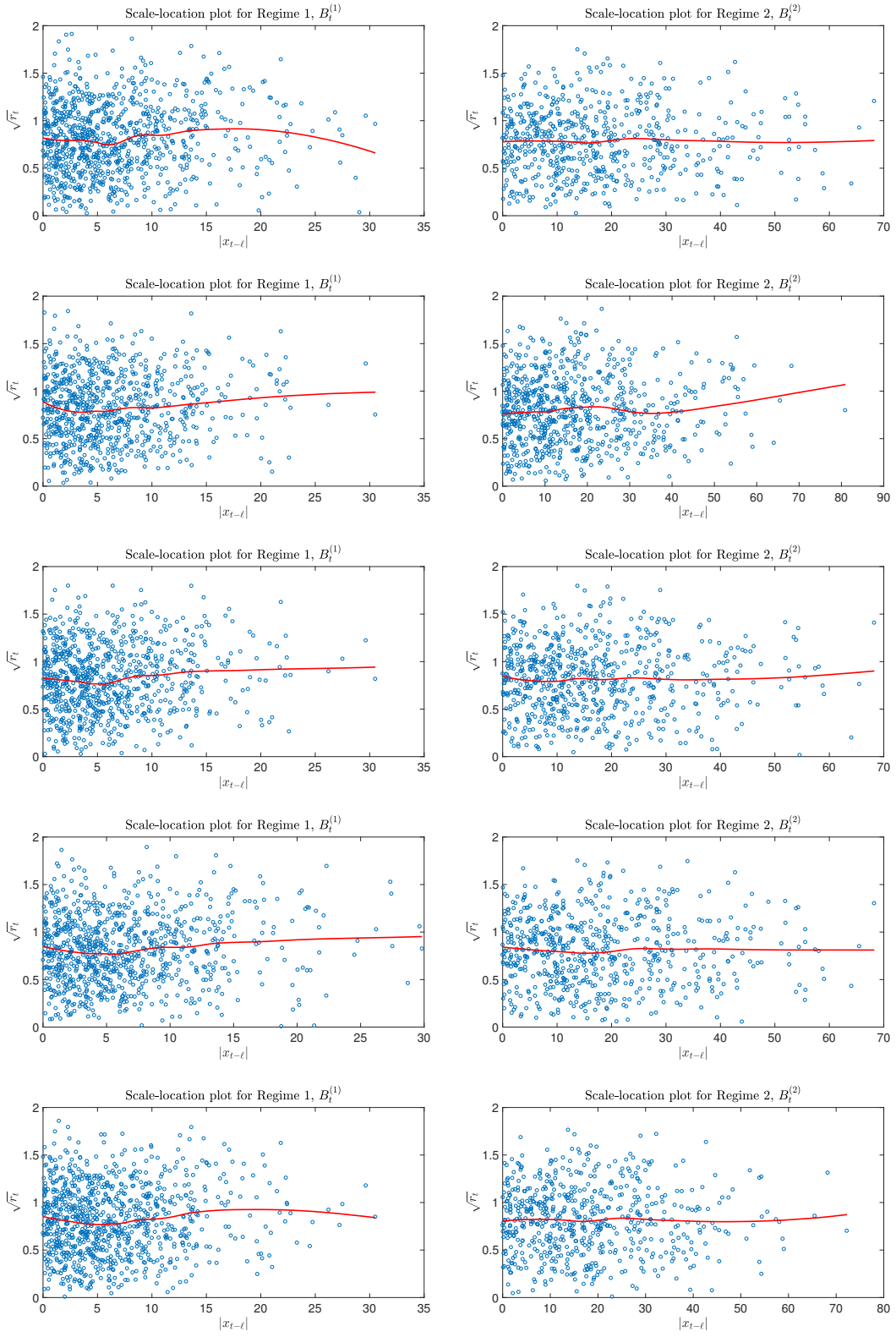
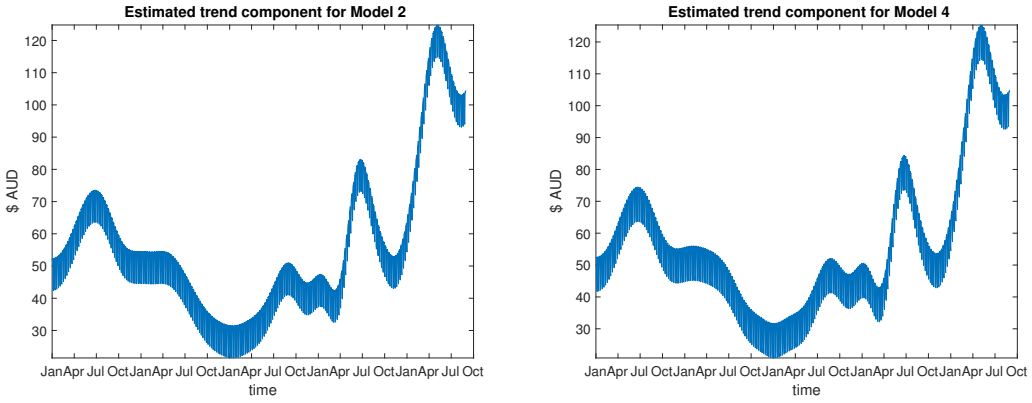


FIGURE B.8: A sample of five scale-location PPCs for each AR(1) regime in Model 4 of Type III. These PPCs show no obvious signs that self-dependent-homoscedasticity assumptions are violated for either regime.



(A) Model 2 of Type III

(B) Model 4 of Type III

FIGURE B.9: Estimated trend components of Models 2 (Left) and 4 (Right) of Type III.

B.2 Maximum likelihood for Type III models

We consider Models 2 and 4 of Type III, and use our EM methodology to fit these models to the data. We classify prices using the soft classification given by the EM algorithm and produce QQ-plots and residuals plots to check model assumptions, see Figures B.10-B.13. In Table B.3 we summarise our findings and in Table B.4 the MLEs for Models 2 and 4 are shown.

	Model 2	Model 4
	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{LN}(\mu_3, \sigma_3^2)$.	$X_t = \begin{cases} B_t^{(1)}, & \text{if } R_t = 1, \\ B_t^{(2)}, & \text{if } R_t = 2, \\ Y_t^{(3)}, & \text{if } R_t = 3, \\ Y_t^{(4)}, & \text{if } R_t = 4, \end{cases}$ $Y_t^{(3)} - q_3 \sim \text{Gamma}(\mu_3, \sigma_3^2)$, $Y_t^{(4)} - q_4 \sim \text{Gamma}(\mu_4, \sigma_4^2)$.
QQ-plots	Figure B.10: distributional assumptions are suitable for Regimes 1 and 2, but questionable for Regime 3.	Figure B.12: distributional assumptions are reasonable for all Regimes.
Residuals vs time	Figure B.11 (A) and (B): the time-homoscedasticity assumption is reasonable for Regime 1. There are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance is due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2.	Figure B.13 (A) and (B): the time-homoscedasticity assumption is reasonable for Regime 1. There are only slight indications that the time-homoscedasticity assumption may be unsuitable for Regime 2, and some of the apparent change in variance is due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2.
Scale-location	Figure B.11 (C) and (D): self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.	Figure B.13 (C) and (D): self-dependent-homoscedasticity assumptions are suitable for Regimes 1 and 2.

TABLE B.3: A summary of our likelihood-based analysis for Type III MRS models for the SA dataset.

(A) (B) (C)

FIGURE B.10: QQ-plots of residuals from each regime for Model 2 of Type III, estimated by maximum likelihood. The QQ-plots for the base regimes, (A) and (B), suggest the distributional assumptions of these regime are reasonable. The QQ-plot of the shifted-log-normal spike regime, (C), shows some deviation from linear which suggests the log-normal assumption may be unreasonable.

Parameter	Model 2	Model 4
α_1	-0.122	-0.64
ϕ_1	0.503	0.555
σ_1^2	44.8	48.0
α_2	0.0373	0.689
ϕ_2	0.421	0.419
σ_2^2	420	401
q_3	14	18.0
μ_3	3.85	2.50
σ_3^2	1.19	26.1
q_4	-	150
μ_4	-	2.50
σ_4^2	-	104
Transition matrix	$\begin{pmatrix} 0.919 & 0.006 & 0.075 \\ 0.000 & 0.913 & 0.087 \\ 0.245 & 0.153 & 0.602 \end{pmatrix}$	$\begin{pmatrix} 0.924 & 0.027 & 0.049 & 0.000 \\ 0.028 & 0.890 & 0.080 & 0.002 \\ 0.279 & 0.347 & 0.317 & 0.057 \\ 0.081 & 0.038 & 0.435 & 0.446 \end{pmatrix}$

TABLE B.4: MLEs of the parameter of Type III Models 2 and 4 for the SA dataset.

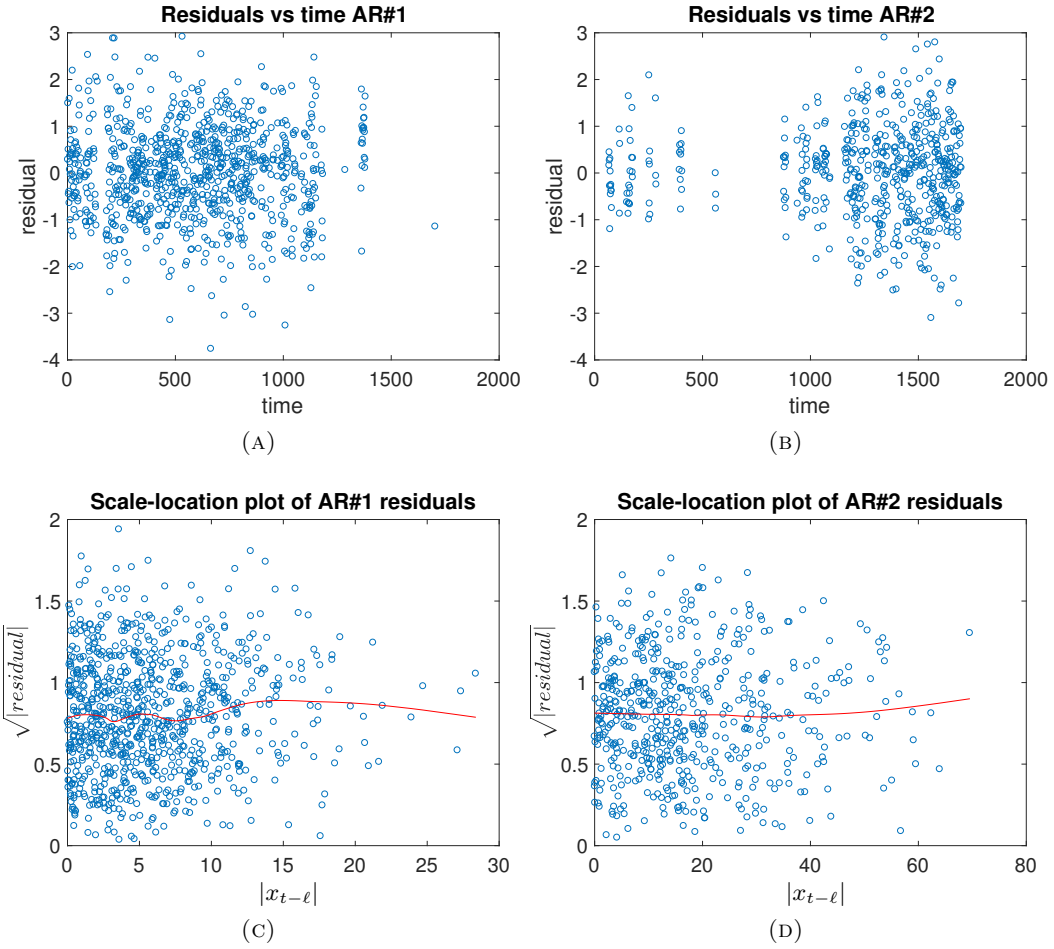
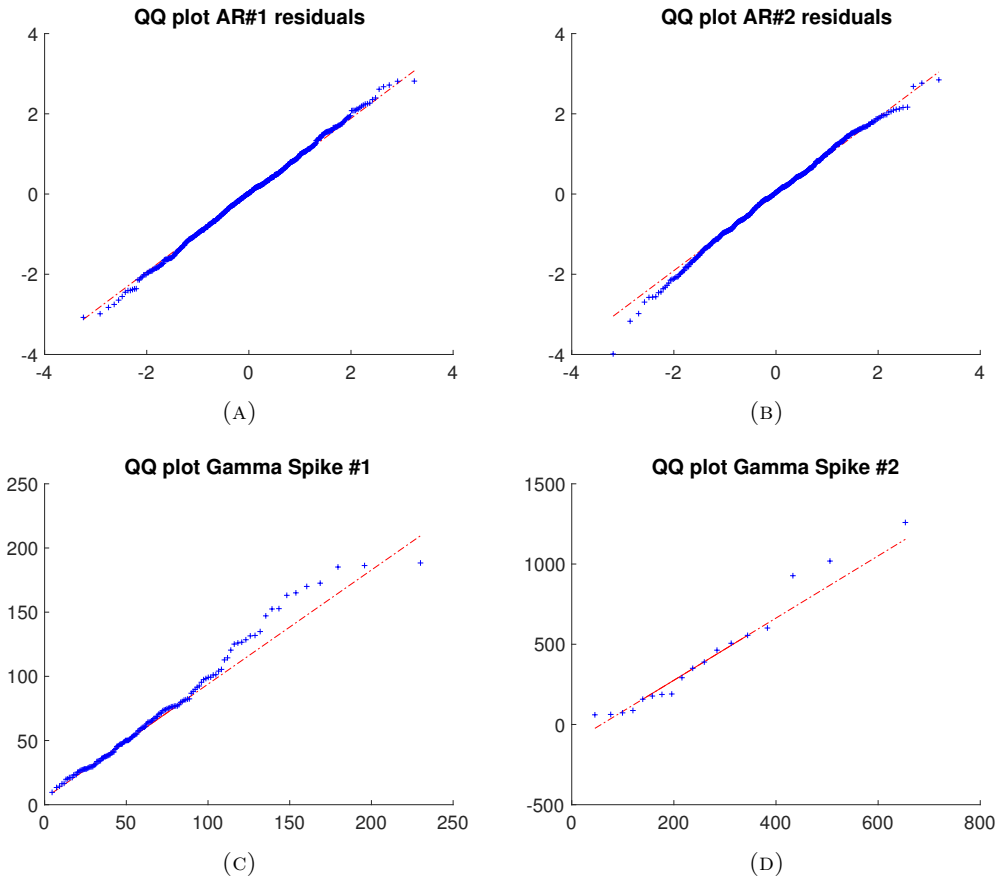


FIGURE B.11: Residuals plots for AR(1) regimes of Model 2 of Type III, estimated by maximum likelihood. Figures (A) and (B) plot the raw residual against time for Regimes 1 and 2 respectively. Figures (C) and (D) plot $\sqrt{|r_t|}$ against the absolute value of the last observed value from the same regime, before time t, $|x_{t-\ell}|$. Figure (A) suggests there is no issue with the time-homoscedasticity assumption for Regime 1. Figure (B) shows slight evidence that the variance of Regime 2 may increase over time, although it is not clear how much change in variation is due to time-heteroscedasicity, or due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2. Figures (C) and (D) suggest no obvious violation of the self-dependent-homoscedasticity assumptions.



(A) (B) (C) (D)

FIGURE B.12: QQ-plots of residuals for each regime of Model 4 of Type II, estimated by maximum likelihood. In plots (A), (B) and (D), the points lie in a relatively straight line, suggesting the distributional assumptions are reasonable for these regimes. In plot (C) there is slight some deviation from a straight line, suggesting the Gamma distribution assumption for Regime 3 may not be entirely appropriate.

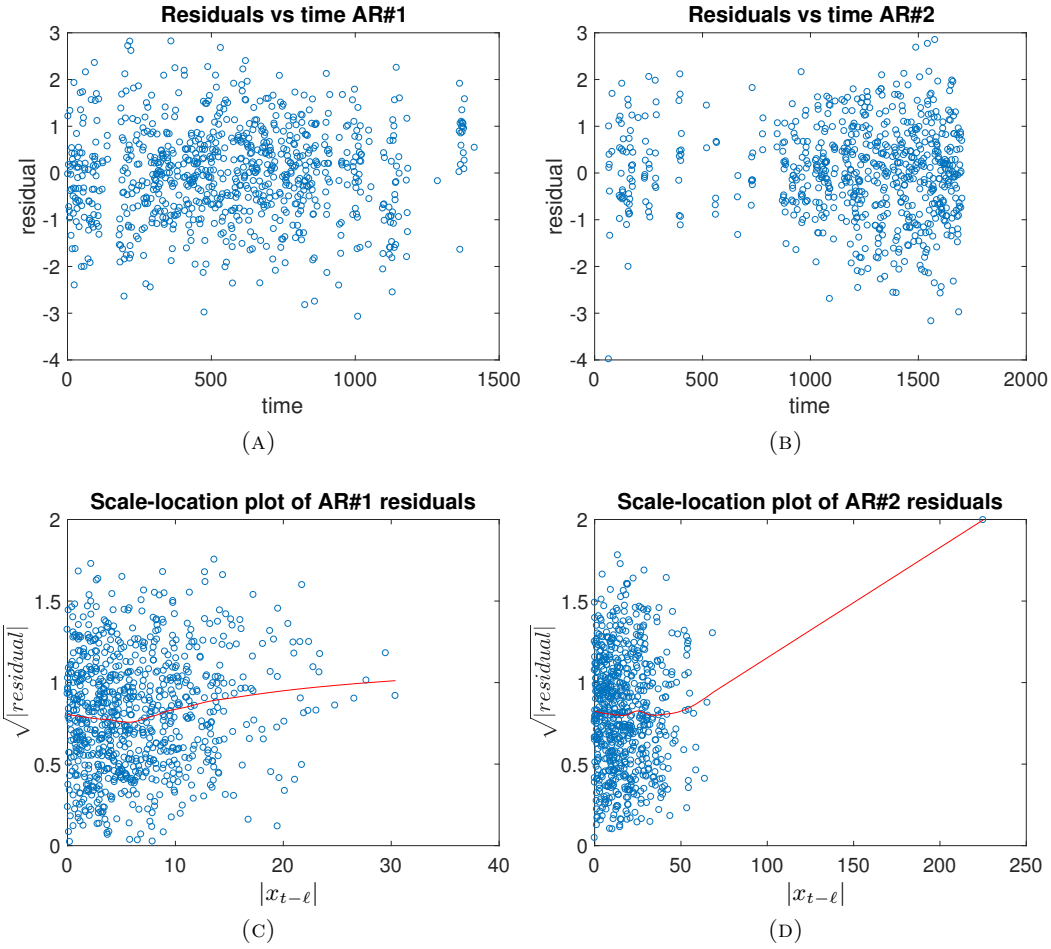


FIGURE B.13: Residuals plots for AR(1) regimes of Model 4 of Type II, estimated by maximum likelihood. Figures (A) and (B) plot the raw residual against time for regimes 1 and 2 respectively. Figures (C) and (D) plot $\sqrt{|r_t|}$ against the absolute value of the last observed value from the same regime, before time t, $|x_{t-\ell}|$. Figure (A) suggests there is no issue with the time-homoscedasticity assumption for Regime 1. Figure (B) shows slight evidence that the variance of Regime 2 may increase over time, although it is not clear how much change in variation is due to time-heteroscedasicity, or due to fewer observations at earlier times. We conclude that the time-homoscedasticity assumption is reasonable for Regime 2. Figures (C) and (D) show no obvious evidence that the variance of the base regimes increases as a function of lagged values.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer New York, New York, NY, 1998.

[2] P. Albert. A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 47(4):13711381, December 1991.

[3] E. Alvaro, I. P. J., and V. Pablo. Modelling electricity prices: International evidence*. *Oxford Bulletin of Economics and Statistics*, 73(5):622–650, 2002.

[4] Australian Energy Market Operator. National Electricity Market fact sheet. Technical report, AEMO, 2016. Accessed: 2017-08-09.

[5] Australian Energy Market Operator. AEMO annual report. Annual report, AEMO, 2017. Accessed: 2018-05-07.

[6] Australian Energy Market Operator. Five-minute settlement: high level design. Technical report, AEMO, September 2017. Accessed: 2018-07-20.

[7] Australian Energy Market Operator. South Australian electricity report. Technical report, AEMO, 2017. Accessed: 2018-02-07.

[8] Australian Energy Market Operator. Data dashboard. https://www.aemo.com.au/Electricity/National-Electricity-Market-NEM/Data-dashboard#aggregated-data, 2018. Accessed: 2018-02-17.

[9] L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363, 05 1967.

[10] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966.

[11] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41(1):164–171, 02 1970.

[12] R. Becker, A. Clements, and W. Zainudin. Modeling electricity price events as point processes. *The Journal of Energy Markets*, 6, 7 2013.

[13] R. Becker, S. Hurn, and V. Pavlov. Modelling spikes in electricity prices. *Economic Record*, 83(263):371–382, 2007.

[14] F. E. Benth, J. S. Benth, and S. Koekebakker. *Stochastic modelling of electricity and related markets*. Advanced series on statistical science and applied probability. World Scientific, Singapore ; Hackensack, N.J., 2008.

[15] F. E. Benth, J. Kallsen, and T. Meyer-Brandis. A non-Gaussian Ornstein-Uhlenbeck process for electricity spot price modeling and derivatives pricing. *Applied Mathematical Finance*, 14(2):153–169, 2007.

[16] F. E. Benth, R. Kiesel, and A. Nazarova. A critical empirical study of three electricity spot price models. *Energy Economics*, 34(5):1589–1616, 2012.

[17] M. Bierbrauer, C. Menn, S. T. Rachev, and S. Trück. Spot and derivative pricing in the EEX power market. *Journal of Banking & Finance*, 31(11):3462–3485, 2007. Risk Management and Quantitative Approaches in Finance.

[18] M. Bierbrauer, S. Trück, and R. Weron. Modeling electricity prices with regime switching models. In M. Bubak, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, editors, *Computational Science - ICCS 2004*, pages 859–867, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[19] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

[20] S. P. Brooks and G. O. Roberts. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8:319–335, 1997.

[21] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[22] D. Chen and D. Bunn. The forecasting performance of a finite mixture regime-switching model for daily electricity prices. *Journal of Forecasting*, 33(5):364–375, 2014.

[23] X. Chen, Z. Y. Dong, K. Meng, Y. Xu, K. P. Wong, and H. W. Ngan. Electricity price forecasting with extreme learning machine and bootstrapping. *IEEE Transactions on Power Systems*, 27(4):2055–2062, Nov 2012.

[24] T. Christensen, A. Hurn, and K. Lindsay. Forecasting spikes in electricity prices. *International Journal of Forecasting*, 28(2):400 – 411, 2012.

[25] T. Christensen, S. Hurn, and K. Lindsay. It never rains but it pours: Modeling the persistence of spikes in electricity prices. *The Energy Journal*, Volume 30(1):25–48, 2009.

[26] A. Clements, R. Herrera, and A. Hurn. Modelling interregional links in electricity price spikes. *Energy Economics*, 51(C):383–393, 2015.

[27] C. de Jong. The nature of power spikes: A regime-switch approach. *Studies in Nonlinear Dynamics & Econometrics*, 10:1361–1361, 02 2007.

[28] C. de Jong and R. Huisman. Option formulas for mean-reverting power prices with spikes. ERIM Report Series Research in Management ERS-2002-96-F&A, Erasmus Research Institute of Management (ERIM), Oct. 2002.

[29] 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.

[30] S. Deng. Stochastic models of energy commodity prices and their applications: Mean-reversion with jumps and spikes. Working Paper PWP-073, University of California Energy Institute, 2000.

[31] M. Dungey, A. Ghahremanlou, and N. V. Long. Strategic bidding of electric power generating companies: Evidence from the Australian National Energy Market. CESifo Working Paper Series 6819, CESifo Group Munich, 2017.

[32] M. Eichler, O. Grothe, H. Manner, and D. Tuerk. Models for short-term forecasting of spike occurrences in Australian electricity markets: a comparative study. *Journal of Energy Markets*, 7:55 – 81, 2014.

[33] C. Erlwein, F. E. Benth, and R. Mamon. HMM filtering and parameter estimation of an electricity spot price model. *Energy Economics*, 32(5):1034–1043, 2010.

[34] R. G. Ethier and T. D. Mount. Estimating the volatility of spot prices in restructured electricity markets and the implications for option values. PSerc Working Paper, Cornell University, 1998.

[35] A. Eydeland and K. Wolyniec. *Energy and Power Risk Management: New Developments in Modeling, Pricing, and Hedging.* Wiley Finance Series. Wiley, 2003.

[36] C. Francq and M. Roussignol. Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32(2):151–173, 1998.

[37] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2nd edition, July 2003.

[38] H. Geman and A. Roncoroni. Understanding the fine structure of electricity prices. *The Journal of Business*, 79(3):1225–1262, 2006.

[39] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov 1984.

[40] J. K. Ghosh, M. Delampady, and T. Samanta. *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media, 2007.

[41] G. Glonek. Personal communication, 2017.

[42] J. Gonzalez, J. Moriarty, and J. Palczewski. Bayesian calibration and number of jump components in electricity spot price models. *Energy Economics*, 65(Supplement C):375–388, 2017.

[43] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.

[44] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[45] J. D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1):39–70, 1990.

[46] W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[47] J. S. Henneke, S. T. Rachev, F. J. Fabozzi, and M. Nikolov. MCMC-based estimation of Markov switching ARMA-GARCH models. *Applied Economics*, 43(3):259–271, 2011.

[48] R. Herrera and N. Gonzlez. The modeling and forecasting of extreme events in electricity spot markets. *International Journal of Forecasting*, 30(3):477 – 490, 2014.

[49] H. Higgs and A. Worthington. Stochastic price modeling of high volatility, mean-reverting, spike-prone commodities: The Australian wholesale spot electricity market. *Energy Economics*, 30(6):3172 – 3185, 2008. Technological Change and the Environment.

[50] B. M. Hill. The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. *Journal of the American Statistical Association*, 58(301):72–84, 1963.

[51] R. Huisman. The influence of temperature on spike probability in day-ahead power prices. *Energy Economics*, 30(5):2697 – 2704, 2008.

[52] R. Huisman and C. de Jong. Option pricing for power prices with spikes. *Energy Power Risk Management*, 7(11):12–16, 2003.

[53] R. Huisman and R. Mahieu. Regime jumps in electricity prices. *Energy economics*, 25(5):425–434, 2003.

[54] A. S. Hurn, A. Silvennoinen, and T. Teräsvirta. A smooth transition logit model of the effects of deregulation in the electricity market. *Journal of Applied Econometrics*, 31(4):707–733, 2016.

[55] M. Hürzeler and H. R. Künsch. Approximating and maximising the likelihood for a general state-space model. In *Sequential Monte Carlo Methods in Practice*, pages 159–175. Springer New York, New York, NY, 2001.

[56] D. J. Swider and C. Weber. Extended ARMA models for estimating price developments on day-ahead electricity markets. *Electric Power Systems Research*, 77:583–593, 04 2007.

[57] J. Janczura, S. Trück, R. Weron, and R. C. Wolff. Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Economics*, 38:96–110, 2013.

[58] J. Janczura and R. Weron. Regime-switching models for electricity spot prices: Introducing heteroskedastic base regime dynamics and shifted spike distributions. In *2009 6th International Conference on the European Energy Market*, pages 1–6, May 2009.

[59] J. Janczura and R. Weron. An empirical comparison of alternate regime-switching models for electricity spot prices. *Energy Economics*, 32(5):1059–1073, 2010.

[60] J. Janczura and R. Weron. Efficient estimation of Markov regime-switching models: An application to electricity spot prices. *Advances in Statistical Analysis*, 96(3):385–407, 2012.

[61] J. Janczura and R. Weron. Inference for Markov-regime switching models of electricity spot prices. HSC Research Reports HSC/12/01, Hugo Steinhaus Center, Wroclaw University of Technology, 2012.

[62] J. Janczura and R. Weron. Goodness-of-fit testing for the marginal distribution of regime-switching models with an application to electricity spot prices. *Advances in Statistical Analysis*, 97(3):239–270, 2013.

[63] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461, 1946.

[64] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*. New York Wiley, 2nd edition, 1994.

[65] T. Kanamura and K. Ōhashi. On transition probabilities of regime switching in electricity prices. *Energy Economics*, 30(3):1158 – 1172, 2008.

[66] N. V. Karakatsani and D. W. Bunn. Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. *International Journal of Forecasting*, 24(4):764 – 785, 2008. Energy Forecasting.

[67] C.-J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2):1–22, January-February 1994.

[68] O. Knapik and P. Exterkate. A regime-switching stochastic volatility model for forecasting electricity prices. Working Papers 2017-02, University of Sydney, School of Economics, 2017.

[69] B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127 – 143, 1992.

[70] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, April 1983.

[71] F. Lisi and F. Nan. Component estimation for electricity prices: Procedures and comparisons. *Energy Economics*, 44:143–159, 07 2014.

[72] T. Mark and K. George. Modeling long-term persistence in hydroclimatic time series using a hidden state Markov model. *Water Resources Research*, 36(11):3301–3310, 2000.

[73] X.-L. Meng and D. B. Rubin. On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and its Applications*, 199:413 – 425, 1994. Special Issue Honoring Ingram Olkin.

[74] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[75] S. Meyn, R. L. Tweedie, and P. W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2009.

[76] S. Mitra and P. Date. Regime switching volatility calibration by the Baum–Welch method. *Journal of Computational and Applied Mathematics*, 234(12):3243 – 3260, 2010.

[77] T. D. Mount, Y. Ning, and X. Cai. Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics*, 28(1):62–80, 2006.

[78] W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier, 1994.

[79] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2):175–240, 1928.

[80] V. Norén. *Modelling Power Spikes with Inhomogeneous Markov-switching Models*. Master's theses in mathematical sciences. Lund University, 2013.

[81] J. Nowotarski, J. Tomczyk, and R. Weron. Modeling and forecasting of the long-term seasonal component of the EEX and Nord Pool spot prices. In *2013 10th International Conference on the European Energy Market (EEM)*, pages 1–8, May 2013.

[82] J. Nowotarski, J. Tomczyk, and R. Weron. Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. *Energy Economics*, 39:13 – 27, 2013.

[83] J. Nowotarski and R. Weron. On the importance of the long-term seasonal component in day-ahead electricity price forecasting. *Energy Economics*, 57, 06 2016.

[84] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, volume 67. Springer-Verlag, 01 2002.

[85] A. Panagiotelis and M. Smith. Bayesian density forecasting of intraday electricity prices using multivariate skew-t-distributions. *International Journal of Forecasting*, 24(4):710–727, 2008.

[86] C. Pape, A. Vogler, O. Woll, and C. Weber. Forecasting the distributions of hourly electricity spot prices. EWL Working Papers 1705, University of Duisburg-Essen, Chair for Management Science and Energy Economics, May 2017.

[87] F. Regland and E. Lindström. Independent spike models: Estimation and validation. *Finance a Uver: Czech Journal of Economics & Finance*, 62(2):180–196, 2012.

[88] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer texts in statistics. Springer, New York, 2nd. ed. edition, 2004.

[89] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[90] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[91] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[92] P. Robinson. Estimation of a time series model from unequally spaced data. *Stochastic Processes and their Applications*, 6(1):9 – 24, 1977.

[93] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172, 12 1984.

[94] M. Shahidehpour, H. Yamin, and Z. Li. *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management*. John Wiley & Sons, 1st edition, 2002.

[95] M. J. Stevenson, L. F. M. do Amaral, and M. Peat. Risk management and the role of spot price predictions in the Australian retail electricity market. *Studies in Nonlinear Dynamics & Econometrics*, 10(3), 2006.

[96] S. Trueck, R. Weron, and R. Wolff. Outlier treatment and robust approaches for modeling electricity spot prices. *Paper presented at the 56th Session of the International Statistical Institute, Invited Paper Meeting IPM71 Statistics of risk aversion, Lisbon, Aug. 22-29, 2007.*, 09 2007.

[97] C. Valens. A really friendly guide to wavelets. Technical report, University of New Mexico, 1999. Accessed: 2018-05-29.

[98] M. Ventosa, A. Baillo, A. Ramos, and M. Rivier. Electricity market modeling trends. *Energy Policy*, 33(7):897–913, 2005.

[99] Z. Walter and G. Peter. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

[100] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[101] R. Weron. Heavy-tails and regime-switching in electricity prices. *Mathematical Methods of Operations Research*, 69(3):457–473, 2009.

[102] R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014.

[103] R. Weron, M. Bierbrauer, and S. Trück. Modeling electricity prices: jump diffusion and regime switching. HSC Research Reports HSC/03/01, Hugo Steinhaus Center, Wroclaw University of Technology, 2003.

[104] R. Weron, A. Misiorek, et al. Forecasting spot electricity prices with time series models. In *Proceedings of the European Electricity Market EEM-05 Conference*, pages 133–141, 2005.

[105] C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[106] S.-Z. Yu. *Hidden Semi-Markov Models*. Elsevier, Boston, 1st edition, 2016.