



COPYRIGHT AND USE OF THIS THESIS

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author's moral rights if you:

- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author's reputation

For further information contact the University's Copyright Service.

sydney.edu.au/copyright

Multi-regime models involving Markov chains

Matthew Fitzpatrick

March 2016

A thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy

School of Mathematics and Statistics

University of Sydney

Contents

1	Introduction	5
2	Geometric ergodicity of the Gibbs sampler for the Poisson change-point model	13
2.1	Introduction	14
2.2	Model specification	17
2.3	Estimation of the model parameters	19
2.4	Geometric ergodicity of the Gibbs sampler	21
2.5	Applications to Victorian driver fatality count data	27
2.6	Conclusions	30
3	Efficient Bayesian estimation of the multivariate double chain Markov model	31
3.1	Introduction	32
3.2	Model specification	37
3.3	Estimation of the model parameters	40
3.3.1	Sampling from $P(\mathbf{x} \mathbf{y}, \theta)$	41
3.3.2	Sampling from $P(\theta \mathbf{x}, \mathbf{y})$	44
3.3.3	Extra permutation step	45
3.3.4	Post-processing algorithm	48
3.4	Simulation studies	49
3.4.1	Parameters Estimation	49
3.4.2	Posterior Density Estimation	53
3.4.3	Comparison to MARCH 3.0 software	56
3.5	Applications to Standard and Poor's credit rating data	58
3.6	Discussion	65
3.6.1	Conclusions	65

3.6.2	Further Research	67
4	Mixtures of Markov chains	68
4.1	The continuous-time Markov chain	69
4.2	General finite mixtures of continuous-time Markov chains	76
4.3	Testing between 1 and 2 mixture components	81
4.3.1	A parametric bootstrap procedure to test for the presence of a mixture	83
4.3.2	Non-identifiability of the likelihood ratio test	86
4.3.3	Divergence of the log-likelihood ratio test statistic	90
4.3.4	A special case with 2 states	99
5	Censored exponential mixture detection	102
5.1	An overview of the testing problem	104
5.2	Testing homogeneity in censored exponential mixture models	106
5.3	Details of original proofs	116
5.3.1	Proof of Lemma 5.1	122
5.3.2	Proof of Lemma 5.2	124
5.3.3	Proof of Theorem 5.3	128
5.3.4	Proof of Lemma 5.4	130
5.3.5	Proof of Lemma 5.5	132
5.3.6	Proof of Theorem 5.6	134
6	Conclusion	135
A	Covariance calculation for stochastic integrals	138

Dr. Dobrin Marchev
Department of Mathematics
New York City College of Technology
300 Jay Street, Brooklyn NY 11201
September 22, 2015

Faculty of Science
University of Sydney
Level 2, Carlaw Building, Eastern Avenue
Camperdown, NSW 2050, Australia

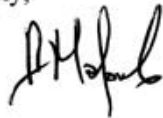
Dear Faculty of Science:

I refer to the thesis submitted by Matthew Fitzpatrick titled “Multi-regime models involving Markov chains” and in particular, Chapter 3 (the chapter) – “Efficient Bayesian estimation of the multivariate double chain Markov model”.

The chapter is largely identical to a publication, co-authored by Matthew Fitzpatrick and myself, titled “Efficient Bayesian estimation of the multivariate double chain Markov model” in the journal *Statistics and Computing*, (2013) Volume 23, pages 467-480. The research and construction of this publication was conducted during the PhD. candidature of Matthew Fitzpatrick as a key component to the thesis with the overarching theme of multi-regime models involving Markov chains, where I acted in the capacity of his Associate Supervisor.

Matthew is responsible for the majority of the research and written work in Chapter 3 including the discovery of the motivating application and model design, under a standard supervisory arrangement where we collaboratively discussed ideas and shared references to solve the problems posed. The ideas for subsection 3.3.3 “Extra permutation step” was largely a contribution by myself as an application to this problem of some of my previous research work. The post-processing algorithm that follows was suggested to Matthew by myself and he implemented it in the theoretical work. Matthew programmed the algorithm in R for the simulation studies.

Sincerely,



Dr. Dobrin Marchev

1 Introduction

In this work, we explore the theory and applications of various multi-regime models involving Markov chains. Markov chains are an elegant way to model path-dependent data. We study a series of problems with non-homogeneous data and the various ways that Markov chains come into play. Non-homogeneous data can be modelled using multi-regime models, which apply a distinct set of parameters to distinct population sub-groups, referred to as *regimes*. Such models essentially allow for a practitioner to understand the nature (and in some cases the existence) of particular regimes within the data without the need to split the population into assumed sub-groups. Examples of problems involving non-homogeneous data include the problem of modelling business outcomes in different economic states (without explicitly using economic variables) or studying rainfall patterns as the seasons change across geographies. The problems we discuss here involve multiple regimes in two different ways and they also involve Markov chains in two different ways. Different regimes can apply to an entire population at different times, which we see in our first two problems, and different regimes can also apply to different subsections of the population over the whole observed time, which we see in our second two problems. Markov chains are involved via the estimation procedure or within models for the observed data. We first study multi-regime problems with Markov chains used in the estimation procedure. These are conducted from a Bayesian approach and we utilise the properties of Markov chains to discover and establish efficiencies in the estimation algorithms. Following this, we explore the uses of Markov chains as components of models applied to non-homogeneous data. Note that our second problem involves Markov chains in both the estimation procedure as well as the model. Although this work is largely focussed on addressing the theoretical issues of each problem, the motivation behind each of the problems studied comes from real datasets, which possess levels of complexity

that are insufficiently described through more standard procedures.

Our first problem is motivated by a simple form of non-homogeneous data. We study a single discrete time series representing quarterly driver fatality counts for the state of Victoria, Australia. Upon inspection of the data, it is clear that there are shifts in the levels of the counts over different time periods. Thus, there is a need to model the non-homogeneous dataset, allowing for multiple regimes. We apply a Bayesian Poisson change-point model to the data, using a Gibbs sampler, and note that there is no way of knowing how many iterations of the sampler will be required for a sufficient level of convergence. We derive a key property of the Markov chain involved in the Gibbs sampler procedure to estimate the parameters of a Poisson change-point model, which provides a significant insight into the nature of the convergence rate of the sampler. This enables us to have greater confidence around the model estimates and the resulting insights gained on the phenomena driving the multiple regimes in the data.

We continue with the use of Bayesian estimation algorithms for our second problem, which is motivated by the regime-switching nature of credit rating migration dynamics for a homogeneous population of firms. This is a problem with more complex discrete time-series data, with multiple series of different lengths. This dataset is modelled using the double chain Markov model (DCMM), where we have a hidden Markov chain that drives the switching process between two Markov chains that drive the observed data. Similar to the first problem, we also estimate the model using a Markov chain Monte Carlo procedure and show how it can be applied to model credit rating migration data over discrete time and identify where the key *regime switches* occur, which aligns remarkably well with notable economic events of the past few decades in the United States. We exploit the properties of the Markov chain underlying the estimation procedure to enhance the efficiency of the sampling algorithm. We show using simulation studies

that we are able to improve the estimation efficiency, when compared to existing estimation procedures.

The application of credit rating migration modelling is also the motivation for our third problem. However, instead of supposing that the different regimes occur over time, we look at different regimes that drive a particular proportion of the population over the whole of the finite observation window. We are thus looking at a Markov chain mixture model and focus on the problem of testing for the number of mixture components. We prove that the log-likelihood ratio test statistic, for the test between 1 and 2 Markov chain components, diverges to infinity with probability 1.

We then outline a simplified version of the model, where we only have 2 possible states for each Markov chain component, one for *non-default* and an absorbing state for *default*, and state a theorem that gives the exact limiting distribution of the log-likelihood ratio test statistic for this version of the problem. This test is equivalent to the test between 1 and 2 components in a mixture of censored exponentials. We ultimately find the exact limiting distribution of the log-likelihood ratio test statistic for this challenging problem, which would allow us to test for the presence of a mixture for this class of models.

Our first problem is explored in Chapter 2, where we apply the Poisson change-point model to driver fatality counts for the state of Victoria, Australia. The different regimes arise from evolving policy settings with some causing the fatalities to drop significantly. We fit this model with a Bayesian approach using the Gibbs sampler, a commonly used Markov chain Monte Carlo procedure. Our sampler starts with initial parameter estimates that are sampled from their respective prior distributions, which are used to sample a subset of the parameters from the conditional distributions that arise from knowing the complementary subset, then conditioning on these new samples to re-sample the initial subset. This iterative

procedure continues until the resulting samples of each parameter have distributions that resemble their true marginal distributions. If these sample distributions are in a steady state and have low values of the autocorrelation function at each lag $k \geq 1$ with respect to the index of the sampled chain of estimates, then we say that the algorithm has converged. Note here that the conditional distribution of future samples, conditional on the current and past samples, is only dependent on the current sample and not the samples preceding it. This is the Markov property of the Gibbs sampler. The chain is the series of samples for the full parameter vector and the state space of the chain is the corresponding combined parameter space of the model. In order to generate appropriate parameter estimates (and distributions around each), we require that the Markov chain of the Gibbs sampler is able to explore all possibilities in the parameter space. That is, we require that the Markov chain be ergodic. If there was an absorbing state, for example, the Markov chain would not be ergodic. This could mean that a particular Gibbs sampler may eventually sample the value that results in the absorbing state and all subsequent samples for that parameter would be the same. This would not allow the sampler to explore all areas of the parameter space but only a small section of it. Sometimes a Markov chain can be ergodic but the chance of an arbitrary chain exploring a particular part of the distribution is so low that it is barely sampled from, even after many iterations of the algorithm. In a practical setting, we require algorithms to be fast and thus need to know the rate at which the Markov chain in the Gibbs sampler has explored all areas of the parameter space sufficiently. We utilise some key results in the literature to show that if a particular sampler has certain properties, we can show that the Markov chain in the sampler is geometrically ergodic. That is, it explores all areas of the parameter's sample space at a geometric rate, meaning that only a moderate amount of iterations are necessary to have a sufficiently rich sample from the parameter space to fit the model.

Key results of this chapter have been published in Fitzpatrick (2014), which was produced as a key component of this thesis with the overarching theme of multi-regime models involving Markov chains. Here, our observations derive from underlying regimes that change over time and we estimate the parameters using a procedure involving a Markov chain. Our key result is on finding a particular property of this Markov chain, geometric ergodicity, which has important implications for our estimation procedure and hence the reliability of our results.

In Chapter 3, as well as using Markov chain Monte Carlo for estimation, we explore a model that uses Markov chains to describe the data dynamics directly. We are modelling the credit rating dynamics of hundreds of financial services firms in the United States of America across a time period that spans many different economic states. We note that the rating dynamics vary widely enough to warrant a multi-regime model. Since the broader economy is often described as a cycle, with growth and contraction periods, we choose to fit two regimes and also model the switching process between these regimes with a Markov chain. This is known as the double chain Markov model (DCMM). The observed data is driven by a Markov chain at each time point; however, the particular Markov chain that drives the data is selected by a hidden Markov chain, which models the switching dynamics. We estimate all of the parameters with an efficient Bayesian algorithm to ensure that all areas of the parameter space are sufficiently explored to allow for effective convergence of the Gibbs sampler as in Chapter 2. After fitting the model to the credit rating data, we find that not only do the two regimes clearly represent *good* and *bad* credit migration dynamics but they are selected for the time periods that are well known to be the *good* and *bad* times of the United States economy. This is a remarkable finding, given that only the credit rating migration dataset was used with no economic information used *a priori*. It has always been a challenge for practitioners to model business dynamics, particularly

when it comes to rare events such as defaults of highly rated firms. The double chain Markov model allows for a few parameters to describe complex dynamics that can assist in understanding the credit risk taken by banks and large investment firms. When we allow for multiple regimes, we are able to estimate the dynamics that occur during times of economic stress. We know from the recent financial crisis of 2008-2009, which had truly global effects, that economic conditions can vary quite dramatically from the long-term average. Thus, we are working in an area that is in great need of further exploration. The iterative model estimation algorithms, the data, computing power, model consistency and general theory all must be explored further to extend the tools available for understanding these dynamics.

Key results of this chapter have been published in Fitzpatrick and Marchev (2013), which was produced as a key component of this thesis with the overarching theme of multi-regime models involving Markov chains. The observed data are driven by different underlying regimes, which switch between each other over time, and Markov chains are involved in a number of ways. Firstly, in a similar way to the previous chapter, the estimation is performed by running a Markov chain. Secondly, the series of regimes that are selected over time is a Markov chain, meaning that, conditional on the current regime, the regime we select for the next time point is independent to the previous regimes. Finally, the parameters of the observed model are also Markov chains. This is because the credit rating data that we study have a discrete state at each time point and the dynamics of their potential migration to other ratings in the future, given a selected regime, is only dependent on their current state.

We study a different need for multi-regime models in Chapter 4 where the different regimes apply to different subsets of the population. We continue with the motivating problem of modelling the credit rating migration dynamics of firms;

however, we look into the theory behind the test for the number of Markov chains required to satisfactorily fit a particular dataset. We explore this problem with mixtures of continuous-time Markov chains and specifically develop the theory for the test between 1 and 2 Markov chain components in the mixture. We conjecture that, similarly to Hartigan (1985), the log-likelihood ratio test statistic diverges to infinity with the sample size, contrary to the claim from Frydman (2005) that we can use standard theory to apply a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the 1 component and 2 component mixture models. We provide evidence for our conjecture with the use of a parametric bootstrap procedure and then adapt the theory of Fukumizu (2003) to our case to definitively prove that the log-likelihood ratio test statistic does in fact diverge to infinity with the sample size. In order to develop a test for the presence of a Markov chain mixture, the next step would be to derive the limiting distribution of the log-likelihood ratio test statistic. We pursue this for a special case in the following chapter.

In Chapter 5, we focus on a simple case of the model in Chapter 4, where each Markov chain component consists of a *non-default* state and an absorbing *default* state. We derive the exact limiting distribution of the log-likelihood ratio test statistic for the test between 1 and 2 Markov chain mixture components. This test is equivalent to the test between 1 and 2 components in a censored exponential mixture problem. We show that the log-likelihood ratio test statistic is asymptotically equivalent to the square of the maximum of a Gaussian process over an interval whose length increases as the logarithm of the sample size. We prove that this Gaussian process is *locally stationary* so that we may utilise the extreme value theory developed in Hüsler (1990) to ultimately derive the exact limiting distribution of the log-likelihood ratio test statistic. These developments allow us to conduct a two sided test between 1 and 2 censored exponential mixture

components, which has applications beyond our original motivating example. We provide some conclusions and ideas for future research following our results.

2 Geometric ergodicity of the Gibbs sampler for the Poisson change-point model

In order to understand the changing rates of driver fatalities over the past 20 years in the state of Victoria, Australia, we observe a discrete time series of quarterly counts between March 1989 and December 2010, shown in Figure 1. From inspection, we can see that there is an initial sharp drop in the counts for each quarter, before a levelling off followed by another drop in the counts and a further levelling off. Although the more recent data is generally lower than the previous years, it does not seem to be following a linear trend, nor a gradual geometric decline. There seems to be multiple levels in the data for various time intervals but it isn't entirely obvious where these levels are. If the count data seemed to have one level of propensity, then we could fit a simple Poisson model. However, due to the multiple levels of counts, it is appropriate to apply a Poisson change-point model to the data. This will allow us to estimate where the change-points are, where we shift to a new regime and what the fatality rates are in each regime.

Poisson change-point models are used for modelling inhomogeneous time-series of count data. There are a number of methods available for estimating the parameters in these models using iterative techniques such as Markov chain Monte Carlo (MCMC). Many of these techniques share the common problem that there does not seem to be a definitive way of knowing the number of iterations required to obtain sufficient convergence. In this chapter, we show that the Gibbs sampler of the Poisson change-point model is geometrically ergodic. Establishing geometric ergodicity is crucial from a practical point of view as it implies the existence of a Markov chain central limit theorem, which can be used to obtain standard error estimates. We prove that the transition kernel is a trace-class operator, which implies geometric ergodicity of the sampler (see Khare and Hobert (2011) for de-

Fatal Crashes in Victoria

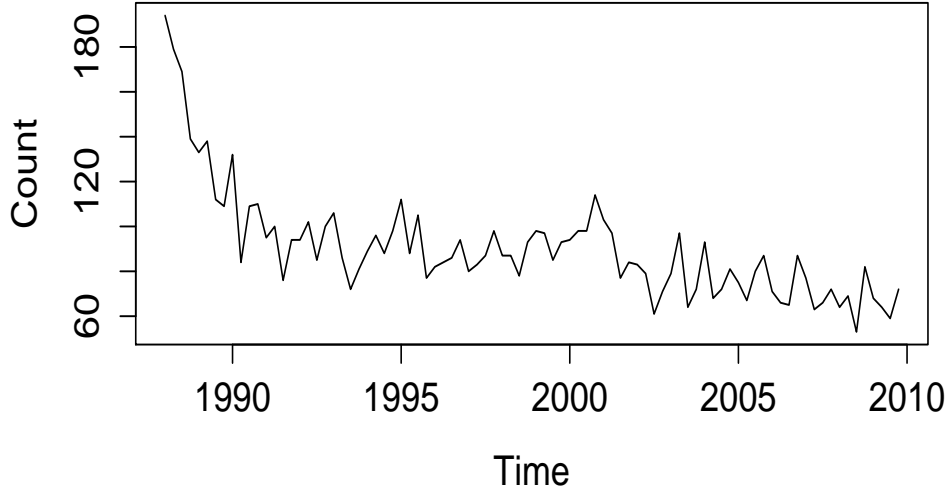


Figure 1: The count of driver fatalities in Victoria for each quarter between March 1989 and December 2010. Source: TAC (2011)

tails). We then examine the application of the sampler to a Poisson change-point model for quarterly driver fatality counts for the state of Victoria, Australia.

2.1 Introduction

Under the Poisson change-point model, we observe a non-homogeneous sequence of T independent Poisson random variables X_1, \dots, X_T . More specifically, we consider the case when the rate λ changes from λ_1 to λ_2 at an unknown point τ_1 , then from λ_2 to λ_3 at a later unknown point τ_2 , and so on, until the rate changes to λ_K , where it remains for the observation periods $\tau_K + 1$ to T . This model has been widely studied (see Carlin et al. (1992) and Raftery and Akman (1986), among others). Each of these models have a fixed K , which means the number of change-points is known *a priori*. The Bayesian Poisson change-point model

studied in Raftery and Akman (1986) assumed conjugate priors and has a single change-point at an unknown time. The model is applied to a well known data set consisting of intervals between coal-mining disasters given by Jarrett (1979). Carlin et al. (1992) present a general approach to hierarchical Bayesian change-point models, including a version of the Poisson change-point model that we apply to our data, and describes a Gibbs sampler procedure in great detail. Although Carlin et al. (1992) acknowledge the need to derive the number of iterations and sampler replications required for sufficient convergence, the convergence of the algorithm is concluded through inspection of the posterior distribution for the parameters after applying up to 50 iterations and 100 replications. Further understanding of the properties of convergence of the Gibbs sampler for the Poisson change-point model will allow for a more precise number of iterations and replications to be directly derived.

Here, we use a Poisson change-point model for detecting the shifts and levels of quarterly driver fatality counts for the state of Victoria, Australia. Within this application, the timing and size of the shifts in the dynamics of the data provide insight into the effectiveness of particular government policies in reducing the number of road fatalities.

In this study of non-homogeneous count data for driver fatalities in Victoria, we utilise the results from Khare and Hobert (2011) to show a theoretical result on the convergence of the Gibbs sampler for estimating the model parameters that is of great importance to practitioners. In cases where these models are utilised for providing objective evidence to influence future policy-making, we must have confidence that the iterative algorithm for estimating the model parameters has converged. It is an interesting approach to providing statistical evidence of shifts in outcomes. Traditionally, a hypothesis would be set that assumes a particular effect is or is not present and this hypothesis is tested as to whether we should adopt the

defined alternative. This essentially requires us to know what the alternatives are. For example testing whether data could be derived from a particular model (such as the standard normal distribution), we would produce a test statistic that has a particular distribution under the null hypothesis and infer with a particular level of confidence whether we should reject this hypothesis in favour of a more general alternative. However, with the class of models discussed here, we are only assuming a model form and then using the data to allow us to discover the potential causes for shifts in the rates of driver fatalities. This differs to us needing to guess the potential causes first then test for whether we should guess again. If the results from this more exploratory approach align with independent prior ideas as to what could be driving the data, our understanding can be further verified.

In Section 2.2, we outline the model specification and introduce some notation. We then discuss the estimation of the model parameters in Section 2.3. The main result is presented in Section 2.4 where we show that the Gibbs sampler is geometrically ergodic. This is a specific application of the results of Khare and Hobert (2011) to our model chosen here due to its practical significance. These theoretical results are used in practice in Section 2.5, where we apply the model to quarterly driver fatality counts for the state of Victoria, Australia. Our main interest is in estimating the non-constant fatality rate λ and the change-points $\tau = (\tau_1, \dots, \tau_K)$ by obtaining a sample from their posterior distributions. We are interested in estimating both the timing of the change-points as well as the size of the shift in fatality rates. The significant shifts in the driver fatality counts, which are thankfully being reduced over time, align with some key policy implementations and public campaigns, providing evidence for their impact. A comparison of the fatality rates in each regime provides a measurement of their effectiveness, despite the natural variation in the data from year to year. We then discuss some conclusions and potential avenues for future research.

2.2 Model specification

For the application of modelling the quarterly driver fatality counts, we are presented with a time series of count data. That is, a series of $0 < T < \infty$ positive integers Y_1, Y_2, \dots, Y_T representing the number of driver fatalities in each quarter. Upon inspection of Figure 1, we see that these numbers vary over the series within a reasonably controlled range and we see immediately that the earlier data points tend to have higher counts than the later data points. We are modelling these data in an exploratory fashion, to understand the features of the data, any patterns that emerge and the resulting insights this can give us about what to expect with future data points given relationships with causal factors that are not directly captured in the data (such as road safety regulations, number of cars, size and density of the population, types of vehicles on the roads, quality of the roads, quality of the drivers, weather and natural disasters etc.). Note that it is impossible to discern exactly what the causes are but we can show evidence that supports or challenges a particular claim. We could look to capture other information that may be related to the data and find a statistical relationship such as fitting a generalised linear model of sorts; however, this requires access to many other sources of data for the same time period and region involved. Given that our analysis is largely exploratory, we would be required to gather much more data than an eventual model as we should keep an open mind as to what may have the strongest relationship with our dependent variables. Alternatively, we can find patterns in our count data and use these patterns to point us in the right direction what could be causing these patterns to emerge. This approach is key. It starts with the data and we are guided to a greater understanding of what drives it.

Let us refer to the probability of a driver fatality within a particular time period with a particular *risk* level. Focussing again on the actual counts, we note that although the counts are greater for the earlier years than the later years, there

does not seem to be a steady decline. In fact, there seems to be a single step down in the counts and a levelling out before another step down. This multi-level effect points to a shift in risk levels that are constant for a certain period before shifting to a new level for the next period and so on. If we modelled all of the data with a regular Poisson model, we would not have a good idea of the level of risk at each time point but rather a view of the average risk over the entire observable period. From inspection, we see that a constant level of risk is certainly not appropriate. Poisson models may work to describe the count data but we must allow for the shift in the risk levels.

We thus consider the Poisson change-point model, where

$$\begin{aligned}
 Y_i | \boldsymbol{\lambda}, \tau &\stackrel{\text{ind.}}{\sim} \begin{cases} \text{Po}(\lambda_1) & \text{for } i = 1, \dots, \tau_1; \\ \text{Po}(\lambda_2) & \text{for } i = \tau_1 + 1, \dots, \tau_2; \\ \vdots & \\ \text{Po}(\lambda_K) & \text{for } i = \tau_{K-1} + 1, \dots, \tau_K; \\ \text{Po}(\lambda_{K+1}) & \text{for } i = \tau_K + 1, \dots, T. \end{cases} \quad (1) \\
 \lambda_i | \boldsymbol{\beta}, \tau &\stackrel{\text{ind.}}{\sim} \text{G}(a_i, \beta_i), i = 1, \dots, K + 1 \\
 \beta_i | \tau &\stackrel{\text{ind.}}{\sim} \text{IG}(c_i, \rho_i), i = 1, \dots, K + 1
 \end{aligned}$$

$0 < K \leq T - 1$ is a known constant and τ_1, \dots, τ_K are distributed as the order statistics from a random sample of size K taken without replacement from the set $\{1, 2, \dots, T - 1\}$.

Here $X \sim \text{G}(a, b)$ implies that X follows the gamma density

$$f_X(x) = \frac{x^{a-1}}{b^a \Gamma(a)} e^{-\frac{x}{b}}, \quad x > 0 \text{ and } X \sim \text{IG}(c, \rho) \text{ implies that } X \text{ follows the inverse gamma density } f_X(x) = \frac{1}{\rho^c x^{c+1} \Gamma(c)} e^{-\frac{1}{\rho x}}, \quad x > 0.$$

The particular form of this model is consistent with the literature. In fact, if we fix $K = 1$, then we have the Poisson change-point model that was studied in Carlin et al. (1992). It is also constructed for a Bayesian approach. Given the

data, there is no way for us to produce consistent maximum likelihood estimates as we do not know when the shifts in λ take place. If we knew when the shifts were (or guessed) then fitting the model with a frequentist approach would be trivial. However, with this approach we allow the timing of the shifts to vary, thus allowing the data to provide guidance as to where these could be. We may also analyse the posterior distribution of the parameters, given their prior distributions and the information provided by the data, which can give us a greater idea of our level of certainty with each of the parameters and the potential that there may be something quite different going on. The choice of prior distributions is consistent with the sort of data that we are analysing (count data that occurs where there are multiple experiments with a low risk of a particular outcome being experienced). These distributions are also conjugate prior distributions. That is, they retain their form in the posterior distribution after being combined with the data likelihood distribution.

We will firstly explore some theoretical properties of this general model before applying it specifically to our practical task at hand. This is the first time that this particular dataset has been analysed in this way, so our findings will be of use to policy makers seeking to further understand the drivers of the data. We also extend the theory to further our understanding of the rate of convergence of the Gibbs sampler for this model, which gives us some guidance as to the running time required for the iterative algorithm to fit the model before we can analyse the parameters and extract practical insights.

2.3 Estimation of the model parameters

Recall that our main interest is in estimating the vector $\boldsymbol{\lambda}$ and the change-points $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)$ by obtaining a sample from their posterior distributions. From

(1) we obtain the joint density

$$f(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{1}{\binom{T-1}{K}} \prod_{h=1}^{\tau_1} \frac{\lambda_1^{y_h} e^{-\lambda_1}}{y_h!} \prod_{i=2}^K \left\{ \prod_{j=\tau_{i-1}+1}^{\tau_i} \frac{\lambda_i^{y_j} e^{-\lambda_i}}{y_j!} \right\} \prod_{k=\tau_{K+1}}^T \frac{\lambda_K^{y_k} e^{-\lambda_K}}{y_k!} \quad (2)$$

$$\prod_{l=1}^{K+1} \frac{\lambda_l^{a_l-1}}{\beta_l^{a_l} \Gamma(a_l)} e^{-\frac{\lambda_l}{\beta_l}} \prod_{m=1}^{K+1} \frac{1}{\rho_m^{c_m} \beta_m^{c_m+1} \Gamma(c_m)} e^{-\frac{1}{\rho_m \beta_m}}.$$

Then, the complete posterior density is

$$f(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{y}) \propto \lambda_1^{\sum_{i=1}^{\tau_1} y_i + a_1 - 1} \prod_{k=2}^K \left\{ \lambda_{k+1}^{\sum_{i=\tau_{k-1}+1}^{\tau_k} y_i + a_{k+1} - 1} \right\} \lambda_{K+1}^{\sum_{i=\tau_{K+1}}^T y_i + a_2 - 1}$$

$$\times \frac{e^{-\lambda_1(\tau_1 + \frac{1}{\beta_1})} \prod_{k=2}^K \left\{ e^{-\lambda_k(\tau_k - \tau_{k-1} + \frac{1}{\beta_k})} \right\} e^{-\lambda_{K+1}(T - \tau_{K+1} + \frac{1}{\beta_{K+1}})} \prod_{k=1}^{K+1} \left\{ e^{-\frac{1}{\rho_k \beta_k}} \right\}}{\beta_1^{a_1 + c_1 + 1} \beta_2^{a_2 + c_2 + 1}}.$$

The desired sample will be obtained by running a two-stage Gibbs sampler that iterates between

$$f(\boldsymbol{\lambda} | \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y}) \quad \text{and} \quad f(\boldsymbol{\beta}, \boldsymbol{\tau} | \boldsymbol{\lambda}, \mathbf{y}),$$

where the sequence of $\boldsymbol{\beta}$'s will be simply ignored.

From (2), it is clear that conditional on $\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y}$, the parameters $\lambda_1, \dots, \lambda_{K+1}$ are independent with

$$\lambda_1 | \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y} \sim \text{G} \left(\sum_{i=1}^{\tau_1} y_i + a_1, \frac{\beta_1}{\tau_1 \beta_1 + 1} \right);$$

$$\lambda_k | \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y} \sim \text{G} \left(\sum_{i=\tau_{k-1}+1}^{\tau_k} y_i + a_k, \frac{\beta_k}{(\tau_k - \tau_{k-1}) \beta_k + 1} \right) \text{ for } k = 2, \dots, K; \quad (3)$$

$$\lambda_{K+1} | \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y} \sim \text{G} \left(\sum_{i=\tau_{K+1}}^T y_i + a_{K+1}, \frac{\beta_{K+1}}{(T - \tau_{K+1}) \beta_{K+1} + 1} \right).$$

Again, from (2) it is clear that conditional on $\boldsymbol{\lambda}, \mathbf{y}$, the parameters $\beta_1, \dots, \beta_{K+1}$

and $\boldsymbol{\tau}$ are independent with

$$\beta_k | \boldsymbol{\lambda}, \mathbf{y} \sim \text{IG} \left(a_k + c_k, \frac{\rho_k}{\rho_k \lambda_k + 1} \right) \text{ for } k = 1, \dots, K + 1; \quad (4)$$

$$f(\boldsymbol{\tau} | \boldsymbol{\lambda}, \mathbf{y}) = \frac{f(\mathbf{y} | \boldsymbol{\tau}, \boldsymbol{\lambda})}{\sum_{\tau'_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau'_i=\tau'_{i-1}+1}^{T-K-1+i} f(\mathbf{y} | \boldsymbol{\tau}', \boldsymbol{\lambda})}.$$

Remarks:

1. Note that we intentionally chose the parametrization of the gamma and inverse gamma densities so that equations (3) and (4) agree perfectly with the complete conditional distributions derived in Carlin et al. (1992).
2. It is possible to integrate out the $\boldsymbol{\beta}$ variables from the posterior density. For example, in the case of one change-point, we see that

$$f(\boldsymbol{\lambda}, \tau | \mathbf{y}) \propto \frac{\lambda_1^{\sum_{i=1}^{\tau} y_i + a_1 - 1} \lambda_2^{\sum_{i=\tau+1}^T y_i + a_2 - 1} e^{-\lambda_1 \tau} e^{-\lambda_2 (T-\tau)}}{(\rho_1 \lambda_1 + 1)^{a_1 + c_1} (\rho_2 \lambda_2 + 1)^{a_2 + c_2}}.$$

However, the above density, although available in closed form (apart from a normalizing constant), is not easy to draw from. Moreover, the introduction of more than one change point makes sampling from $f(\boldsymbol{\lambda}, \boldsymbol{\tau} | \mathbf{y})$ even harder, whereas with our approach the algorithm is essentially the same.

2.4 Geometric ergodicity of the Gibbs sampler

In this section we prove that the Gibbs sampler, originally described by Geman and Geman (1984), applied to the Poisson change-point model, specified in the previous section, is *geometrically ergodic*. A geometrically ergodic Gibbs sampler converges to its target distribution at a geometric rate. We do this by using the results in Khare and Hobert (2011) about data augmentation (DA) algorithms which are trace-class. DA algorithms involve the introduction of unobserved or

latent variables to sampling or iterative optimisation procedures. Stochastic DA algorithms constructed for posterior sampling can take the form of a two-block Gibbs sampler, such as the one used for our model.

Definition 2.1. If a DA algorithm based on a joint density $f(x, y)$ satisfies

$$\int_{\Theta} K^{mo}(\theta|\theta)d\theta = \int_{\mathcal{Y}} \int_{\mathcal{X}} f_{X|Y}(x|y)f_{Y|X}(y|x)\mu(dx)\nu(dy) < \infty, \quad (5)$$

then the Markov operator, K^{mo} , associated with the chain is a *trace-class* operator.

Here, $\Theta = \mathcal{X} \times \mathcal{Y}$ and also $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are the densities for the parameter subsets \mathcal{X} and \mathcal{Y} with measures μ and ν respectively.

Furthermore, if K^{mo} is a trace-class operator then it is *compact* and its norm $\|K^{mo}\| < 1$, so by Roberts and Rosenthal (1997), the corresponding Markov chain must be geometrically ergodic. Further details about trace-class operators can be found, for example, in Conway (1990).

We can prove geometric ergodicity of the Gibbs sampler for our model via the following theorem.

Theorem 2.2. *For the Poisson change-point model (1), the two conditional densities (3) and (4) satisfy*

$$\sum_{\tau_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau_i=\tau_{i-1}+1}^{T-K-1+i} \int_{\mathbb{R}_+^{K+1}} \int_{\mathbb{R}_+^{K+1}} f(\boldsymbol{\lambda}|\boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{y})f(\boldsymbol{\beta}, \boldsymbol{\tau}|\boldsymbol{\lambda}, \mathbf{y})d\boldsymbol{\beta}d\boldsymbol{\lambda} < \infty. \quad (6)$$

Therefore, the Gibbs sampler for the Poisson change-point model (1) is geometrically ergodic.

We can see from (5) that (6) implies that the Markov operator associated with the Gibbs sampler for the Poisson change-point model (1) is a trace-class operator.

We can then use the results of Roberts and Rosenthal (1997) to see that the Gibbs sampler for (1) is geometrically ergodic.

Proof. From (3) and (4) we can see that the left hand side becomes

$$\begin{aligned} & \sum_{\tau_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau_i=\tau_{i-1}+1}^{T-K-1+i} \int_0^\infty \cdots \int_0^\infty f(\lambda_1|\beta_1, \boldsymbol{\tau}, \mathbf{y}) \cdots f(\lambda_{K+1}|\beta_{K+1}, \boldsymbol{\tau}, \mathbf{y}) f(\beta_1|\lambda_1) \\ & \times \cdots f(\beta_{K+1}|\lambda_{K+1}) f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \mathbf{y}) d\beta_1 \cdots d\beta_{K+1} d\lambda_1 \cdots d\lambda_{K+1}. \end{aligned}$$

Note that

$$f(\boldsymbol{\tau}^*|\boldsymbol{\lambda}, \mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\tau}^*, \boldsymbol{\lambda})}{\sum_{\tau_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau_i=\tau_{i-1}+1}^{T-K-1+i} f(\mathbf{y}|\boldsymbol{\tau}, \boldsymbol{\lambda})} \leq 1 \text{ for all possible } \boldsymbol{\tau}^*,$$

which implies that the left hand side of the expression in the theorem is bounded above by

$$\begin{aligned} & \sum_{\tau_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau_i=\tau_{i-1}+1}^{T-K-1+i} \int_0^\infty \cdots \int_0^\infty f(\lambda_1|\beta_1, \boldsymbol{\tau}, \mathbf{y}) \cdots f(\lambda_{K+1}|\beta_{K+1}, \boldsymbol{\tau}, \mathbf{y}) \\ & \quad \times f(\beta_1|\lambda_1) \cdots f(\beta_{K+1}|\lambda_{K+1}) d\beta_1 \cdots d\beta_{K+1} d\lambda_1 \cdots d\lambda_{K+1} \\ & = \sum_{\tau_1=1}^{T-K-1} \sum_{i=2}^K \sum_{\tau_i=\tau_{i-1}+1}^{T-K-1+i} \left\{ \int_0^\infty \int_0^\infty f(\lambda_1|\beta_1, \boldsymbol{\tau}, \mathbf{y}) f(\beta_1|\lambda_1) d\beta_1 d\lambda_1 \right\} \\ & \quad \times \cdots \\ & \quad \times \left\{ \int_0^\infty \int_0^\infty f(\lambda_{K+1}|\beta_{K+1}, \boldsymbol{\tau}, \mathbf{y}) f(\beta_{K+1}|\lambda_{K+1}) d\beta_{K+1} d\lambda_{K+1} \right\}. \end{aligned}$$

Thus, since T is finite and since $f(\lambda_k|\beta_k, \boldsymbol{\tau}, \mathbf{y})$ and $f(\beta_k|\lambda_k)$ are distributed similarly for all $k = 1, \dots, K+1$, then it will suffice to prove that

$$\int_0^\infty \int_0^\infty f(\lambda_1|\beta_1, \boldsymbol{\tau}, \mathbf{y}) f(\beta_1|\lambda_1) d\beta_1 d\lambda_1 < \infty \quad \forall \tau_1 = 1, \dots, T-K-1. \quad (7)$$

Note also that letting $M(\tau_1) = \tau_1$, $M(\tau_{K+1}) = T - \tau_K$, and $M(\tau_i) = \tau_i - \tau_{i-1}$ for $i = 2, \dots, K$, and letting $N(\tau_1) = \sum_{j=1}^{\tau_1} y_j$, $N(\tau_{K+1}) = \sum_{j=\tau_K+1}^T y_j$, and $N(\tau_i) = \sum_{j=\tau_{i-1}+1}^{\tau_i} y_j$ for $i = 2, \dots, K$, we have

$$\begin{aligned} f(\lambda_i | \mathbf{y}, \boldsymbol{\tau}, b_i) &\sim \text{G} \left(a_i + N(\tau_i), \left(\frac{b_i}{b_i M(\tau_i) + 1} \right) \right) \\ f(b_i | \lambda_i) &\sim \text{IG} \left(a_i + c_i, \left(\frac{\rho_i}{\rho_i \lambda_i + 1} \right) \right) \end{aligned}$$

where a_i , c_i , and ρ_i are known positive constants, for $i = 1, \dots, K + 1$. Thus, taking the model specification into account, for any general $\tau \in \{1, \dots, T - 1\}$ and $N(\tau) = \sum_{i=1}^{\tau} y_i$, we are required to prove

$$\int_0^\infty \int_0^\infty \frac{\lambda^{a+N(\tau)-1} e^{-\lambda(\tau+\frac{1}{b})}}{\left(\frac{b}{b\tau+1}\right)^{a+N(\tau)} \Gamma(a+N(\tau))} \frac{e^{-\frac{1}{b}(\lambda+\frac{1}{\rho})}}{\left(\frac{\rho}{\rho\lambda+1}\right)^{a+c} \Gamma(a+c) b^{a+c+1}} db d\lambda < \infty. \quad (8)$$

We will do this by bounding the left hand side by integrable functions. We will bound the left and right tails of each of the b and λ supports by a different function and show that the result is still finite.

$$\begin{aligned} LHS &= \int_0^\infty \int_0^\infty I_1 db d\lambda \\ &= \int_0^\infty \int_0^\infty \frac{\lambda^{a+N(\tau)-1} (\lambda + \frac{1}{\rho})^{a+c} e^{-\lambda(\tau+\frac{2}{b})}}{\Gamma(2a+N(\tau)+c+1) \left(\frac{b}{b\tau+2}\right)^{2a+N(\tau)+c+1}} d\lambda \\ &\quad \times \left[\frac{\Gamma(2a+N(\tau)+c+1) \left(\frac{b}{b\tau+2}\right)^{2a+N(\tau)+c+1} e^{-\frac{1}{b}(\frac{1}{\rho})}}{\left(\frac{b}{b\tau+1}\right)^{a+N(\tau)} \Gamma(a+c) \Gamma(a+N(\tau)) b^{a+c+1}} \right] db \\ &< \int_0^\infty \int_0^M I_1 d\lambda db + \int_0^\infty \int_M^\infty \frac{\lambda^{2a+N(\tau)+c} e^{-\lambda(\tau+\frac{2}{b})}}{\Gamma(2a+N(\tau)+c+1) \left(\frac{b}{b\tau+2}\right)^{2a+N(\tau)+c+1}} d\lambda \\ &\quad \times \left[\frac{Q_1 \left(\frac{b}{b\tau+2}\right)^{2a+N(\tau)+c+1} e^{-\frac{1}{b}(\frac{1}{\rho})}}{\left(\frac{b}{b\tau+1}\right)^{a+N(\tau)} b^{a+c+1}} \right] db \end{aligned}$$

where $Q_1 = \frac{\Gamma(2a+N(\tau)+c+1)}{\Gamma(a+c)\Gamma(a+N(\tau))}$, since there exists $M > 0$ such that $(x + \frac{1}{\rho})^{a+c} <$

$x^{a+c+1} \forall \rho, a, c > 0$ and $x > M$. So now, we have

$$\begin{aligned}
LHS &< \int_0^\infty \int_0^M I_1 d\lambda db + Q_1 \int_0^\infty \frac{(b\tau + 1)^{a+N(\tau)} e^{-\frac{1}{b}(\frac{1}{\rho})}}{(b\tau + 2)^{2a+N(\tau)+c+1} b^{a+c+1}} db \\
&< \int_0^\infty \int_0^M I_1 d\lambda db + Q_1 \int_0^\infty \frac{e^{-\frac{1}{b}(\frac{1}{\rho})}}{(b\tau + 2)^{a+c+1}} db \\
&< \int_0^\infty \int_0^M I_1 d\lambda db + Q_1 \rho^{a+c} \Gamma(a+c) \int_0^\infty \frac{e^{-\frac{1}{b}(\frac{1}{\rho})}}{b^{a+c+1} \rho^{a+c} \Gamma(a+c)} db \\
&= \int_0^\infty \int_0^M I_1 d\lambda db + Q_2
\end{aligned}$$

where $Q_2 = Q_1 \rho^{a+c} \Gamma(a+c)$. Thus we can now focus on the first term, so

$$\begin{aligned}
LHS &< Q_2 + \int_0^M \int_0^\infty \frac{(b\tau + 1)^{a+N(\tau)} e^{-\frac{1}{b}(2\lambda + \frac{1}{\rho})}}{b^{2a+N(\tau)+c+1} \Gamma(2a + N(\tau) + c) (\frac{\rho}{2\rho\lambda+1})^{2a+N(\tau)+c}} db \\
&\quad \times \left[\frac{\Gamma(2a + N(\tau) + c) (\frac{\rho}{2\rho\lambda+1})^{2a+N(\tau)+c} \lambda^{a+N(\tau)-1} e^{-\lambda\tau}}{(\frac{\rho}{\rho\lambda+1})^{a+c} \Gamma(a+c) \Gamma(a+N(\tau))} \right] d\lambda
\end{aligned}$$

and since there exists $N > 0$ such that $(b\tau + 1)^{a+N(\tau)} < b^{[a+N(\tau)+1]} \forall b, \tau, a, N(\tau) > 0$ and $b > N$,

$$\begin{aligned}
&< Q_2 + \int_0^M \int_0^N I_1 dbd\lambda \\
&\quad + \int_0^M \mathbf{E}[b^{[a+N(\tau)+1]}] \frac{\Gamma(2a + N(\tau) + c) (\frac{\rho}{2\rho\lambda+1})^{2a+N(\tau)+c} \lambda^{a+N(\tau)-1} e^{-\lambda\tau}}{(\frac{\rho}{\rho\lambda+1})^{a+c} \Gamma(a+c) \Gamma(a+N(\tau))} d\lambda \\
&< Q_2 + \int_0^M \int_0^N I_1 dbd\lambda \\
&\quad + \int_0^M \frac{\Gamma(2a + N(\tau) + c - [a + N(\tau) + 1]) (\frac{\rho}{2\rho\lambda+1})^{a+N(\tau)} \lambda^{a+N(\tau)-1} e^{-\lambda\tau}}{(\frac{\rho}{2\rho\lambda+1})^{[a+N(\tau)+1]} \Gamma(a+c) \Gamma(a+N(\tau))} d\lambda \\
&< Q_2 + \int_0^M \int_0^N I_1 dbd\lambda \\
&\quad + \int_0^M \frac{\Gamma(2a + N(\tau) + c - [a + N(\tau) + 1])}{\Gamma(a+c) \Gamma(a+N(\tau))} \left(2\lambda + \frac{1}{\rho}\right)^{[a+N(\tau)]} e^{-\lambda\tau} d\lambda \\
&= Q_3 + \int_0^M \int_0^N I_1 dbd\lambda
\end{aligned}$$

where $Q_3 = Q_2 + M \frac{\Gamma(2a+N(\tau)+c-[a+N(\tau)+1])}{\Gamma(a+c)\Gamma(a+N(\tau))} \left(\sup_{0<\lambda<M} (2\lambda + \frac{1}{\rho})^{[a+N(\tau)]} e^{-\lambda\tau}\right)$, and under the condition that $2a+c-[a] \neq 1, 0, -1, -2, \dots$, which is not very restrictive.

Therefore, the last thing to prove is that

$$A = \int_0^M \int_0^N I_1 dbd\lambda < \infty. \quad (9)$$

For this, we note that the mode of an $\text{IG}(\alpha, \beta)$ distribution is $\frac{1}{\beta(\alpha+1)}$ and the mode of a $\text{G}(\alpha, \beta)$ distribution is $\frac{\alpha-1}{\beta}$. Thus, we have

$$\begin{aligned}
A &< \int_0^M \int_0^N \frac{(a + N(\tau) - 1)b}{b\tau + 1} \frac{\rho\lambda + 1}{\rho(a + c + 1)} dbd\lambda \\
&= \int_0^M \int_0^N I_2 dbd\lambda
\end{aligned}$$

and since

$$\frac{\partial I_2}{\partial b} = \frac{(a + N(\tau) - 1)(\rho\lambda + 1)}{\rho(a + c + 1)(b\tau + 1)^2} > 0 \text{ for all } 0 \leq b \leq N,$$

we have

$$A < N \int_0^M \frac{(a + N(\tau) - 1)(\rho\lambda + 1)}{\rho(a + c + 1)(N\tau + 1)^2} d\lambda.$$

Also,

$$\frac{\partial}{\partial \lambda} \left[\frac{(a + N(\tau) - 1)(\rho\lambda + 1)}{\rho(a + c + 1)(N\tau + 1)^2} \right] = \frac{(a + N(\tau) - 1)\rho}{\rho(a + c + 1)(N\tau + 1)^2} > 0 \forall 0 \leq \lambda \leq M,$$

so we have

$$A < MN \frac{(a + N(\tau) - 1)(M\rho + 1)}{\rho(a + c + 1)(N\tau + 1)^2} < \infty.$$

Therefore,

$$\int_{\Theta} K^{mo}(\theta|\theta) d\theta < \infty$$

as required. □

2.5 Applications to Victorian driver fatality count data

A popular application of the Poisson change-point model is in the assessment of the effectiveness of government policies. We analyse count data for the number of fatal crashes in each calendar quarter, recorded in the state of Victoria, Australia. The data is from the Australian Government - Department of Infrastructure and Transport, TAC (2011). This particular time series is of interest due to the fact that the frequency of fatal crashes in Victoria has reduced dramatically over the

past twenty years, despite an increase in the number of drivers over the same period. It is important to assess the nature of the reductions, whether it is a steady downward trend or whether there are sudden drops due to various effective policies or other discrete influences.

Our data set consists of $T = 88$ quarterly observations, ranging from the March quarter of 1989 to the December quarter of 2010. We fit the Poisson change-point model from (1), with $K = 2$ change-points, to the data. The constant hyperparameters are set with $a_1 = 170$, $a_2 = 120$, $a_3 = 80$, $c_1 = c_2 = c_3 = 1$, and $\rho_1 = \rho_2 = \rho_3 = 1$, so that the sampled rate parameters λ_i are more likely to begin with values that roughly mirror the values in Figure 1. The algorithm runs for 100,000 iterations, after a burn-in period of 10,000 iterations. The results are as shown in Figure 2.

The estimated change-points are after the June quarter of 1990 and the March quarter of 2002. The first change-point was a major drop that followed the implementation of the Road Safety Act (1986), which governs road use and deals with licensing and road related offences in Victoria. This also led to the establishment of the Transport Accident Commission (TAC), which is the statutory insurer of third-party personal liability. In 1989 there was also a federal Ten Point Plan to reduce the number of deaths on Australian roads. The TAC have also launched successful TV advertising campaigns in Victoria throughout the 1990's and early 2000's. In 2000, the National Road Safety Strategy 2001-2010 was developed, with a target of a 40 per cent reduction in the population rate of road fatalities from 9.3 to 5.6 per 100,000. The Strategy was supported by a series of two-year action plans. The introduction of this targeted focus on road fatality reduction coincided with the second change-point seen in the data.

It is difficult to discern exactly what causes the change-points. However, it is evident that significant shifts in the rate of road fatalities in Victoria have been

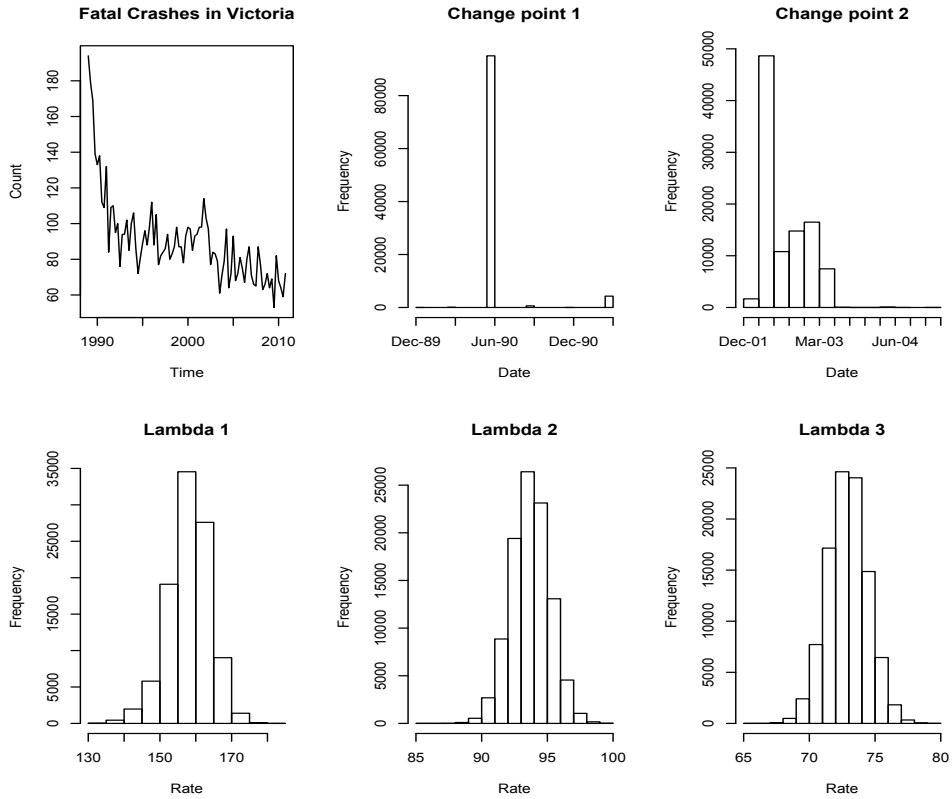


Figure 2: Crash statistics data with Poisson change-point model parameters

observed and assuming that the behaviour of the general population is consistent over time, it is likely that the introduction and implementation of these major policy introductions have lead to a significant reduction in road fatality rates. The model parameter estimates allow us to measure the difference in rates. From the ratios of the posterior estimates of λ_2/λ_1 and λ_3/λ_2 , we see that the first change-point lead to a drop in road fatalities of between 36 and 45 per cent and the second change-point lead to a drop in road fatalities of between 18 and 26 per cent, at the 95 per cent confidence level. This form of objective feedback to the effectiveness of major policy decisions is vital to continued high levels of governance.

2.6 Conclusions

On a practical level, we have seen how the Poisson change-point model can be estimated for modelling road fatalities, where there are various policies and laws in place at different times. The model clearly shows shifts in the count levels towards a lowered rate of fatalities, despite a rise in the population of drivers. Given the validity of the datapoints, we can conclude that the level of fatalities has significantly reduced over the past 20 years. It would be of interest to see if these results held true for other states and territories around Australia, which could help identify the cause of the decline in fatalities. For example, if the downward shifts occur around the same time, it could indicate the cause of the decline in fatalities was due to a federal intervention. On a theoretical level, the results in this chapter imply that the Gibbs sampler for the Poisson change-point model will converge at a geometric rate. Thus, given a specific convergence level, the minimum number of iterations required can be calculated. Although we have identified a key quality of the convergence rate of the sampler, the calculation of the specific rate of convergence is left for further research. It would also be of interest to see if the bounding technique of section 2.4 can be used to prove geometric ergodicity of MCMC algorithms for other models.

3 Efficient Bayesian estimation of the multivariate double chain Markov model

Key results of this chapter have been published in Fitzpatrick and Marchev (2013), which was produced as a key component of this thesis with the overarching theme of multi-regime models involving Markov chains. The observed data are driven by different underlying regimes, which switch between each other over time, and Markov chains are involved in a number of ways. Firstly, in a similar way to the previous chapter, the estimation procedure is a Markov chain. Secondly, the series of regimes that are selected over time is a Markov chain, meaning that, conditional on the current regime, the regime we select for the next time point is independent to the previous regimes. Finally, the parameters of the observed model are also Markov chains. This is because the credit rating data that we study have a discrete state at each time point and the dynamics of their potential migration to other ratings in the future, given a selected regime, is only dependent on their current state.

The *double chain Markov model* (DCMM) is used to model an observable process $Y = \{Y_t\}_{t=1}^T$ as a Markov chain with transition matrix, P_{x_t} , dependent on the value of an unobservable (hidden) Markov chain $\{X_t\}_{t=1}^T$. We present and justify an efficient algorithm for sampling from the posterior distribution associated with the DCMM, when the observable process Y consists of independent vectors of (possibly) different lengths. Convergence of the Gibbs sampler, used to simulate the posterior density, is improved by adding a random permutation step. Simulation studies are included to illustrate the method. The problem that motivated our model is presented at the end. It is an application to real data, consisting of the credit rating dynamics of a portfolio of financial companies where the (unobserved) hidden process is the state of the broader economy.

3.1 Introduction

Let \mathcal{Y} be a set of J elements. For convenience we will denote them with the first J positive integers; i.e., $\mathcal{Y} = \{1, \dots, J\}$. Consider a stochastic process $\{Y_t\}_{t=0}^T$, where each Y_t takes values in \mathcal{Y} for $t = 0, \dots, T$. Dependence among such Y_t 's, taking values in a finite state space, can be modeled by Markov chains. For example, the first order simple Markov chain model can be described as follows:

$$\begin{aligned} \mathbb{P}((Y_1, \dots, Y_T)^\top = (y_1, y_2, \dots, y_T)^\top | Y_0 = y_0, \theta) \\ = \prod_{t=1}^T \theta_{y_{t-1}y_t}, \end{aligned}$$

where θ is a $J \times J$ transition matrix such that $\theta_{ij} = \mathbb{P}(Y_t = j | Y_{t-1} = i, \theta)$ for $t = 1, \dots, T$, $i, j = 1, \dots, J$ and the elements in each row of θ sum to 1. In other words, regardless of any external factors that may affect the observations, given the state y_t of the random variable at the current time, it migrates with the same multinomial distribution of probabilities $(\theta_{y_t 1}, \dots, \theta_{y_t J})$ to the other possible states. A more rigorous definition of Markov chains in a general state space can be found in Meyn and Tweedie (1993).

There have been different extensions to the simple Markov chain model that have emerged in the literature. One of the most important has been the Hidden Markov model (HMM) which was first presented in the late 1960's in Baum and Petrie (1966) and can be regarded as a Markov chain observed with noise. More precisely, a HMM is a stochastic process $\{(Y_t, X_t)\}_{t=0}^T$, where $\{X_t\}_{t=0}^T$ is a hidden Markov chain (i.e. unobservable), and $\{Y_t\}_{t=0}^T$ is a sequence of (observable) independent random variables such that the distribution of Y_t depends on $X_t, t = 0, \dots, T$. An excellent book on inference in HMM's is Cappé et al. (2005).

Various applications, in areas such as meteorology, biotechnology, finance and speech recognition, have motivated the exploration of the properties of HMM's.

For example, Churchill (1989) uses HMM's to study the sequences of bases on a DNA molecule and Hughes et al. (1999) study the relationship between observed rainfall occurrence and broad scale atmospheric circulation patterns via HMM's. These models have also been popular in their application to credit modeling in recent years. Studies such as Giampieri et al. (2005), and Korolkiewicz and Elliott (2008) use HMM's to model credit rating dynamics, by making the assumption that the observed ratings are not dependent upon previous observed ratings but rather on the hidden variables, representing the effects of the broader economy. A good summary of the bibliography on HMM's can be found in Cappé (2001).

Since the model we consider in this chapter is a version of a HMM, we now describe the HMM in more detail. Assume that the hidden process $\{X_t\}_{t=0}^T$ evolves independently of $\{Y_t\}_{t=0}^T$ and is a Markov chain with first-order transition matrix Π of dimension $a \times a$ and initial state distribution $\Pi_0 := (\pi_{01}, \dots, \pi_{0a})^\top$. Assume further that at each time point $t = 0, \dots, T$, depending on the value of the hidden process x_t , there are a finite number, a , of possible distributions of the random variable Y_t that takes values in the set \mathcal{Y} . We write the mass function of Y_t as $P(Y_t = y_t | X_t = x_t, \Theta) = \theta_{x_t, y_t}$, where $\Theta = \{\theta_{k,l}, k = 1, \dots, a, l \in \mathcal{Y}\}$, are unknown parameters. That is, letting $\theta = \{\Pi_0, \Pi, \Theta\}$, the HMM can be described as

$$\begin{aligned} P(y_0, \dots, y_T, x_0, \dots, x_T | \theta) &= P(y_0, \dots, y_T | x_0, \dots, x_T, \theta) P(x_0, \dots, x_T | \theta) \\ &= P(x_0 | \Pi_0) P(y_0 | \theta_{x_0}) \prod_{t=1}^T [P(y_t | x_t, \Theta) P(x_t | x_{t-1}, \Pi)] \\ &= \pi_{0x_0} \theta_{x_0, y_0} \prod_{t=1}^T \theta_{x_t, y_t} \pi_{x_{t-1}, x_t}. \end{aligned}$$

These models work well for modeling the heterogeneity of the observed process over time. However, they do not incorporate any direct dependence between observations. The logical extension is to allow the hidden Markov process to select

one of a finite number of Markov chains to drive the observed process at each time point. This sort of model is known as the *double chain Markov model* (DCMM) and was first formally presented in Berchtold (1999). It is basically designed for modeling non-homogeneous time series. If a time series can be decomposed into a finite mixture of Markov chains, then the DCMM can be applied to describe the switching process between these chains. This idea is not entirely new. The first extension was to combine the HMM with an autoregressive model for the observed process in Poritz (1982) and later in Kenny et al. (1990). Then Wellekens (1987) and Paliwal (1993) presented a model, similar to the DCMM for both the continuous case of the HMM and the discrete case respectively. Berchtold (1999) differed from Paliwal (1993) with a more rigorous derivation of the forward-backward and Viterbi algorithms involved in the model estimation and also by interpreting the relation between observed outputs of the model as a non-stationary Markov chain.

There have been extensions to the DCMM presented in Berchtold (1999), increasing the order of the Markov chains as in Eisenkopf (2008). However, this leads to an explosion in the number of parameters. There exist alternatives to modeling higher order dependence in Markov chains, such as in the mixture transition distribution (MTD) model presented in Raftery (1985), which presents the conditional probability of the current state as a linear combination of contributions from each of a fixed number of past states. An iterative algorithm for the estimation of these models was described in Berchtold (2002). The DCMM was extended in Eisenkopf (2008) using the theory of MTD's in Berchtold (2002) to show that the DCMM can handle higher order relationships among the hidden states as well as the observed outputs.

There are alternative generalizations of the HMM, which also take into account the heterogeneity of mixture models over time. In Lanchantin et al. (2008), the triplet Markov chain (TMC) model is presented, which can be viewed as an al-

ternative generalization of HMM's that is slightly different to the DCMM, where the non-stationary distribution of the hidden Markov chain was modelled by an auxiliary process governing the switching of the transition matrix over time (in the DCMM, the hidden process is time-homogeneous). There are several explorations of the TMC, such as in Pieczynski and Desbouvries (2005) and Pieczynski (2007). Finally, we point the interested reader to Kirshner (2005), where there is a detailed description of all levels of generalization from HMM's to models such as the DCMM, where there are direct relationships between the observed states, to non-homogeneous hidden Markov models with autoregressive observed states, similar to the TMC.

The computational estimation of the DCMM is explored in Berchtold (1999). Due to the structure of the DCMM, there is no direct formula to compute the log-likelihood. The problem is solved using an iterative procedure known as the forward-backward algorithm. The estimation of the model parameters is traditionally obtained by an *expectation-maximisation* (EM) algorithm known in the speech recognition literature as the Baum-Welch algorithm. Finally, the optimal sequence of hidden states is computed using another iterative procedure called the Viterbi algorithm presented in Forney (1973).

This chapter is focused specifically on multivariate time series data, which is especially relevant in the context of modeling vectors of observations of different lengths for each time point, such as in credit portfolio applications. The estimation of the hidden states, the model parameters and the hidden Markov process parameters is from a Bayesian perspective and is carried out using an efficient extension of the techniques presented in Chib (1996). In order to improve the convergence speed of the Gibbs sampler used to simulate the posterior density, we employ the random permutation sampler presented in Frühwirth-Schnatter (2001). During each iteration of the sampling process, the hidden states are sampled from

their joint distribution, given the current parameter estimates and the observed data. Then, we randomly permute the current labelling of the states of the hidden process. This permutation of the labels is justified and is shown to be optimal using the recent results of Hobert and Marchev (2008). After obtaining the MCMC sample, a post-processing algorithm from Stephens (2000), as presented in Boys and Henderson (2002), is utilised to find the most suitable permutation of the labels at each run of the sampler so that a consistent form of the model results, without the non-identifiability arising from label switching.

Our work was motivated by the lack of appropriate models in the context of credit portfolio modelling. In this setup the hidden Markov process represents the effects of the broader economy and governs the particular regime driving the transitions of credit ratings in a large portfolio of firms for each time point. We apply our model on a dataset comprised of monthly Standard and Poor's (S&P) credit rating transitions for a portfolio of globally sourced financial institutions and insurance companies from the 1st of January 1981 to the 1st of January 2010. The estimated switching behavior of the hidden Markov regimes selected for each time point bear remarkable similarities to the behaviour of the global economy over the last three decades, as explained in Section 5.

The chapter is organised as follows. In Section 3.2 we specify the model and introduce the notation and background to the theory of DCMM's. In Section 3.3, we estimate the model parameters from a Bayesian perspective, using an efficient Data Augmentation (DA) algorithm in combination with the post-processing algorithms of Stephens (2000) and Boys and Henderson (2002). Section 3.4 displays the results of the model when applied to simulated data. In Section 3.5 we apply our model on real data from Standard and Poor's. Finally, Section 3.6 provides conclusions and ideas for further research in this area.

3.2 Model specification

In this section, we describe our multivariate Bayesian DCMM and its parameters and derive the density function of the complete data set, consisting of both hidden and observed variables.

Consider data of n random variables observed discretely over time, each of potentially different lengths. That is, for each $i = 1, \dots, n$, we observe a vector, $(y_{i,u_i}, \dots, y_{i,m_i})^\top$, where $u_i < m_i$. Define

$$u_0 := \min_{1 \leq i \leq n} \{u_i\}, \text{ and } M := \max_{1 \leq i \leq n} \{m_i\}$$

and note that the times u_i and m_i may vary over the entire observation period from u_0, \dots, M with the only restriction that $m_i - u_i \geq 1, i = 1, \dots, n$.

Assume that for each $i = 1, \dots, n$ and each time point $t = u_i, \dots, m_i$, the random variable $Y_{i,t} \in \mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, J\}$, and is modelled as dependent upon the value at the previous time point, $y_{i,(t-1)}$, as well as on a hidden state, x_t , for that time point. We assume that the hidden process $\mathbf{X} = \{X_t\}_{t=u_0}^M$ is a Markov chain with first-order transition matrix Π of dimension $a \times a$, where $\pi_{gh} = P(X_t = h | X_{t-1} = g)$ for $g, h = 1, \dots, a$ and $t = u_0 + 1, \dots, M$. We let the first hidden state X_{u_0} be selected from a multinomial distribution with vector of probabilities $r = (r_1, \dots, r_a)^\top$. We also assume that the observable process is a Markov chain with a possible transition matrices P_1, \dots, P_a , each of order $J \times J$, such that for a given hidden state x_t , the elements of the transition matrix P_{x_t} are

$$p_{x_t, jk} = P(Y_{i,t} = k | Y_{i,(t-1)} = j, X_t = x_t),$$

for $i = 1, \dots, n, t = u_i + 1, \dots, m_i$.

For each random variable, we consider the time of initial observation u_i , the

initial observed state y_{i,u_i} and the number of consecutive time-points that it was observed $m_i - u_i + 1$ as fixed so all inference is conditional on those values. We denote the collection of all parameters in our model by θ and observe that $\theta \in \Theta$, where Θ is the d -dimensional hypercube with d equal to the number of free parameters in the model, since all parameters are probabilities between 0 and 1.

Conditional on \mathbf{X} , each of the random variables are modelled independently of each other. For each $i = 1, \dots, n$, if we define $\mathbf{y}_i := (y_{i,(u_i+1)}, \dots, y_{i,m_i})^\top$, then we consider the following hierarchical Bayesian model:

$$\begin{aligned}
P(\mathbf{y}_i | y_{i,u_i}, \mathbf{x}, \theta) &= P(y_{i,(u_i+1)} | y_{i,u_i}, x_{u_i+1}) \times \dots \times \\
&\quad \times P(y_{i,m_i} | y_{i,(m_i-1)}, x_{m_i}) \\
P(\mathbf{x} | \theta) &= P(x_{u_0}) P(x_{u_0+1} | x_{u_0}) \times \dots \times \\
&\quad \times P(x_M | x_{M-1}) \\
P(\theta) &= P(r) P(\Pi) P(P_1) \dots P(P_a),
\end{aligned} \tag{10}$$

where, similarly to Chib (1996), the priors on r , Π , P_1, \dots, P_a are Dirichlet as follows:

$$\begin{aligned}
r &\sim D(\alpha_{01}, \dots, \alpha_{0a}) \\
(\pi_{i1}, \dots, \pi_{ia}) &\stackrel{\text{ind}}{\sim} D(\alpha_{i1}, \dots, \alpha_{ia}), i = 1, \dots, a \\
(p_{1,l1}, \dots, p_{1,lJ}) &\stackrel{\text{ind}}{\sim} D(\alpha_{1,l1}, \dots, \alpha_{1,lJ}), l = 1, \dots, J \\
&\vdots \\
(p_{a,l1}, \dots, p_{a,lJ}) &\stackrel{\text{ind}}{\sim} D(\alpha_{a,l1}, \dots, \alpha_{a,lJ}), l = 1, \dots, J,
\end{aligned}$$

and the α 's are given constants. More details on how priors are chosen for HMM's can be found in Subsection 13.1.2 of Cappé et al. (2005).

Then the model for $\mathbf{Y} := (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ is

$$P(\mathbf{y}|\mathbf{y}_0, \mathbf{x}, \theta) = \prod_{i=1}^n P(\mathbf{y}_i | y_{i,u_i}, \mathbf{x}, \theta), \quad (11)$$

where $\mathbf{y}_0 := (y_{1,u_1}, \dots, y_{n,u_n})^\top$ and $P(\mathbf{x}|\theta)$ and $P(\theta)$ are as specified in (10).

From equation (10) the joint mass function of \mathbf{Y}_i , and \mathbf{X} given y_{i,u_i} and θ is:

$$\begin{aligned} P(\mathbf{y}_i, \mathbf{x} | y_{i,u_i}, \theta) &= r_{x_{u_0}} \pi_{x_{u_0} x_{u_0+1}} \dots \pi_{x_{u_i} x_{u_i+1}} p_{x_{u_i+1}, y_{i,u_i} y_{i,(u_i+1)}} \\ &\quad \times \dots \times \pi_{x_{m_i-1} x_{m_i}} p_{x_{m_i}, y_{i,(m_i-1)} y_{i,m_i}} \dots \pi_{x_{M-1} x_M} \\ &= \left[\prod_{l=1}^a r_l^{I_{\{x_{u_0}\}}(l)} \right] \prod_{t=u_0+1}^M \prod_{g=1}^a \prod_{h=1}^a \pi_{gh}^{I_{\{(x_{t-1}, x_t)\}}(g,h)} \\ &\quad \times \left[\prod_{t=u_i+1}^{m_i} \prod_{l=1}^a \prod_{j=1}^J \prod_{k=1}^J p_{l,jk}^{I_{\{(y_{i,(t-1)}, y_{i,t}, x_t)\}}(j,k,l)} \right], \end{aligned} \quad (12)$$

where $I_A(x)$ is the usual indicator function of a set A .

Next, we utilise the fact that the random vectors \mathbf{Y}_i , for $i = 1, \dots, n$ are independent, conditional on the hidden process \mathbf{X} , when deriving the joint mass function of all random variables \mathbf{Y} and \mathbf{X} :

$$\begin{aligned} P(\mathbf{y}, \mathbf{x} | \mathbf{y}_0, \theta) &= \left[\prod_{l=1}^a r_l^{I_{\{x_{u_0}\}}(l)} \right] \left[\prod_{t=u_0+1}^M \prod_{g=1}^a \prod_{h=1}^a \pi_{gh}^{I_{\{(x_{t-1}, x_t)\}}(g,h)} \right] \\ &\quad \times \left[\prod_{i=1}^n \prod_{t=u_i+1}^{m_i} \prod_{l=1}^a \prod_{j=1}^J \prod_{k=1}^J p_{l,jk}^{I_{\{(y_{i,(t-1)}, y_{i,t}, x_t)\}}(j,k,l)} \right]. \end{aligned} \quad (13)$$

We are interested in exploring the posterior density $f(\theta|\mathbf{y}) := \frac{f(\mathbf{y}, \theta)}{f(\mathbf{y})}$, where $f(\mathbf{y}, \theta)$ and $f(\mathbf{y})$ are defined from (13) and (10) as

$$f(\mathbf{y}, \theta) = \sum_{\mathbf{x} \in \mathcal{X}^m} P(\mathbf{y}, \mathbf{x} | \mathbf{y}_0, \theta) P(\theta) \quad (14)$$

and $f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y}, \theta) d\theta$ with \mathcal{X}^m being the m -tuple product of the set $\{1, \dots, a\}$ with itself. Here, $m = M - u_0 + 1$. Of course, given the nature of $P(\mathbf{y}, \mathbf{x}|\mathbf{y}_0, \theta)$ in (13) and the summation in (14), direct calculation of $f(\theta|\mathbf{y})$ is impossible; however, as explained in the next section, it is possible to construct an efficient MCMC algorithm to obtain approximate draws from it.

3.3 Estimation of the model parameters

The target density $f(\theta|\mathbf{y})$, as defined in the previous section, is not available in closed form, but can be presented as the θ -marginal density of $f(\theta, \mathbf{x}|\mathbf{y})$, where $f(\theta, \mathbf{x}|\mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{x}|\mathbf{y}_0, \theta)P(\theta)}{f(\mathbf{y})}$. Therefore, we will employ the data augmentation (DA) algorithm of Tanner and Wong (1987) to obtain approximate draws from it. Thus, we run a two-stage Gibbs sampler that alternates between sampling from

$$P(\mathbf{x}|\mathbf{y}, \theta) := \frac{f(\theta, \mathbf{x}|\mathbf{y})}{f(\theta|\mathbf{y})} \quad (15)$$

and

$$P(\theta|\mathbf{x}, \mathbf{y}) := \frac{f(\theta, \mathbf{x}|\mathbf{y})}{f(\mathbf{x}|\mathbf{y})}, \quad (16)$$

where $f(\mathbf{x}|\mathbf{y})$ is the \mathbf{x} -marginal density of $f(\theta, \mathbf{x}|\mathbf{y})$. The exact forms of (15) and (16) are derived in the next two subsections.

It is well-known that in Bayesian mixture models, there is the so called problem of “label switching”, which means that the target posterior density is multi-modal and the sampler can easily get stuck in one of the modes (or explore the modes irregularly). We remedy this issue by performing an additional step, which randomly permutes the labels after sampling from (15). This random permutation step was introduced in Frühwirth-Schnatter (2001) and further analyzed in Hobert et al. (2011) as being a special case of the general scheme for improvement of DA algorithms, presented in Hobert and Marchev (2008). What we will show in Subsection

3.3.3 is that this extra step is also optimal in the sense of Hobert and Marchev (2008), as being constructed via a group action on \mathcal{X}^m with the appropriate Haar measure.

Finally, the parameters θ are estimated as posterior means, calculated from the output of the modified DA algorithm, after a post-processing step is applied (as detailed in Subsection 3.3.4).

3.3.1 Sampling from $\mathbf{P}(\mathbf{x}|\mathbf{y}, \theta)$

Chib (1996) developed a method for simulating the hidden states \mathbf{x} , given a particular $i \in \{1, \dots, n\}$ and a single vector of observed data, \mathbf{y}_i . We will generalise this algorithm to a set of multiple vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$, of different lengths.

For $u_0 < t < M$ define

$$\begin{aligned}\mathbf{x}_{-t} &:= (x_{u_0}, \dots, x_t) \\ \mathbf{x}^t &:= (x_t, \dots, x_M) \\ \mathbf{y}_{,t} &:= \bigcup_{i:u_i < t} \{(y_{i,u_i}, \dots, y_{i,\min\{t,m_i\}})\} \\ \mathbf{y}^t &:= \bigcup_{i:t < m_i} \{(y_{i,\max\{t+1,u_i\}}, \dots, y_{i,m_i})\} \\ \mathbf{y}(t) &:= \{y_{i,t}\} \text{ for all } i \in \{1, \dots, n\} \text{ with } u_i \leq t \leq m_i,\end{aligned}$$

and note that the first, third and fifth definitions are also valid for $t = M$.

The following lemma is used to derive $\mathbf{P}(\mathbf{x}|\mathbf{y}, \theta)$:

Lemma 3.1. *For $t = u_0, \dots, M - 1$ we have*

$$P(x_{t+1}|x_t, \mathbf{y}_{,t}, \theta) = P(x_{t+1}|x_t, \Pi).$$

Proof. For all $t = u_0, \dots, M - 1$ we have

$$\begin{aligned} \text{LHS} &= P(x_{t+1}|x_t, \mathbf{y}_t, \theta) \\ &= \frac{P(\mathbf{y}_t|x_t, x_{t+1}, \theta)P(x_{t+1}|x_t, \theta)}{P(\mathbf{y}_t|x_t, \theta)} = P(x_{t+1}|x_t, \theta) \end{aligned}$$

since, by the definition of the model in (10), the observable data up to any such time t is not dependent upon the unobservable data at time $t + 1$ and only dependent on the unobservable data up to time t and θ . Then,

$$P(x_{t+1}|x_t, \theta) = P(x_{t+1}|x_t, \Pi)$$

since, by the definition of the model in (10), the unobservable data is driven by a hidden Markov chain with transition matrix Π that is not dependent upon the other parameters. \square

Our main result about sampling from $P(\mathbf{x}|\mathbf{y}, \theta)$ follows.

Theorem 3.2. *For data of independent vectors $\{\mathbf{Y}_i\}_{i=1}^n$ the joint distribution, $P(\mathbf{x}, \mathbf{y}, \theta)$, of the hidden data, the observed data and the parameters is given by*

$$\begin{aligned} P(\mathbf{x}, \mathbf{y}, \theta) &\equiv P(\mathbf{x}_{-M}|\mathbf{y}_M, \theta) \propto P(x_M|\mathbf{y}_{M-1}, \theta) f(\mathbf{y}(M)|\mathbf{y}_{M-1}, \theta_{x_M}) \\ &\quad \times \prod_{t=u_0+1}^{M-1} P(x_{t+1}|x_t, \Pi) P(x_t|\mathbf{y}_{t-1}, \theta) f(\mathbf{y}(t)|\mathbf{y}_{t-1}, \theta_{x_t}) \\ &\quad \times P(x_{u_0+1}|x_{u_0}, \Pi) P(x_{u_0}|r) \end{aligned}$$

with

$$P(x_t|\mathbf{y}_{t-1}, \theta) = \sum_{l=1}^a P(x_t|x_{t-1} = l, \Pi) P(x_{t-1} = l|\mathbf{y}_{t-1}, \theta).$$

Remark: Note that from the above only $P(x_t|\mathbf{y}_{t-1}, \theta)$ needs to be calculated

and the remaining components are from the model specified in equations (10) and (11).

Proof. The joint mass function of the hidden states, given the parameters and the observed data as vectors at each time point is

$$P(\mathbf{x}_{-M}|\mathbf{y}_{,M}, \theta) = P(x_M|\mathbf{y}_{,M}, \theta) \times \dots \times P(x_t|\mathbf{y}_{,M}, \mathbf{x}^{t+1}, \theta) \dots \times P(x_{u_0}|\mathbf{y}_{,M}, \mathbf{x}^{u_0+1}, \theta).$$

The ‘‘typical term’’ can be written as

$$\begin{aligned} P(x_t|\mathbf{y}_{,M}, \mathbf{x}^{t+1}, \theta) &= P(x_t|\mathbf{y}_{,t}, \mathbf{y}^{t+1}, \mathbf{x}^{t+1}, \theta) \\ &= \frac{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, x_t, \theta)P(x_t|\mathbf{y}_{,t}, \theta)}{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, \theta)} \\ &= P(x_t|\mathbf{y}_{,t}, \theta) \frac{P(x_{t+1}, \mathbf{x}^{t+2}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, x_t, \theta)}{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, \theta)} \\ &= P(x_t|\mathbf{y}_{,t}, \theta)P(x_{t+1}|x_t, \mathbf{y}_{,t}, \theta) \frac{P(\mathbf{x}^{t+2}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, x_t, x_{t+1}, \theta)}{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, \theta)} \\ &= P(x_t|\mathbf{y}_{,t}, \theta)P(x_{t+1}|x_t, \Pi) \frac{P(\mathbf{x}^{t+2}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, x_t, x_{t+1}, \theta)}{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, \theta)} \end{aligned}$$

by Lemma 3.1. Now, $\frac{P(\mathbf{x}^{t+2}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, x_t, x_{t+1}, \theta)}{P(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}|\mathbf{y}_{,t}, \theta)}$ depends only on x_{t+1} , and is therefore independent of x_t and thus can become the normalising constant. That is,

$$P(x_t|\mathbf{y}_{,M}, \mathbf{x}^{t+1}, \theta) \propto P(x_t|\mathbf{y}_{,t}, \theta)P(x_{t+1}|x_t, \Pi). \quad (17)$$

We continue, in more detail, to show

$$\begin{aligned} P(x_t|\mathbf{y}_{,t}, \theta) &= P(x_t|\mathbf{y}_{,t-1}, \mathbf{y}(t), \theta) \\ &= \frac{P(x_t|\mathbf{y}_{,t-1}, \theta)P(\mathbf{y}(t)|x_t, \mathbf{y}_{,t-1}, \theta)}{P(\mathbf{y}(t)|\mathbf{y}_{,t-1}, \theta)} \\ &\propto P(x_t|\mathbf{y}_{,t-1}, \theta)f(\mathbf{y}(t)|\mathbf{y}_{,t-1}, \theta_{x_t}). \end{aligned} \quad (18)$$

By the law of total probability and Lemma 3.1, we have

$$\begin{aligned} P(x_t|\mathbf{y}_{,t-1}, \theta) &= \sum_{l=1}^a P(x_t|x_{t-1} = l, \mathbf{y}_{,t-1}, \theta)P(x_{t-1} = l|\mathbf{y}_{,t-1}, \theta) \\ &= \sum_{l=1}^a P(x_t|x_{t-1} = l, \Pi)P(x_{t-1} = l|\mathbf{y}_{,t-1}, \theta) \end{aligned}$$

and consequently from (17) and (18),

$$\begin{aligned} P(x_t|\mathbf{y}_{,M}, \mathbf{x}^{t+1}, \theta) &\propto P(x_{t+1}|x_t, \Pi) \left[\sum_{l=1}^a P(x_t|x_{t-1} = l, \Pi)P(x_{t-1} = l|\mathbf{y}_{,t-1}, \theta) \right] \\ &\quad \times f(\mathbf{y}(t)|\mathbf{y}_{,t-1}, \theta_{x_t}). \end{aligned}$$

This is initialized at $t = u_0$ by setting $P(x_0|\mathbf{y}_{,M}, \theta) = P(x_{u_0}|r)$ to be the same as the Dirichlet prior on $D(\alpha_{01}, \dots, \alpha_{0a})$. \square

3.3.2 Sampling from $P(\theta|\mathbf{x}, \mathbf{y})$

Define

$$\begin{aligned} n_{0,l} &:= I_{\{x_{u_0}\}}(l), \quad l = 1, \dots, a, \\ n_{gh} &:= \sum_{t=u_0+1}^M I_{\{(x_{t-1}, x_t)\}}(g, h), \quad g, h = 1, \dots, a, \\ n_{l,jk} &:= \sum_{i=1}^n \sum_{t=u_i+1}^{m_i} I_{\{(y_{i,(t-1)}, y_{i,t}, x_t)\}}(j, k, l), \end{aligned}$$

for $j, k = 1, \dots, J$, $l = 1, \dots, a$.

Then from Equation (13), combined with the Dirichlet priors, it can be seen that $P(\theta|\mathbf{x}, \mathbf{y})$ can be simulated separately and independently for r , Π and all the

P 's as follows:

$$\begin{aligned}
r|\mathbf{x}, \mathbf{y} &\sim D(\alpha_{0,1} + n_{0,1}, \dots, \alpha_{0,a} + n_{0,a}) \\
\pi_{11}, \dots, \pi_{1a}|\mathbf{x}, \mathbf{y} &\sim D(\alpha_{11} + n_{11}, \dots, \alpha_{1a} + n_{1a}) \\
&\vdots \\
\pi_{a1}, \dots, \pi_{aa}|\mathbf{x}, \mathbf{y} &\sim D(\alpha_{a1} + n_{a1}, \dots, \alpha_{aa} + n_{aa}).
\end{aligned}$$

For the parameters for the observed process in each regime $l = 1, \dots, a$, this yields

$$\begin{aligned}
p_{l,11}, \dots, p_{l,1J}|\mathbf{x}, \mathbf{y} &\sim D(\alpha_{l,11} + n_{l,11}, \dots, \alpha_{l,1J} + n_{l,1J}) \\
&\vdots \\
p_{l,J1}, \dots, p_{l,JJ}|\mathbf{x}, \mathbf{y} &\sim D(\alpha_{l,J1} + n_{l,J1}, \dots, \alpha_{l,JJ} + n_{l,JJ}).
\end{aligned}$$

3.3.3 Extra permutation step

To improve the convergence properties of the DA algorithm at each iteration of the Gibbs sampler, we conduct a random permutation of the labels, as detailed in Frühwirth-Schnatter (2001). Here we show that this extra step is justified and is optimal in the sense of Hobert and Marchev (2008). What they denote by \mathbf{Y} is our \mathcal{X}^m and what they denote by \mathbf{X} is our Θ .

From (14) it can be seen that the posterior of interest, $f(\theta|\mathbf{y})$, is the θ -marginal density of $f(\mathbf{x}, \theta|\mathbf{y})$; i.e., $f(\theta|\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}^m} f(\mathbf{x}, \theta|\mathbf{y}) = \int_{\mathcal{X}^m} f(\mathbf{x}, \theta|\mathbf{y}) \mu(d\mathbf{x})$, where μ is the counting measure on \mathcal{X}^m . Clearly, this form of the target density allows for construction of the optimal Haar PX-DA algorithm, as defined in Hobert and Marchev (2008). Here we will show that the random permutation sampler of Frühwirth-Schnatter (2001) is a specific case of the Haar PX-DA algorithm.

In our case $\mathcal{X} = \{1, \dots, a\}$ and the space \mathcal{X}^m is the m -tuple product of \mathcal{X} with

itself:

$$\begin{aligned}\mathcal{X}^m &= \{(x_1, \dots, x_m) : x_i \in \mathcal{X}, i = 1, \dots, m\} \\ &= \{1, \dots, a\}^m\end{aligned}$$

where $m = M - u_0 + 1$. Notice that \mathcal{X} , as any other discrete space, is a particularly simple topological space, equipped with the discrete metric

$$d(\mathbf{x}, \tilde{\mathbf{x}}) = \begin{cases} 0, & \mathbf{x} = \tilde{\mathbf{x}} \\ 1, & \mathbf{x} \neq \tilde{\mathbf{x}} \end{cases}, \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}^m.$$

Any discrete space with discrete metric is separable and locally compact and all subsets are open (and closed). In addition, we need to define a group action on \mathcal{X}^m . Let G be the symmetric group on the set \mathcal{X} ; i.e.,

$$G := S_{\mathcal{X}} := \{\text{permutations of } (1, \dots, a)\},$$

again equipped with the discrete metric. Finally, define the group action on \mathcal{X}^m as

$$F(g, \mathbf{x}) = g\mathbf{x} = (g(x_1), \dots, g(x_m)),$$

which just permutes the values of the labels. (For example, if $J = 3$, $\mathbf{x} = (3, 2, 1, 1) \in \mathcal{X}^4$, and $g = (2, 3, 1)$, then $g\mathbf{x} = (1, 3, 2, 2)$.) Then for the identity permutation e , we have $e\mathbf{x} = (x_1, \dots, x_m) = \mathbf{x}$, and for any two permutations g_1 and g_2 , $(g_1 g_2)\mathbf{x} = g_1(g_2\mathbf{x})$. As a multiplier we take $\chi(g) = 1$, $\forall g \in S_{\mathcal{X}}$. Then, obviously, $\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$. It is easy to see that μ is relatively invariant,

since for any integrable function h we have:

$$\begin{aligned} \chi(g) \int_{\mathcal{X}^m} h(g\mathbf{x})\mu(d\mathbf{x}) &= \sum_{\mathbf{x} \in \mathcal{X}^m} h(g\mathbf{x}) = \sum_{\mathbf{x} \in \mathcal{X}^m} h(\mathbf{x}) \\ &= \int_{\mathcal{X}^m} h(\mathbf{x})\mu(d\mathbf{x}). \end{aligned} \quad (19)$$

Note that (19) follows from the fact that $\mathcal{X}_g^m := \{g\mathbf{x} : \mathbf{x} \in \mathcal{X}^m\} = \mathcal{X}^m$, since for any $\mathbf{x} \in \mathcal{X}^m$, there exists $\tilde{\mathbf{x}} = g^{-1}\mathbf{x}$ such that $g\tilde{\mathbf{x}} = \mathbf{x}$. The last piece of the Haar PX-DA setup is the function $j : g \times \mathcal{X}^m \rightarrow \mathbb{R}_+$, defined as $j(g, \mathbf{x}) = 1$. Also notice that the Haar measure on the symmetric group is the counting measure $\nu(dg)$. After all this, the Haar PX-DA will iterate between the following three steps to move from the current θ' to the next θ :

1. $\mathbf{x} \sim P(\mathbf{x}|\theta', \mathbf{y})$;
2. $g \sim \chi(g)P(g\mathbf{x}|\mathbf{y}) = P(g(x_1), \dots, g(x_m)|\mathbf{y})$. Set $\mathbf{x}' = g\mathbf{x}$;
3. $\theta \sim f(\theta|\mathbf{x}')$.

Notice that step 2. in the above algorithm reduces to choosing a random permutation on \mathcal{X} with probability $\frac{1}{a!}$, as long as the \mathbf{x} -marginal density $P(\mathbf{x}|\mathbf{y})$ is symmetric under permuting the values of x_1, \dots, x_m . This may sound very restrictive and impractical but it is equivalent to the well-known random permutation sampler developed in Frühwirth-Schnatter (2001) and used to remedy convergence problems of MCMC algorithms used in mixture models and HMM's. The condition that $P(\mathbf{x}|\mathbf{y})$ is symmetric under permutations is easy to check and is generally satisfied under model (10) combined with Dirichlet priors with equal parameters. Since the above derivations show that the extra step of randomly permuting the labels is obtained under the conditions of Theorem 4 from Hobert and Marchev (2008), the resulting Markov chain will have smaller asymptotic variance and bet-

ter mixing than the regular DA algorithms. Further theoretical developments about the extra step can be found in Khare and Hobert (2011) and Roy (2012).

3.3.4 Post-processing algorithm

In general, the posterior densities of each of the parameters can have a different modes, meaning that simply taking the posterior means would not yield a useful estimate. Since we cannot directly use the posterior mean as an estimate, it is necessary to either impose artificial identifiability constraints to each of the components in the model or to employ a post-processing algorithm to ensure that the labels of the hidden states are consistent for all iterations. Stephens (2000) details how the first approach generally fails to deal with the problem of label switching in mixtures. We therefore use the post-processing algorithm of Stephens (2000) to ensure consistency in the labeling of the components in our model.

This algorithm attempts to relabel the parameters for each iteration $k = 1, \dots, N$ so as to minimize the expected loss under a class of loss functions. These loss functions are defined in the decision theoretic framework outlined in Stephens (2000). The specific version we use is as described in the relabelling algorithm in Figure 3 on page 247 of Boys and Henderson (2002). Initially for each iteration $k = 1, \dots, N$ of the original Gibbs sampler, the post-processing algorithm seeks out the particular permutation of the labels that minimizes the number of labels that differ from the selected labels of the previous iteration $k - 1$. Then this permutation of the labels is applied to the parameters that were sampled at iteration k . Finally, for the selected labels at each time point t for $t = u_0, \dots, M$, the label to be selected is the mode of all of the labels at t that have been selected over all of the previous iterations $1, \dots, k$.

3.4 Simulation studies

In this section we illustrate the performance of our algorithm on simulated data. It consists of two parts - estimating parameters and estimating the whole posterior density. We also compare various performance measures of the Haar PX-DA estimation procedure to the DCMM estimation procedure outlined in Berchtold (2002). We used R to program our algorithm and we used MARCH 3.0 for Berchtold's method.

3.4.1 Parameters Estimation

In this subsection we opted for large sample sizes, so that the parameter estimates, calculated as the posterior means, will be close to the values used to simulate the data. We tried many combinations of a and J and all of them performed similarly. Here we present the result for one of these settings.

We simulated data from the multivariate DCMM using equations (11) and (10) with $n = 500$ random vectors, a total of $m = 300$ possible time points, $a = 2$ hidden states and $J = 4$ possible observed states. For $i = 1, \dots, n$, the random vectors \mathbf{Y}_i , were simulated with different starting times u_i and ending times m_i , selected uniformly from 2 to m . The data were simulated using the following transition matrices:

$$\begin{aligned}
\Pi^* &= \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}, P_1^* = \begin{pmatrix} 0.80 & 0.16 & 0.03 & 0.01 \\ 0.15 & 0.65 & 0.15 & 0.05 \\ 0.05 & 0.10 & 0.60 & 0.25 \\ 0.01 & 0.01 & 0.01 & 0.97 \end{pmatrix}, \\
P_2^* &= \begin{pmatrix} 0.80 & 0.15 & 0.04 & 0.01 \\ 0.10 & 0.85 & 0.04 & 0.01 \\ 0.01 & 0.15 & 0.75 & 0.09 \\ 0.01 & 0.01 & 0.01 & 0.97 \end{pmatrix}.
\end{aligned} \tag{20}$$

In this case Π was motivated by our real data example, presented in the next section, and P_1^* and P_2^* were intentionally chosen to be similar to each other so that it would be difficult for the algorithm to distinguish between the two regimes. The α 's of each Dirichlet prior were all set to 1.

We ran our Haar PX-DA procedure for 40,000 iterations. Even with such a high number of iterations and with such large values for n and m , this took only 22 minutes on a Pentium E6550 processor at 2.33 Ghz using our non-optimised R code. The estimates of the parameters, Π , P_1 and P_2 , were then obtained as the posterior means of the corresponding distribution (after the post-processing was

applied). The results for $a = 2, J = 4$ were as follows:

$$\begin{aligned}
 \hat{\Pi} &= \begin{pmatrix} 0.4323 & 0.5677 \\ 0.1725 & 0.8275 \end{pmatrix}, \\
 \hat{P}_1 &= \begin{pmatrix} 0.7935 & 0.1610 & 0.0346 & 0.0109 \\ 0.1508 & 0.6551 & 0.1485 & 0.0456 \\ 0.0535 & 0.1047 & 0.6114 & 0.2304 \\ 0.0082 & 0.0135 & 0.0093 & 0.969 \end{pmatrix}, \\
 \hat{P}_2 &= \begin{pmatrix} 0.8064 & 0.1407 & 0.0416 & 0.0112 \\ 0.0991 & 0.8465 & 0.0437 & 0.0107 \\ 0.0087 & 0.1465 & 0.7477 & 0.0971 \\ 0.0108 & 0.0107 & 0.0111 & 0.9675 \end{pmatrix}.
 \end{aligned} \tag{21}$$

We notice that upon comparison to the true parameter values in (20), the estimated values in (21) are remarkably accurate, especially since the true values of the two components driving the simulated data were quite similar.

The eventual level of accuracy of the parameter estimates clearly depends on the level of separation between the two components. We conduct this simulation with similar components, such as in the real example in Section 3.5, to illustrate the approximate accuracy of the parameter estimates that can be obtained with an appropriately large amount of data.

It is also worth mentioning that the posterior estimates obtained without the extra permutation step were not very different from those obtained with the permutation step. The reason for this is the use of the post-processing algorithm, which plays a similar role to that of the extra step - it reduces the autocorrelation of the Markov chain. However, as we demonstrate in the next subsection, the extra permutation step greatly improves the overall posterior density estimates.

Furthermore, while the extra permutation has been completely justified theoretically in Hobert and Marchev (2008), there is very little proved in the literature about the post-processing algorithm. Lastly, it should be mentioned that the post-processing benefits come at a very “steep” price in terms of computation time - roughly the same as to run the DA algorithm, whereas the Haar PX-DA’s extra step just takes a couple of extra seconds overall.

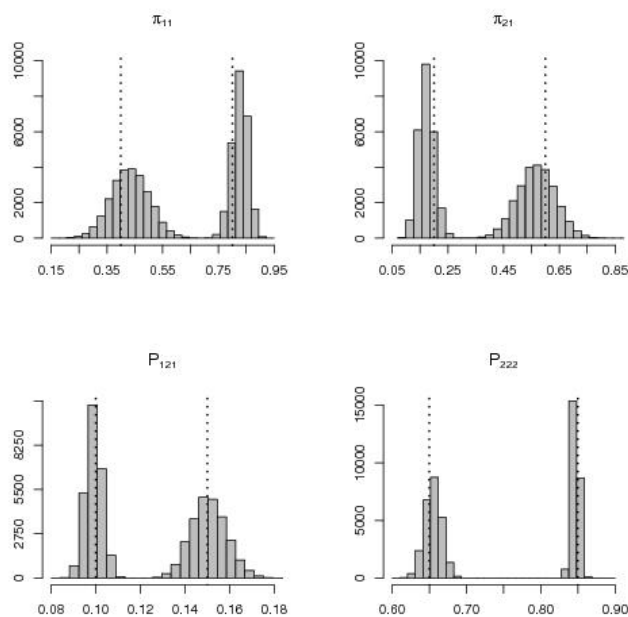


Figure 3: Histograms of some of the posterior densities, before post-processing algorithm is applied. The dotted lines in the graphs are the locations of the true values used to simulate the data

In Figure 3, we see that, before post-processing is applied, there is a clear bi-modal form to the posterior densities. The dotted lines in the graphs are at the true values used to simulate the data. We see in Figure 4 that after the post-processing algorithm of Stephens (2000) as described in Boys and Henderson

(2002) is run, the bi-modal form of the posterior densities is no longer present, so the mean of the posterior samples over all iterations would be a suitable estimate for each parameter in the model.

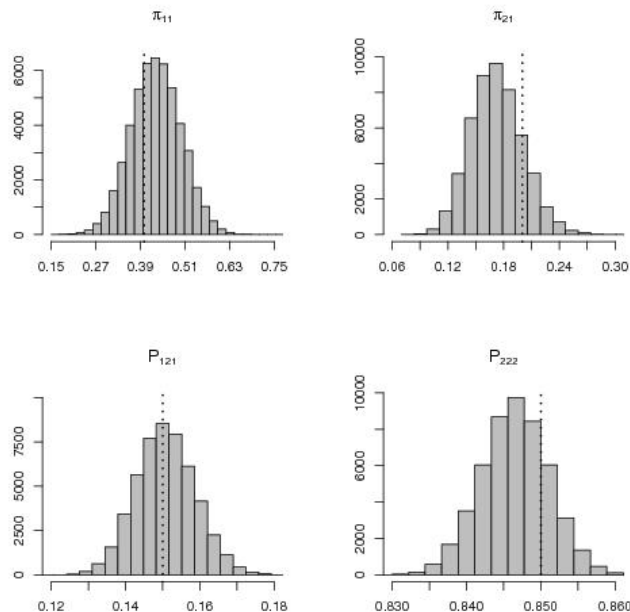


Figure 4: Histograms of some of the posterior densities, after post-processing algorithm is applied

3.4.2 Posterior Density Estimation

To illustrate the benefit of employing the extra permutation step in our Haar PX-DA procedure compared to the standard DA (no extra step) procedure in obtaining estimates for the posterior densities of each of the parameters, we conducted further simulations with a smaller amount of data and a smaller number of iterations, varying the number of hidden components a and the number of observed states J .

The metric for comparing the posterior density estimates $\hat{f}(x)$, from the algorithms with and without the extra permutation step, to the “true” posterior $f(x)$ was the *integrated squared error* (ISE), defined as $\text{ISE} = \int [\hat{f}(x) - f(x)]^2 dx$. We simulated data from the multivariate DCMM using equations (11) and (10) with $n = 30$ random vectors, and a total of $m = 400$ possible time points. The number of hidden states a and the number of observable states J were varied between 2 and 4 to investigate the efficiency of our algorithm with 10,000 iterations. Various transition matrices were used to simulate the data in each case. Since we know in the theory from Hobert and Marchev (2008) that the Haar PX-DA procedure is at least as efficient as the standard DA procedure in this setting, the posterior distributions are assumed to converge towards those obtained by using the Haar PX-DA procedure with 300,000 iterations, which we will call the “true” posterior distribution. No post-processing was used as we did not need the posterior means in this study. Table 1 summarizes all of the simulations, where the sum of the ISE for the elements in each matrix is displayed for comparison.

a	J	Perm	Π	P_1	P_2	P_3
2	3	Yes	0.05147	0.34813	0.30698	–
2	3	No	2.65665	18.70867	19.00292	–
3	3	Yes	0.20895	2.28660	2.41501	2.33394
3	3	No	11.0752	153.974	31.3304	25.1975
3	4	Yes	2.03327	3.90679	3.52574	3.03290
3	4	No	14.2823	237.698	43.7094	41.4116

Table 1: Sum of the ISE for each element of the transition matrices for various simulations, using the estimation procedure with and without the extra permutation step, denoted by Perm=Yes and Perm=No respectively.

We see in Table 1 that by using our Haar PX-DA procedure, with the extra permutation step, we can obtain much better approximations of the posterior density estimates for the parameters than with the standard DA procedure with

the same number of iterations. Although the results display the sum of the ISE for the elements in each matrix in the simulations, it must be noted that the Haar PX-DA procedure was superior for estimating parameters from the posterior density for every parameter in the model. We also note that if we use the standard DA with 300,000 iterations as a benchmark, then the results for the case $a = 2, J = 3$ are almost unchanged. However, in the other two cases the standard DA was still stuck in one of the modes even after 300,000 iterations.

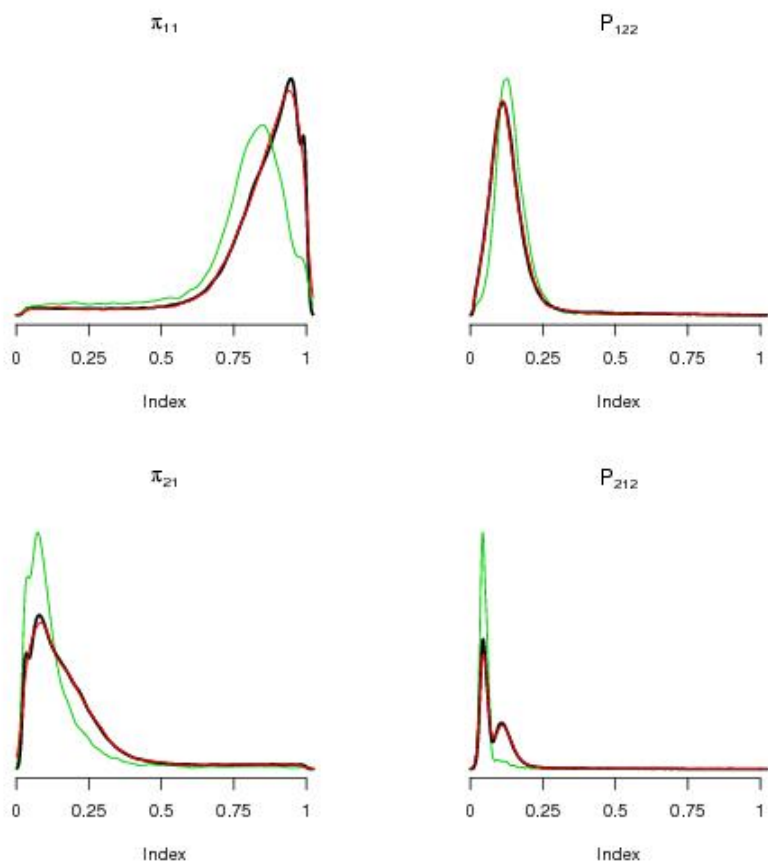


Figure 5: Posterior density estimates for parameters in the simulation where $n = 30$, $m = 400$, $a = 2$, and $J = 3$. The true posterior density is in black, the estimate with the permutation step is red and the estimate without the permutation step is green

We further illustrate the difference between the estimation procedures in Figure 5, which displays the estimated posterior densities obtained by our Haar PX-DA procedure, the standard DA procedure and the true posterior estimates for a selection of the parameters. These parameters are from the simulation where $n = 30$, $m = 400$, $a = 2$ and $J = 3$, corresponding to the first two rows of Table 1. It is clear in Figure 5 that the standard DA procedure is not as efficient in sampling from all modes of the posterior distribution as the Haar PX-DA procedure, which employs the extra permutation step. In some instances, such as in P_1 for the $a = 2$, $J = 3$ simulation, the standard DA procedure estimates the posterior densities very poorly, compared to our Haar PX-DA procedure. This is due to the fact that the posterior densities for the parameters in such mixture models are multi-modal due to the invariance of the likelihood under label permutations. However, it can so happen that the standard DA procedure predominantly samples from one of the modes, leading to a poor estimate of the posterior density.

3.4.3 Comparison to MARCH 3.0 software

In this subsection we compare the performance of our Haar PX-DA procedure to the standard DCMM estimation procedure outlined in Berchtold (2002). The algorithm for our estimation procedure was written using the R programming language. The procedure in Berchtold (2002) has been implemented into a publicly available software package called **MARCH 3.0**.

The first comparison between the two procedures was done using a real dataset, presented in Azzalini and Bowman (1990). This dataset consists of a sequence of 299 successive observations of either long or short duration eruptions of the Old Faithful geyser in Yellowstone Park, USA, during the period 1-15 August, 1985. Using 100 expectation-maximisation (EM) iterations, the **MARCH 3.0** software can fit a first order DCMM in under 10 seconds on a standard PC to this dataset with

$a = 2$, $J = 2$, $n = 1$ and $m = 299$. The log-likelihood is -124.421 and the BIC is 271.582 . The first order DCMM was also fit using our Haar PX-DA procedure with 100 iterations in R. This was also completed in under 10 seconds on a standard PC and yielded a comparable fit, with a log-likelihood of -128.151 and a BIC of 290.424 . Upon calculating the log-likelihood of the observations, conditional on the estimated path of hidden states, the Haar PX-DA procedure yields a slightly better result of -33.792 compared to -40.779 from using the `MARCH 3.0` software.

We also compared the two estimation procedures on a simulated data set from a first order DCMM with known parameters. With $a = 2$, $J = 2$, $n = 1$ and $m = 300$, the true parameters were as follows:

$$\begin{aligned}
 \Pi^* &= \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix}, \\
 P_1^* &= \begin{pmatrix} 0.85 & 0.15 \\ 0.2 & 0.8 \end{pmatrix}, P_2^* = \begin{pmatrix} 0.4 & 0.6 \\ 0.75 & 0.25 \end{pmatrix}.
 \end{aligned}$$

We then estimated a first order DCMM on this simulated dataset, using both the Haar PX-DA procedure and the `MARCH 3.0` software. The performance results were as follows:

	<code>MARCH 3.0</code>	Haar PX-DA
SSE	0.1168266	0.0398213
Cond. Log-Likelihood	-150.1	-155.5
Log-Likelihood	-201.1	-191.8
BIC	436.4	417.8
Regime Error Rate	0.179	0.179
Computation Time	63 seconds	55 seconds
Software	<code>MARCH 3.0</code>	R

Table 2: Performance comparison between our Haar PX-DA procedure to the

standard DCMM estimation procedure outlined in Berchtold (2002) on a simulated dataset.

Note: The *sum of squared errors* (SSE) is just the sum of squared differences of all parameters used to simulate the data and their estimates.

The results in Table 2 show that our estimation method not only allows further practical applications due to its flexibility with data types but it can also improve the accuracy and fit of parameter estimates in some cases, compared to the procedure presented in Berchtold (2002). This is of great importance to practitioners who constantly deal with irregular data sets and also require accurate model estimation.

3.5 Applications to Standard and Poor’s credit rating data

In this section we use our model on a real dataset. The data to be analysed are the monthly Standard & Poor’s credit rating transitions of $n = 3,918$ firms, ranging from the 1st of January 1981 to 1st of January 2010. Since the economic conditions vary across different industries, our model is most meaningful when applied to similar firms only; in this case all firms are financial institutions and insurance companies.

We decided to fit a model with $a = 2$ since it is a well-accepted theory that the economic cycle fluctuates between two regimes: “expansion” and “contraction”. The value of J is 10, corresponding to 10 levels of credit ratings: AAA, AA, A, . . . , D. We ran our MCMC algorithm for 50,000 iterations and the results are shown in Table 3 with Component 1, corresponding to “contraction” and Component 2 to “expansion”:

Hidden Matrix		
	Component 1	Component 2
Component 1	0.6897	0.3103
Component 2	0.1083	0.8917

Component 1					
	AAA	AA	A	BBB	BB
AAA	0.9859	0.0126	0.0003	0.0003	0.0002
AA	0.0006	0.9793	0.0196	0.0002	0.0001
A	0	0.0012	0.9881	0.0098	0.0005
BBB	0.0001	0.0004	0.0028	0.9856	0.0094
BB	0.0002	0.0002	0.0003	0.0044	0.9745
B	0.0003	0.0003	0.0005	0.0003	0.006
CCC	0.0012	0.0012	0.0012	0.0035	0.0024
CC	0.011	0.0111	0.011	0.011	0.011
C	0.052	0.0516	0.052	0.0523	0.0517
D	0.0103	0.0103	0.0102	0.0103	0.0102
	B	CCC	CC	C	D
AAA	0.0001	0.0001	0.0001	0.0001	0.0001
AA	0.0001	0.0001	0.0001	0.0001	0.0001
A	0.0001	0	0	0	0.0001
BBB	0.0006	0.0003	0.0002	0.0001	0.0006
BB	0.0165	0.0022	0.0008	0.0002	0.0009
B	0.9651	0.017	0.0038	0.0003	0.0066
CCC	0.0096	0.9298	0.0147	0.0023	0.0343
CC	0.0111	0.0282	0.6694	0.011	0.2252
C	0.052	0.0522	0.0527	0.4295	0.1541
D	0.0102	0.0103	0.0103	0.0102	0.9077

Component 2					
	AAA	AA	A	BBB	BB
AAA	0.9944	0.0049	0.0003	0.0001	0.0001
AA	0.0005	0.9948	0.0044	0.0002	0
A	0	0.0029	0.9943	0.0024	0.0001
BBB	0	0.0003	0.0053	0.9912	0.0027
BB	0	0.0003	0.0003	0.0072	0.9863
B	0.0001	0.0001	0.0002	0.0008	0.0096
CCC	0.0004	0.0004	0.0004	0.0004	0.0017
CC	0.0067	0.0067	0.0066	0.0066	0.0067
C	0.0338	0.034	0.0337	0.0338	0.0336
D	0.0037	0.0037	0.0037	0.0037	0.0038
	B	CCC	CC	C	D
AAA	0	0	0	0	0
AA	0	0	0	0	0
A	0	0	0	0	0.0001
BBB	0.0003	0	0	0	0.0001
BB	0.005	0.0004	0.0001	0	0.0004
B	0.9834	0.004	0.0007	0.0001	0.001
CCC	0.0206	0.9568	0.0028	0.0004	0.015
CC	0.0201	0.0225	0.8102	0.0067	0.1073
C	0.0339	0.0337	0.0338	0.6609	0.0689
D	0.0037	0.0037	0.0037	0.0037	0.9664

Table 3: Estimated transition probabilities from the Standard and Poor's credit rating data (that is, $\hat{\pi}$, \hat{P}_1 , \hat{P}_2).

Note that the transition probabilities between the hidden economic cycles are estimated remarkably close to the well-established transition probabilities given in Bangia et al. (2002). Note also that the two estimated transition matrices for the observed process under the two different hidden regimes are quite different to the

transition matrix obtained by assuming the simple Markov chain model in Table 4.

Simple Markov Chain Transition Matrix					
	AAA	AA	A	BBB	BB
AAA	0.9929	0.0068	0.0002	0.0001	0
AA	0.0005	0.9914	0.0079	0.0001	0
A	0	0.0025	0.9931	0.004	0.0002
BBB	0	0.0003	0.0048	0.9903	0.0041
BB	0	0.0002	0.0002	0.0064	0.9842
B	0	0.0001	0.0002	0.0006	0.0086
CCC	0	0	0	0.0006	0.0013
CC	0	0	0	0	0
C	0	0	0	0	0
D	0	0	0	0	0
	B	CCC	CC	C	D
AAA	0	0	0	0	0
AA	0	0	0	0	0
A	0	0	0	0	0.0001
BBB	0.0004	0.0001	0	0	0.0002
BB	0.0076	0.0007	0.0002	0	0.0004
B	0.9803	0.0068	0.0013	0.0001	0.0021
CCC	0.017	0.9547	0.0055	0.0003	0.0205
CC	0.009	0.0179	0.8161	0	0.157
C	0	0	0	0.8966	0.1034
D	0	0	0	0	1

Table 4: Estimated transition probabilities using a simple Markov chain (SMC) model from the Standard and Poor's credit rating data

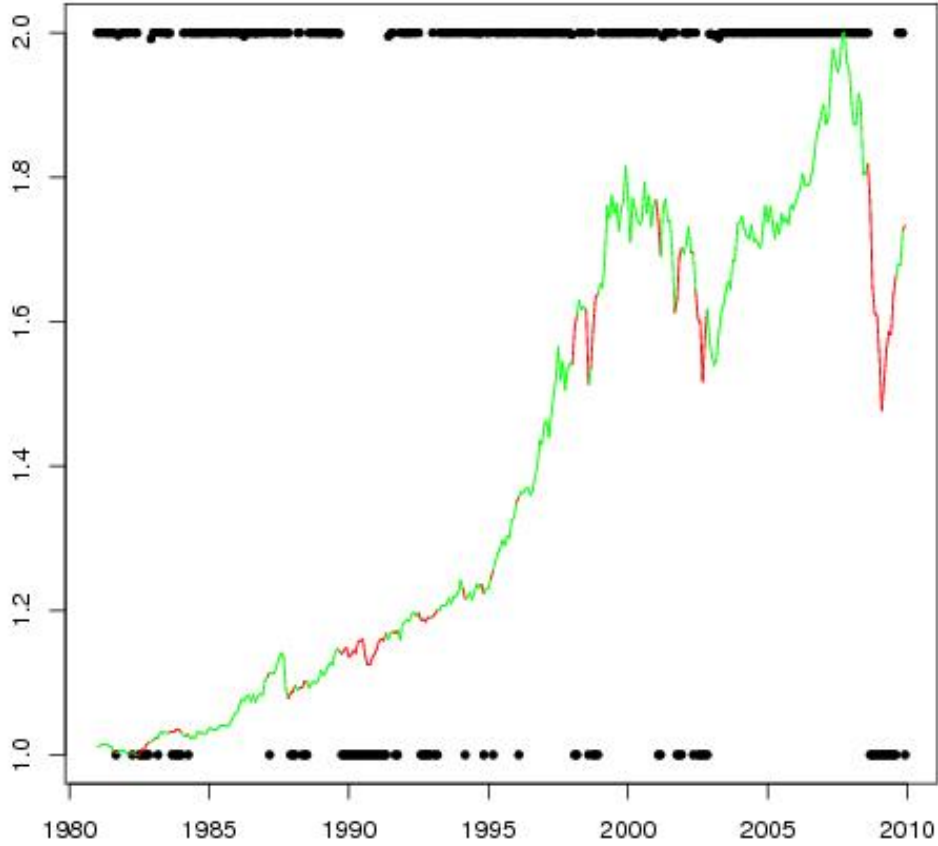


Figure 6: Mean value of selected components over all iterations with S&P data along with the Dow Jones Industrial Average closing price linearly scaled for comparison. The DJIA is red for periods that our model estimated as a “contraction” and green for periods that our model estimated as an “expansion”

We also draw attention to Figure 6 representing the mean estimated regime of the hidden process for each time point superimposed with the Dow Jones Industrial Average (DJIA) index. In particular, it can be seen that Component 1 (corresponding to downturns of the economy) was selected, at times correspond-

ing to well-known economic downturns in the financial services industry. Similarly, Component 2 tends to be selected during growth periods. More specifically, in the last 30 years of finance history there are a number of notable events. These include:

1. Savings and Loan Crisis (Early 1980s)
2. Black Monday (October 1987)
3. Economic Recession (1990-1991)
4. Asian financial crisis (1997-1998)
5. Dot-com bubble (1995-2001)
6. Dot-com bust and September 11 terrorist attacks after effects (2001-2002)
7. United States housing bubble (2002-2008)
8. Global Financial Crisis (2008-2010).

It must be noted that although the parameters of the model are estimated from the observed credit rating data, the hidden states sampled at each time point bear a remarkable resemblance to the above-mentioned events in the history of finance. Therefore, the intuitive expectation of the effect that a *hidden economic state* has on the migration behavior of a large portfolio of firms is captured in the model. This is of extreme importance for practitioners before confidence can be placed in a model's forecasting abilities.

In comparison to our results, Bangia et al. (2002) considers a DCMM-type model applied to credit modeling, where the hidden states are *known* to be either an expansion or a contraction of the economy. The observed credit ratings of firms are driven by a process, which switches between two Markov chains (one for each state of the economy). The hidden states are directly observed from macro-economic data and the parameters of the hidden transition matrix are estimated

independently of the observed credit ratings. However, this model is not a single integrated HMM but rather two separate simple Markov chain models that may not be able to detect true heterogeneity in transition behavior. We were able to estimate the underlying transition matrix of the hidden process quite well without relying on observations of economic variables but entirely from the observed credit ratings data.

Since practitioners are often concerned with the one year default probabilities of a portfolio of firms (i.e. migration probabilities into the *D* credit rating category), we now utilise the estimated model parameters and the simple Markov model in Table 4 to predict the expected proportion of defaults in each credit rating from *AAA* to *C*, after 12 months. This method is employed in Jarrow et al. (1997) and can potentially incorporate a model for the term structure of default rates. The results are displayed in Table 5 for the following scenarios:

1. Firms migrate according to the estimated hidden Markov model, conditional on the first month migrating by Component 1 (C1).
2. Firms migrate according to the estimated hidden Markov model, conditional on the first month migrating by Component 2 (C2).
3. Firms migrate according to the Simple Markov Chain model (SMC).
4. Firms migrate according to the estimated hidden Markov model, conditional on all of the next 12 months migrating by Component 1 (Worst).
5. Firms migrate according to the estimated hidden Markov model, conditional on all of the next 12 months migrating by Component 2 (Best).

Annual Expected Default Rates					
	C1	C2	SMC	Worst	Best
AAA	0.0017	0.0002	0	0.0025	0
AA	0.0017	0.0002	0.0001	0.0024	0
A	0.0015	0.0012	0.0012	0.0017	0.0012
BBB	0.0074	0.002	0.0027	0.0103	0.0013
BB	0.0203	0.0072	0.0079	0.0285	0.0057
B	0.0901	0.0269	0.039	0.1211	0.0178
CCC	0.3121	0.1844	0.2253	0.3654	0.1636
CC	0.6895	0.5672	0.7936	0.716	0.5407
C	0.3591	0.2785	0.7301	0.3772	0.2645

Table 5: Forecasted 12-month default rates from the Standard and Poor’s data. We see in Table 5 that with the DCMM, we are able to provide more information about future default rates of firms with each of the Standard and Poor’s credit ratings, than with the traditional SMC approach. Since the DCMM has estimated two migration matrices for the observed states, one with favourable (stable) migrations and the other with unfavourable (unstable) migrations, we are able to obtain lower and upper bounds of yearly default rate outcomes for each of the ratings, denoted by the Worst and Best columns of Table 5 respectively. Note that the default rates in the columns for C1, C2 and SMC all lie within the forecasted range, apart from the CC and C credit ratings, where there are not very many historical observations.

3.6 Discussion

3.6.1 Conclusions

In the practice of modelling multivariate panel data-sets with a large number of independent random variables observed over many time points, it is often clear that

there is a need to model the time heterogeneity of the migrations of observations. The particular source of this heterogeneity is particularly difficult to estimate and model directly from the observed data. Allowing a hidden state driven by a Markov process to govern the Markov regime driving the observations at each time point can often provide an intuitive way to capture the varying dynamics of the random variable's migrations over time. We have shown in this chapter that a double chain Markov model that has this capability can be estimated efficiently from the observed panel dataset and is not reliant on pin-pointing the exact causes of the hidden effect on the migration dynamics.

The application of modeling the credit ratings of a large portfolio of firms over time is of particular relevance to the contributions of this chapter. The fact that each firm can enter the portfolio at different times, as well as leave the portfolio at different times, means that at each time point, we have a vector of observations that are possibly of different lengths. We have shown that the efficient computational techniques presented in Chib (1996) can be extended to a new sampling algorithm that applies to the exact type of data set that we are dealing with here.

There is a subtle dependence between the random variables of the panel dataset. This dependence is as a result of the fact that all random variables migrate at each time point according to the same transition matrix that is selected by the hidden Markov process. Upon looking back at the selected hidden states in our application to real world credit rating data, over the observed time period, we note that we are able to capture significant events in the history of financial institutions and insurance companies. Therefore the model provides a strong fit to our large data set, with only a small increase in the number of parameters over the simple Markov chain model that does not capture any economic effects at all.

3.6.2 Further Research

This study has brought to light that there are areas of further research that should be explored. In our study, we allowed all parameters to be estimated with a non-zero probability. If practitioners need to impose restrictions on some of the parameters, by forcing an observable state to be an absorbing state for example, then techniques such as the method of Lagrange multipliers should be employed to ensure that the update of the parameter values at each iteration of the Gibbs sampler algorithm is accurate.

Another possible avenue for future research would be to incorporate a more complicated data structure between the random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, rather than assuming that they are conditionally independent given the hidden process \mathbf{X} . Perhaps there could be different groups of \mathbf{Y}_i 's, corresponding to different types of observations (e.g., different types of companies are affected differently during the different periods of economic cycles).

It should also be pointed out that although we allow for different start and end points of the random vectors $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, we assume there are no missing data. That is, our method doesn't work correctly with vectors observed with gaps.

On the theoretical side, it would be interesting (and challenging) to develop more specific results about the convergence rate of the modified Haar PX-DA algorithm and to compare it to the convergence rate of the regular DA without the extra permutation step.

4 Mixtures of Markov chains

In this chapter, we continue with the application to modelling the non-homogeneous credit rating dynamics of firms; however, instead of allowing for different regimes in the data over time, we will instead allow for different regimes to apply to different parts of the population. That is, we approach the problem by detecting and modelling the non-homogeneity amongst the population rather than changing dynamics over time. Although the estimation procedure used in this chapter (the E-M Algorithm of Dempster et al. (1977)) is similar to the Bayesian MCMC estimation methods employed in the previous two chapters, we do not focus on the properties of the Markov chains used in the estimation. Instead, we will focus on the Markov chains used in the parameters for the observed data to develop the theory behind testing whether a single homogeneous Markov chain model is appropriate for the data.

A key difference to the Markov chains used to model the observed data in this chapter, as opposed to the previous chapter is that we are studying continuous time Markov chains here compared to the previously studied discrete time models. The choice of model lies with the type of data studied (whether it is observed at discrete time points or if it is continuously observed) but there are some important differences to the theory, discussed in Frydman (2005). We first introduce our notation and some key concepts around continuous time discrete state Markov chains before introducing the concept of mixtures of continuous-time Markov chains and some key considerations for testing between 1 and 2 mixture components. We conduct a parametric bootstrap procedure to test for the presence of a mixture, which yields results that throw into doubt the claim from Frydman (2005) that we can use standard theory to apply a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the 1 component and 2 component mixture models. This motivates us to adapt the theory of Fukumizu

(2003) to our case to prove the divergence of the log-likelihood ratio test statistic for the test between 1 and 2 component mixture components. Finally, we look at the most simple case of 2-state Markov chain components, which each have the second state being an absorbing state, which directly applies to a *default vs. non-default model* in our application (grouping all *non-default* credit ratings into one category) and we derive a theorem for the exact limiting distribution of the log-likelihood ratio test statistic.

4.1 The continuous-time Markov chain

Suppose we have a discrete state, continuous-time Markov chain $\{X(t); 0 \leq t \leq T\}$ with $T < \infty$, that takes values amongst the discrete states in the set $\{1, \dots, w\}$ over time. Furthermore, when certain conditions are satisfied (see (22) and (23) below), $X(t)$ takes values in the space of step functions $x(\cdot) \in \mathcal{X}$ with a finite number of jumps between the discrete states over the fixed observation window $[0, T]$.

We let $\mathbf{d} = (d_1, \dots, d_w)$ be the initial state distribution given by

$$d_i = P[X(0) = i]$$

for $i \in \{1, \dots, w\}$. We also let $P(s, t)$ be the transition matrix with $(i, j)^{\text{th}}$ element

$$P_{ij}(s, t) = P[X(t) = j | X(s) = i]$$

with $s < t$ and $i, j \in \{1, \dots, w\}$. Then, following Albert (1962), we characterise $X(t)$ in terms of \mathbf{d} and $P(s, t)$.

Our process here has stationary transition probabilities,

$$P(s, t) = P(t - s) \tag{22}$$

so that transitions depend only on the difference between start and end times. We also have the infinitesimal generator matrix Q , which solves

$$\frac{\partial P(t)}{\partial t} = P'(t) = QP(t) = P(t)Q; \quad \text{such that } P(0) = I.$$

Thus, we define the continuous-time Markov chain in terms of the transition matrix

$$P(t) = e^{tQ} = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!},$$

so that our stochastic process forms right-continuous paths $x(\cdot) \in \mathcal{X}$. We may write the density as the product

$$f(x(\cdot)|\mathbf{d}, Q) = d_{x(0)}g(x(\cdot)|Q)$$

where $g(x|Q)$ is the conditional density of a continuous-time Markov chain with infinitesimal generator Q , given that the initial state $X(0) = x(0)$.

The elements of Q can be written more formally as

$$\begin{aligned} Q_{ii} &= -\sum_{j \neq i} Q_{ij} = \lim_{h \rightarrow 0} [(1 - P_{ii}(h))/h] \\ Q_{ij} &= \lim_{h \rightarrow 0} [(P_{ij}(h))/h] \quad \forall j \neq i. \end{aligned} \tag{23}$$

for $i, j \in \{1, \dots, w\}$.

A realisation $x(\cdot)$ of a discrete state, continuous-time Markov chain can be described by the sequence $\{z_0, z_1, \dots\}$ of states visited and the corresponding sequence of durations of time the process spends in each state $\{\tilde{t}_0, \tilde{t}_1, \dots\}$, so that $X(t) = z_j$ for $\sum_{b=0}^{j-1} \tilde{t}_b \leq t < \sum_{b=0}^j \tilde{t}_b$, with $j = 1, 2, \dots$ and $x(t) = z_0$ for $0 \leq t < \tilde{t}_0$. The corresponding random sequence $\{Z_j; j \in \mathbb{N}\}$ forms a discrete time Markov

chain with transition matrix J with $(i, j)^{\text{th}}$ element

$$J_{ij} = \begin{cases} \frac{Q_{ij}}{-Q_{ii}} & \text{for } i \neq j \text{ and} \\ 0 & \text{for } i = j. \end{cases}$$

Define

$$\lambda_i = -Q_{ii} = \sum_{j \neq i} Q_{ij}, \text{ for each } i \in \{1, \dots, w\}.$$

Note that given $Z_0 = z_0, Z_1 = z_1, \dots$, the sequence $\tilde{T}_0, \tilde{T}_1, \dots$ are conditionally independent random variables, with each \tilde{T}_j exponentially distributed with rate

$$\frac{1}{\lambda_{z_j}} = \mathbb{E}(\tilde{T}_j) \text{ for } j \in \{1, 2, \dots\}.$$

We then have some further properties of the chain, which are written as follows

$$\begin{aligned} \tilde{T}_0 &= \inf \{t > 0 | X(t) \neq X(0)\} \\ P(\tilde{T}_0 > t | X(0) = i) &= e^{-\lambda_i t} \\ P(X(\tilde{t}_0) = j | X(0) = i) &= J_{ij}. \end{aligned}$$

If $\tilde{T}_0 > T$, we let $\tilde{n} = 0$ and define $T_0 = T$. Otherwise, letting $\tilde{n} = j$ with $j \geq 1$, if $\sum_{i=0}^{j-1} \tilde{T}_i < T$ and $\sum_{i=1}^j \tilde{T}_i > T$, we define $T_0, T_1, \dots, T_{\tilde{n}}$ such that

$$T_j = \begin{cases} \tilde{T}_j & \text{for } j \in \{0, 1, \dots, \tilde{n} - 1\} \text{ and} \\ T - \sum_{i=0}^{\tilde{n}-1} \tilde{T}_i & \text{for } j = \tilde{n} \end{cases}$$

It is useful to reparameterise the chain in terms of $\mathbf{d}, \boldsymbol{\lambda}$, and J . That is $Q = Q(\boldsymbol{\lambda}, J)$ is viewed as a function of $\boldsymbol{\lambda}$ and J . Suppose now that we observe a right-continuous sample path $x(\cdot)$ from the process $\{X(t); 0 \leq t < T\}$, where $T < \infty$ is

the length of our observation window. It follows from (22) and (23) that with probability 1 this is a step function with a finite number of jumps $\tilde{n} < \infty$. See Albert (1962) and Chapter 6 of Doob (1953) for technical details.

We then represent a complete observation $x(\cdot) = \{x(t); 0 \leq t < T\}$ as follows

$$((z_0, t_0), (z_1, t_1), \dots, (z_{\tilde{n}-1}, t_{\tilde{n}-1}), (z_{\tilde{n}}, t_{\tilde{n}})),$$

which is a point in the space

$$\mathcal{W}_{\tilde{n}} = \left[\prod_{j=1}^{\tilde{n}} (\mathcal{W}_0 \times (0, T]) \right] \times \mathcal{W}_0 \quad (24)$$

where $\mathcal{W}_0 = \{1, \dots, w\}$ is the state space of the chain and $(0, T] \subset \mathbb{R}$ is the observation window. We denote σ to be the measure on the space of all sample functions such that

$$\sigma(B) = \sum_{\tilde{n}=0}^{\infty} \sigma^{(\tilde{n})}(B \cap \mathcal{W}_{\tilde{n}})$$

where $B \subset \mathcal{W}$ is an event with $\mathcal{W} = \cup_{\tilde{n}=0}^{\infty} \mathcal{W}_{\tilde{n}}$, and $\sigma^{(\tilde{n})}$ is the measure

$$\sigma^{(\tilde{n})} = \left[\prod_{j=1}^{\tilde{n}} (C \times l) \right] \times C$$

for the Lebesgue measure l on \mathbb{R} and the counting measure C on \mathcal{W}_0 such that $C(\{z\}) = 1$ if $z \in \mathcal{W}_0$ and 0 otherwise.

Note that our initial state is $x(0) = z_0$, so that we have \tilde{n} jumps in total before we reach the end of the observation window and that for $\tilde{n} > 0$, we have $T_{\tilde{n}} = T - \sum_{j=0}^{\tilde{n}-1} T_j$.

With the parameters for the initial state distribution $\mathbf{d} = (d_1, \dots, d_w)$, the rates of leaving each of the states $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_w)$ and the matrix for the jump

process J , as well as the statistics for our initial state $x(0)$, the number of each type of jump

$$n_{jk} = \sum_{i=1}^{\tilde{n}} \mathbb{I}\{z_{i-1} = j, z_i = k\} \text{ for } j, k = 1, 2, \dots, w$$

and the total time spent in each of the states

$$\tau_j = \sum_{i=0}^{\tilde{n}} t_i \mathbb{I}\{z_i = j\} \text{ for } j = 1, 2, \dots, w,$$

we can then write the density with respect to σ to be

$$\begin{aligned} P[B] &= \int_B f(x(\cdot)|\mathbf{d}, \boldsymbol{\lambda}, J) d\sigma(x(\cdot)) \text{ where} \\ f(x(\cdot)|\mathbf{d}, \boldsymbol{\lambda}, J) &= d_{x(0)} g(x(\cdot)|\boldsymbol{\lambda}, J). \end{aligned} \quad (25)$$

Here, $g(x(\cdot)|\boldsymbol{\lambda}, J)$ is the density of the process, conditional on the initial state, which takes the form

$$g(x(\cdot)|\boldsymbol{\lambda}, J) = \begin{cases} e^{-\lambda_{x(0)} T} & \text{if } x(\cdot) = (z_0, T); \\ \prod_{j=1}^w e^{-\lambda_j \tau_j} \prod_{k=1}^w \lambda_j^{n_{jk}} J_{jk}^{n_{jk}} & \text{if } x(\cdot) = ((z_0, t_0), (z_1, t_1), \dots, (z_{n-1}, t_{n-1}), (z_{\tilde{n}}, t_{\tilde{n}})), \\ & \text{with } \tilde{n} > 0, t_i > 0 \text{ (} i = 0, 1, \dots, \tilde{n} - 1 \text{) and } \sum_{i=0}^{\tilde{n}} t_i = T; \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Note that we can write $g(x(\cdot)|\boldsymbol{\lambda}, J) = h^{(1)}(x(\cdot)|\boldsymbol{\lambda})h^{(2)}(x(\cdot)|J)$, where

$$h^{(1)}(x(\cdot)|\boldsymbol{\lambda}) = \prod_{j=1}^w e^{-\lambda_j \tau_j} \prod_{k=1}^w \lambda_j^{n_{jk}} \text{ and} \quad (26)$$

$$h^{(2)}(x(\cdot)|J) = \prod_{j=1}^w \prod_{k=1}^w J_{jk}^{n_{jk}}. \quad (27)$$

Suppose we have n independent and identically distributed (*iid*) realisations of $\{X(t), 0 \leq t \leq T\}$, then we have a log likelihood function defined by

$$\begin{aligned} L_n^{(1)} &= \sum_{k=1}^n \log f(x_k(\cdot)|\mathbf{d}, \boldsymbol{\lambda}, J) \\ &= \sum_{k=1}^n \log [d_{x_k(0)} g(x|\boldsymbol{\lambda}, J)]. \end{aligned}$$

If we let $b_{k,i}$ be 1 where the initial state of x_k is i and 0 otherwise, $n_{k,ij}$ be the number of times that x_k makes an $i \rightarrow j$ transition with $i \neq j$, and $\tau_{k,i}$ be the total time that x_k spends in state i , then the log likelihood becomes

$$\begin{aligned} L_n^{(1)} &= \sum_{k=1}^n \left\{ \sum_{i=1}^w b_{k,i} \log d_i + \sum_{i=1}^w \sum_{j \neq i}^w n_{k,ij} [\log J_{ij} + \log \lambda_i] - \sum_{i=1}^w \tau_{k,i} \lambda_i \right\} \\ &= \sum_{k=1}^n \sum_{i=1}^w b_{k,i} \log d_i + \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i}^w n_{k,ij} \log J_{ij} \\ &\quad + \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i}^w n_{k,ij} \log \lambda_i - \sum_{k=1}^n \sum_{i=1}^w \tau_{k,i} \lambda_i. \end{aligned} \quad (28)$$

As in Albert (1962) we obtain the sufficient statistics

$$\begin{aligned} b_i &= \sum_{k=1}^n b_{k,i} \\ n_{ij} &= \sum_{k=1}^n n_{k,ij} \\ \tau_i &= \sum_{k=1}^n \tau_{k,i} \end{aligned}$$

and the maximum likelihood estimates of the parameters from (28) as

$$\begin{aligned} \hat{d}_i &= \frac{b_i}{\sum_{i=1}^w b_i} \\ \hat{J}_{ij} &= \frac{n_{ij}}{\sum_{j \neq i} n_{ij}} \\ \hat{\lambda}_i &= \frac{\sum_{j \neq i} n_{ij}}{\tau_i}. \end{aligned} \tag{29}$$

Note that when the log likelihood is written in the form (28), we can clearly see that the initial state distribution \mathbf{d} appears in a term that is separate to a term involving the transition probabilities $\{J_{ij}\}$, which is in turn separate to a term involving the rate parameters $\{\lambda_i\}$. When we obtain estimates for each of the parameters by maximising the total likelihood, it suffices to maximise each of these terms separately.

It may be the case that one of our discrete states is an *absorbing state*. This is where the probability of leaving the state is zero. If, for example, state w is an absorbing state, we will assume that there are no observations whose initial state $x(0) = w$, thus we can fix $d_w = 0$ and we will also set the parameters of the w^{th} row of Q to be

$$Q_{wj} = 0 \text{ for } j = 1, \dots, w.$$

Equivalently, we will set $\lambda_w = 0$ and the w^{th} row of J to be

$$J_{wj} = 0 \text{ for } j = 1, \dots, w.$$

This constrains the parameter space. Then, assuming there are no jump observations out of the absorbing state, the maximum likelihood estimates given by (29) above will still obey the constraints.

4.2 General finite mixtures of continuous-time Markov chains

We now consider the case where our stochastic process $\{X(t), 0 \leq t \leq T\}$ is driven by one of a finite number N of possible independent discrete state, continuous-time Markov chain components $\{X_m(t), 0 \leq t \leq T\}$ for $m \in 1, 2, \dots, N$. Thus, we have a random vector $\mathbf{Y} = (Y_1, \dots, Y_N)$, where $Y_m = 1$ if $X(t)$ is driven by the m^{th} component and 0 otherwise. Let

$$P(Y_m = 1) = \pi_m \text{ for } m \in 1, \dots, N,$$

where each $\pi_m \geq 0$ and $\sum_{m=1}^N \pi_m = 1$. We then have

$$X(t) = \sum_{m=1}^N X_m(t) \mathbb{I}\{Y_m = 1\},$$

so that $X(t)$ has a mixture of Markov chains distribution.

We use a form of Markov chain mixtures presented in Frydman (2005), where the mixing is on the transition rates $\boldsymbol{\lambda}$ and the initial state distributions. That is the m^{th} component $X_m(t)$ has parameters $\boldsymbol{\lambda}_m, \mathbf{d}_m$ and J (note that J is common to all components). This parameterisation is different to Frydman (2005) but the model is equivalent as we show with (34) and (36) below.

We have $\{X(t), 0 \leq t \leq T\}$ that takes values in the set $\{1, 2, \dots, w\}$, so $\{X(t)\}$

takes values in the same sample space \mathcal{W} as described in (24) above such that

$$\begin{aligned}
P(X(\cdot) \in B) &= \sum_{m=1}^N P(X(\cdot) \in B | Y_m = 1) P(Y_m = 1) \\
&= \sum_{m=1}^N P(X_m(\cdot) \in B) P(Y_m = 1) \\
&= \sum_{m=1}^N \pi_m \int_B f_m(x(\cdot)) d\sigma(x(\cdot)). \tag{30}
\end{aligned}$$

This implies that the density is as follows

$$\begin{aligned}
f(x(\cdot)) &= \sum_{m=1}^N \pi_m f_m(x(\cdot)) \\
&= \sum_{m=1}^N \pi_m d_{x(0),m} g(x(\cdot) | \boldsymbol{\lambda}_m, J). \tag{31}
\end{aligned}$$

Here, $f(\cdot)$ is the density for a mixture of $N < \infty$ continuous-time Markov chains $\{X_m(t), 0 \leq t \leq T\}$ with initial distributions $\{d_{1,m}, d_{2,m}, \dots, d_{w,m}\}$ and generators $Q_m = Q(\boldsymbol{\lambda}_m, J)$ for $1 \leq m \leq N$.

We write (as in (25)) the density of a single Markov chain in the form

$$\begin{aligned}
f(x(\cdot) | \mathbf{d}, \boldsymbol{\lambda}, J) &= d_{x(0)} g(x(\cdot) | \boldsymbol{\lambda}, J) \\
&= d_{x(0)} h^{(1)}(x(\cdot) | \boldsymbol{\lambda}) h^{(2)}(x(\cdot) | J),
\end{aligned}$$

where $h^{(1)}(x(\cdot) | \boldsymbol{\lambda})$ and $h^{(2)}(x(\cdot) | J)$ are defined with (26) and (27) respectively. We see that the mixture density can be written as

$$\begin{aligned}
f(x(\cdot) | \mathbf{d}, \boldsymbol{\pi}, \boldsymbol{\lambda}, J) &= \sum_{m=1}^N \pi_m d_{x(0),m} g(x(\cdot) | \boldsymbol{\lambda}_m, J) \\
&= h^{(2)}(x(\cdot) | J) \sum_{m=1}^N \pi_m d_{x(0),m} h^{(1)}(x(\cdot) | \boldsymbol{\lambda}_m). \tag{32}
\end{aligned}$$

For a finite number n of *iid* observations, we write the N -component Markov chain mixture log-likelihood as

$$\begin{aligned} L_n^{(N)} &= \log \left\{ \prod_{k=1}^n \left[h^{(2)}(x_k(\cdot)|J) \sum_{m=1}^N \pi_m d_{x_k(0),m} h^{(1)}(x_k(\cdot)|\boldsymbol{\lambda}_m) \right] \right\} \\ &= \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} + \sum_{k=1}^n \log \left\{ \sum_{m=1}^N \pi_m d_{x_k(0),m} h^{(1)}(x_k(\cdot)|\boldsymbol{\lambda}_m) \right\} \end{aligned} \quad (33)$$

Note the log-likelihood can be written as the sum of two terms, one of which only involves J , the other only involving the other parameters. Indeed, the term involving J is identical to that in the single component Markov chain model log-likelihood in (28) and so maximum likelihood estimation for J is the same as in that case, since it again suffices to maximise each term separately.

It is convenient to change parameters slightly, by expressing $\boldsymbol{\pi}, \mathbf{d}_1, \dots, \mathbf{d}_N$ in terms of other parameters $\mathbf{d}, \mathbf{s}_1, \dots, \mathbf{s}_N$, defined as follows. From (30), we again write $\mathbf{d} = (d_1, \dots, d_w)$ as the initial distribution of $X(\cdot)$ under the mixture model, which is given by

$$\begin{aligned} d_i &= P(X(0) = i) \\ &= \sum_{m=1}^N \pi_m P(X_m(0) = i) \\ &= \sum_{m=1}^N \pi_m d_{i,m}. \end{aligned}$$

Defining

$$\begin{aligned} s_{i,m} &= \frac{\pi_m d_{i,m}}{\sum_{m=1}^N \pi_m d_{i,m}} \\ &= P(Y_m = 1 | X(0) = i) \end{aligned} \quad (34)$$

as the conditional mixture proportions given the initial state, we thus write the density in (31) as

$$\begin{aligned} f(x(\cdot)|\mathbf{d}, \mathbf{s}, \boldsymbol{\lambda}, J) &= \sum_{m=1}^N d_{x(0)} s_{x(0),m} h^{(1)}(x(\cdot)|\boldsymbol{\lambda}_m) h^{(2)}(x(\cdot)|J) \\ &= d_{x(0)} h^{(2)}(x(\cdot)|J) \left\{ \sum_{m=1}^N s_{x(0),m} h^{(1)}(x(\cdot)|\boldsymbol{\lambda}_m) \right\}. \end{aligned} \quad (35)$$

The N -component Markov chain mixture log-likelihood of n *iid* observations can then be written as

$$\begin{aligned} L_n^{(N)} &= \log \left\{ \prod_{k=1}^n \left[d_{x_k(0)} h^{(2)}(x_k(\cdot)|J) \left\{ \sum_{m=1}^N s_{x_k(0),m} h^{(1)}(x_k(\cdot)|\boldsymbol{\lambda}_m) \right\} \right] \right\} \\ &= \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} + \sum_{k=1}^n \sum_{i=1}^w b_{k,i} \log d_i \\ &\quad + \sum_{k=1}^n \log \left\{ \sum_{m=1}^N s_{x_k(0),m} h^{(1)}(x_k(\cdot)|\boldsymbol{\lambda}_m) \right\}. \end{aligned} \quad (36)$$

Note that it is clear from the form of (36) that the initial state distribution can now be estimated separately from the other parameters, similarly to how we separately estimate the jump matrix J in (28). We see that the initial state distribution vector \mathbf{d} and the jump matrix J have the same interpretation and maximum likelihood estimates as under a 1-component Markov chain.

There is, however, no easy closed form solution for the maximisation over $\boldsymbol{\lambda}$ and \mathbf{s} so these must be obtained numerically. Frydman (2005) employs the E-M algorithm of Dempster et al. (1977) to obtain maximum likelihood estimates. This is an iterative algorithm with two steps. In the “ E ”-step, we obtain the *E-M log-likelihood* as a function of our parameters and the data, given by the conditional expectation under the current parameter values of the full data log-likelihood given the observed data. Then the “ M ”-step involves finding parameter values which

maximise the *E-M log-likelihood*. These steps are iterated until there is sufficient convergence of the parameter estimates (or some other stopping criterion).

Frydman (2005) considers a conditional version of this model, which does not model the initial states as random. The n observations can be split into w sets based on their initial states, each with b_1, \dots, b_w observations respectively, with densities given as follows

$$\begin{aligned} X_{1,1}(\cdot), \dots, X_{1,b_1}(\cdot) & \text{ are } iid \text{ with density } f(x(\cdot)) = \sum_{m=1}^N s_{1,m} g(x(\cdot) | \boldsymbol{\lambda}_m, J, x(0) = 1) \\ X_{2,1}(\cdot), \dots, X_{2,b_2}(\cdot) & \text{ are } iid \text{ with density } f(x(\cdot)) = \sum_{m=1}^N s_{2,m} g(x(\cdot) | \boldsymbol{\lambda}_m, J, x(0) = 2) \\ & \vdots \\ X_{w,1}(\cdot), \dots, X_{w,b_w}(\cdot) & \text{ are } iid \text{ with density } f(x(\cdot)) = \sum_{m=1}^N s_{w,m} g(x(\cdot) | \boldsymbol{\lambda}_m, J, x(0) = w). \end{aligned}$$

The conditional log-likelihood of all such observations is then given by

$$L_n^{(Nc)} = \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} + \sum_{k=1}^n \log \left\{ \sum_{m=1}^N s_{x(0),m} h^{(1)}(x_k(\cdot) | \boldsymbol{\lambda}_m) \right\}, \quad (37)$$

which is just $L_n^{(N)}$ from (36) with the term involving the d_i 's and b_i 's removed.

Thus, all of the remaining parameters have the same (conditional) maximum likelihood estimates under (37) as in the unconditional case (36). The log-likelihood ratio for comparison with the single component Markov chain that conditions on the initial states,

$$\begin{aligned} L_n^{(1c)} &= \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} \\ &+ \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log \lambda_i - \sum_{k=1}^n \sum_{i=1}^w \tau_{k,i} \lambda_i, \end{aligned} \quad (38)$$

which is the same as $L_n^{(1)}$ from (28) with the term involving the d_i 's and b_i 's

removed. Thus, the log-likelihood ratio statistics for the two formulations $L_n^{(N_c)} - L_n^{(1_c)}$ and $L_n^{(N)} - L_n^{(1)}$ are identical. We choose to use the unconditional (*iid*) formulation as it is more convenient to derive its asymptotic properties.

4.3 Testing between 1 and 2 mixture components

The data set studied in Frydman (2005) consists of the time-series of credit ratings for a sample of 848 corporate bond issuers in the industrial sector, observed each day between January 1985 and December 1995. The original rating categories are grouped into the coarser rating states *Aaa*, *Aa*, *A*, *Baa*, *Ba*, *B*, and *C*. There are also rating states *WR* for *rating withdrawal* and *D* for the *default* state. Note that the initial states of each of the firms are given. They are distributed across each of the ratings states, except for *WR* and *D*, where there are no firms that begin in these states. There are also no firms that migrate out of *D*, thus it is assumed to be an absorbing state.

In practice, credit rating dynamics are often modelled with a simple Markov chain. However, this can fail to pick up some of the more complex dynamics that appear in the data. The new mixture introduced in Frydman (2005) allows for the modelling of population heterogeneity on the rates that firms leave each rating state and argues that this provides a significantly better fit than the simple Markov chain.

In Frydman (2005) there is a likelihood ratio test conducted to test between the null hypothesis of a simple Markov chain model and the alternative hypothesis, where the data is modelled by a mixture of two Markov chain components (which has a jump process J that is common to all of the components in a mixture). In order to test for whether it is necessary to introduce these additional parameters to model the data, as opposed to using a single Markov chain component, Frydman (2005) conducts a likelihood ratio test. As discussed at the end of Section 4.2,

we are considering the (*iid*) unconditional model, which is superficially different to Frydman (2005) but essentially equivalent. The likelihood ratio test between a mixture of 2 Markov chains and a simple Markov chain uses the following statistic

$$\Lambda_n = L_n^{(2)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2, \hat{\mathbf{s}}, \hat{J}) - L_n^{(1)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_0, \hat{J}), \quad (39)$$

where we write the log-likelihood functions for the single component Markov chain and the 2-component Markov chain mixture as

$$\begin{aligned} L_n^{(1)}(\mathbf{d}, \boldsymbol{\lambda}, J) &= \sum_{k=1}^n \sum_{i=1}^w b_{k,i} \log d_i + \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} \\ &+ \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log \lambda_i - \sum_{k=1}^n \sum_{i=1}^w \tau_{k,i} \lambda_i \end{aligned} \quad (40)$$

$$\begin{aligned} L_n^{(2)}(\mathbf{d}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \mathbf{s}, J) &= \sum_{k=1}^n \sum_{i=1}^w b_{k,i} \log d_i + \sum_{k=1}^n \sum_{i=1}^w \sum_{j \neq i} n_{k,ij} \log J_{ij} \\ &+ \sum_{k=1}^n \log \left\{ s_{x_k(0)} \prod_{i=1}^w e^{-\lambda_{1,i} \tau_i} \prod_{j=1}^w \lambda_{1,i}^{n_{ij}} \right. \\ &\left. + (\mathbf{1} - s_{x_k(0)}) \prod_{i=1}^w e^{-\lambda_{2,i} \tau_i} \prod_{j=1}^w \lambda_{2,i}^{n_{ij}} \right\}, \end{aligned} \quad (41)$$

and define maximum likelihood estimates via

$$(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2, \hat{\mathbf{s}}, \hat{J}) = \operatorname{argmax}_{\mathbf{d}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \mathbf{s}, J} L_n^{(2)}(\mathbf{d}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \mathbf{s}, J),$$

and $(\hat{\mathbf{d}}_0, \hat{\boldsymbol{\lambda}}_0, \hat{J}_0) = \operatorname{argmax}_{\mathbf{d}, \boldsymbol{\lambda}, J} L_n^{(1)}(\mathbf{d}, \boldsymbol{\lambda}, J)$, noting that $\hat{\mathbf{d}}_0 = \hat{\mathbf{d}}$ and $\hat{J}_0 = \hat{J}$ as discussed in Section 4.2. Thus, the log-likelihood ratio for the test in Frydman (2005) can be seen to be of the form from (39), although we are also modelling the initial states as random so we have *iid* observations, facilitating our analysis

below.

Frydman (2005) claims that by standard theory, the likelihood ratio test statistic Λ_n is distributed under the null hypothesis as a chi-squared with 8 degrees of freedom, corresponding to the difference in the number of parameters. The resulting statistic of 276.96 is used to indicate a strong rejection of the simple Markov chain in favour of the alternative model. However, it is not obvious that the standard theory applies. We show below that in fact Λ_n diverges to infinity under the *iid* model and indeed the results of the parametric bootstrap in the next section suggest the result of the likelihood ratio test is not even significant.

4.3.1 A parametric bootstrap procedure to test for the presence of a mixture

To investigate this further, we employ a parametric bootstrap procedure. This procedure is used to understand the distributional properties of a statistic using resampling. In our case, we are interested in approximating the distribution of the likelihood ratio test statistic under the null hypothesis. We simulate from our fit and recompute the statistic a large number of times. We can then compare our calculated statistic to this approximate empirical distribution of simulated statistics to assess the evidence against the null hypothesis in favour of the alternative. The parametric bootstrap *p-value* is then calculated as the proportion of simulated statistics exceeding the originally calculated statistic. Although we do not prove here that this test is conclusive, the resultant parametric bootstrap *p-value* we derive is *strongly suggestive* that the result is *not significant* evidence against the null hypothesis in favour of the alternative.

The procedure is conducted as follows:

1. Estimate the simple Markov chain parameters by maximum likelihood estimation from the original sample data

2. Estimate the two component Markov mixture model parameters with the EM algorithm from the original sample data
3. Calculate the likelihood ratio test statistic for the above two models and the original sample data
4. Simulate another dataset from the simple Markov chain estimated in step 1
5. Estimate the simple Markov chain and Markov mixture models from this data and calculate the likelihood ratio test statistic
6. Repeat step 4 and step 5 many times, recording the likelihood ratio test statistics at each stage to form a simulated empirical distribution
7. Compare the likelihood ratio test statistic from step 3 to the distribution of likelihood ratio test statistics in step 6 to obtain a parametric bootstrap p-value. If the parametric bootstrap p-value is not very small then we conclude that the data provides no evidence against the null hypothesis of a single-component Markov chain.

We use the data summaries from Tables 1 and 2 from Frydman (2005) and define the empirical data parameters including the observation window, number of firms in each of the starting states as well as the simple Markov chain to generate the data. Then we extract the parameters from the maximum likelihood estimates in Frydman (2005) and simulate from their single Markov chain fit. Finally, we conduct the parametric bootstrap procedure that yields the following result in Figure 7.

The parametric bootstrap procedure is an important and widely used data-analytic tool. We note that further investigation is warranted into the properties

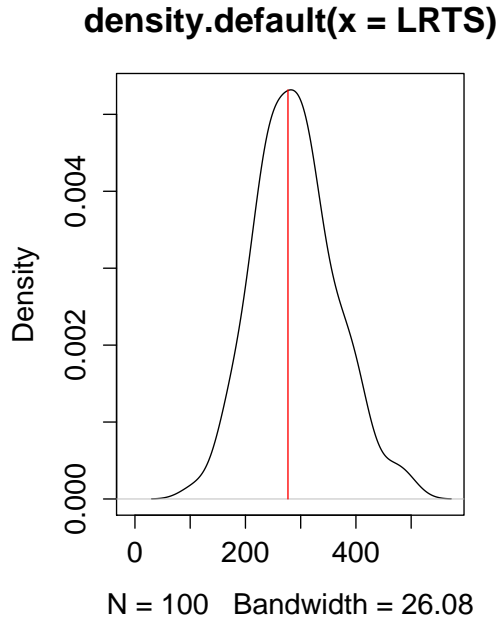


Figure 7: A comparison between the likelihood ratio statistic derived in Frydman (2005) and the distribution of the likelihood ratio statistic derived from a parametric bootstrap procedure. The parametric bootstrap p-value is 0.528, which suggests that we do not have sufficient evidence to reject the null hypothesis.

of the estimates and asymptotic distributions involved in the test. It suffices to say that this result challenges the claim that the data provides strong evidence against the null hypothesis in favour of the alternative, with a likelihood ratio test statistic of 276.96. The parametric bootstrap p-value of 0.528 suggests this is not strong evidence against the null hypothesis of a single Markov chain. This provides significant motivation for us to further explore the properties of the likelihood ratio test between a single component Markov chain model and a two component mixture of Markov chains alternative.

It is pointed out in Frydman (2005) and is widely the case in practice that not all of the censoring times are the same. That is, we have firms that enter our dataset at different times, so are thus observed over varying periods. In Frydman

(2005) no indication is given as to the distribution of starting times for each of the firms, so we have assumed that the initial states of each of the firms are all observed at the beginning of the observation window.

4.3.2 Non-identifiability of the likelihood ratio test

Despite the fact that mixture models have been studied for some time, for many types of models, there remains no definitive way to test for the number of mixture components. Studies such as Hartigan (1985), Ghosh and Sen (1985), Dacunha-Castelle and Gassiat (1997), Liu and Shao (2003) and Garel (2005) develop methods to deal with likelihood ratio tests for the number of components in mixture models. However, they are either applied to mixtures of normal distributions or require certain conditions to hold that can be very difficult to verify in a practical setting. Frydman (2005) claims that by standard theory, the asymptotic distribution of the usual log-likelihood ratio test statistic Λ_n , under H_0 , is chi-squared with $(N - 1) \times w$ degrees of freedom. However, the appropriate regularity conditions for applying the test are not verified. Given the results of our parametric bootstrap procedure, rather than Λ_n having an asymptotic chi-squared distribution, its limiting behaviour is given by the following proposition:

Proposition 4.1. *If Λ_n is given by (39), then $\Lambda_n \rightarrow \infty$ in probability as $n \rightarrow \infty$.*

The problem here is that the null hypothesis H_0 is not identifiable. That is, there are infinitely many ways that the null model (of the simple Markov chain) can be written in terms of the parameters of the alternative model (of a mixture of two continuous-time Markov chains).

Let us define the null model parameters to be $\theta_0 = \{\mathbf{d}_0, \boldsymbol{\lambda}_0, J_0\}$ where $\theta_0 \in \Theta_0$ is a point within the null parameter space, and the alternative model parameters are $\theta = \{\mathbf{d}, \mathbf{s}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, J\}$ where $\theta \in \Theta$ is a point within the higher-dimensional alternative parameter space. If we let $s_i = 0$ for $i \in 1, \dots, w$ then our null model

and alternative model have the same likelihood. Thus, one may propose that we simply test for whether each $s_i = 0$. However, this would not be appropriate since there is the case where $s_i \neq 0$ for $i \in 1, \dots, w$ and we set $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \boldsymbol{\lambda}_0$, which would also mean that the null model and the alternative model have the same likelihood. The point $\{\boldsymbol{d}_0, \boldsymbol{\lambda}_0, J_0\} \in \Theta_0$ does not correspond to a single point in Θ . In fact, each single point in the null model space corresponds to infinitely many points in the alternative model space. It is thus possible that the parameters change but the likelihood doesn't. This is what we refer to as *non-identifiability*.

This non-identifiability issue implies that the Fisher information matrix of the likelihood ratio, under the null hypothesis, is singular, so the “standard theory” of Wilks (1938) does not necessarily apply. The model is a very interesting one, given the need in finance to model the inhomogenous behaviour of large populations over time. Thus from a practitioner's perspective, it is important to understand the considerations that must be taken into account in testing for the number of mixture components required.

The problem of testing for the identification of a mixture using the likelihood ratio test was explored with normal mixtures in Hartigan (1985). For an *iid* sample X_1, X_2, \dots, X_n , Hartigan (1985) examines the asymptotics of the likelihood ratio for the test between

$$\begin{aligned}
 H_0 : X_1 &\sim N(0, 1) \text{ against} \\
 H_1 : X_1 &\sim (1 - p)N(0, 1) + pN(\theta, 1).
 \end{aligned}$$

The asymptotic properties of the likelihood ratio

$$\begin{aligned} L_n &= \sup_{\theta, p} L_n(\theta, p) \\ &= \sup_{\theta, p} \sum_{i=1}^n \log \left[(1-p) + pe^{X_i\theta - \frac{1}{2}\theta^2} \right] \end{aligned}$$

are then derived. For this example, Hartigan (1985) proves that the log-likelihood ratio diverges to infinity in probability and conjectures that the rate of divergence is of the order $\log \log n$. This was later proven in Bickel and Chernoff (1993) and Liu and Shao (2004).

A generalisation of this example is presented in Fukumizu (2003) for locally conic models, using the reparameterisation techniques of Dacunha-Castelle and Gassiat (1997). Fukumizu (2003) proves that under some regularity conditions, if there exists a sequence of standardised score functions that approaches 0 in probability, then the likelihood ratio diverges in probability. The theorem is applied to a practical example of the likelihood ratio for multilayer neural network models. We provide a useful sufficient condition in a special case of this theory for our practical example with the test between 1 and 2 component mixtures, with a focus on the mixtures of Markov chains presented in Frydman (2005).

Recall that the likelihood ratio test statistic Λ_n , for the test between 1 and 2 mixture components presented in Frydman (2005), can be written as in (39). We observe that

$$\begin{aligned} \Lambda_n &= L_n^{(2)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_1, \hat{\boldsymbol{\lambda}}_2, \hat{\mathbf{s}}, \hat{J}) - L_n^{(1)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_0, \hat{J}) \\ &\geq L_n^{(2)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, \tilde{\boldsymbol{\lambda}}_2, \tilde{\boldsymbol{\pi}}\mathbf{1}, J_0) - L_n^{(1)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_0, \hat{J}) \end{aligned} \quad (42)$$

$$= \left[L_n^{(2)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, \tilde{\boldsymbol{\lambda}}_2, \tilde{\boldsymbol{\pi}}\mathbf{1}, J_0) - L_n^{(1)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, J_0) \right] \quad (43)$$

$$- \left[L_n^{(1)}(\hat{\mathbf{d}}, \hat{\boldsymbol{\lambda}}_0, \hat{J}) - L_n^{(1)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, J_0) \right], \quad (44)$$

where

$$(\tilde{\boldsymbol{\lambda}}_2, \tilde{\pi}) = \operatorname{argmax}_{\boldsymbol{\lambda}_2, \pi} L_n^{(2)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_2, \pi \mathbf{1}, J_0).$$

For the parameters of $L_n^{(2)}$ in (42), we set each element of \mathbf{d} to be \mathbf{d}_0 the true value of the initial distribution under the null hypothesis, we set J equal to J_0 the true value of the jump matrix under the null hypothesis and we set (s_1, \dots, s_w) to be some fixed (π, \dots, π) with $0 < \pi < 1$. Note that (44) is written as the difference between two log-likelihood ratio test statistics for simple hypothesis tests, where $\boldsymbol{\lambda}_0$ is the true value of $\boldsymbol{\lambda}$ under the null hypothesis. The first of these is a test between

$$\begin{aligned} H_0 : X(\cdot) &\sim d_{0,x(0)}g(x(\cdot)|\boldsymbol{\lambda}_0, J_0), \quad \mathbf{d}_0, \boldsymbol{\lambda}_0, J_0 \text{ known, against} \\ H_1^{(2)} : X(\cdot) &\sim d_{0,x(0)} \{(1 - \pi)g(x(\cdot)|\boldsymbol{\lambda}_0, J_0) + \pi g(x(\cdot)|\boldsymbol{\lambda}_2, J_0)\}. \end{aligned} \quad (45)$$

The second test has an identical null hypothesis but has a different alternative hypothesis

$$H_1^{(1)} : X(\cdot) \sim d_{x(0)}g(x(\cdot)|\boldsymbol{\lambda}_0, J), \quad \mathbf{d}, \boldsymbol{\lambda}_0, J \text{ unknown.} \quad (46)$$

Twice the log-likelihood ratio test statistic in (44) is for a regular testing problem with alternative hypothesis (46) and so, under the assumption that the true values of J_0 , $\boldsymbol{\lambda}_0$ and \mathbf{d}_0 are interior points in the parameter space, is asymptotically chi-squared distributed with degrees of freedom equal to the difference in the number of free parameters between H_0 and $H_1^{(1)}$. Otherwise a mixture of chi-squared distributions is obtained (see Chernoff (1954) for details). Then defining,

$$\Lambda_1 = \left[L_n^{(2)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, \tilde{\boldsymbol{\lambda}}_2, \tilde{\pi} \mathbf{1}, J_0) - L_n^{(1)}(\mathbf{d}_0, \boldsymbol{\lambda}_0, J_0) \right], \quad (47)$$

we can write

$$\Lambda_n \geq \Lambda_1 + \mathcal{O}_p(1). \quad (48)$$

We show below that $\Lambda_1 \rightarrow \infty$. If Λ_1 diverges, then so too does Λ_n , at least as quickly.

4.3.3 Divergence of the log-likelihood ratio test statistic

Suppose $\{g_\lambda\}$ is a set of density functions with a parameter λ that takes values in the parameter space Λ . Let λ_0 be a fixed, known parameter value within Λ . Suppose further that each of the density functions in $\{g_\lambda\}$ are dominated by g_{λ_0} for all $\lambda \in \Lambda$, so that $g_\lambda(x) > 0 \implies g_{\lambda_0}(x) > 0$. Then, we let $r_\lambda = \frac{g_\lambda}{g_{\lambda_0}}$ and

$$\|r_\lambda\|^2 = \int r_\lambda^2 g_{\lambda_0} d\mu = \int \left(\frac{g_\lambda^2}{g_{\lambda_0}} \right) d\mu. \quad (49)$$

We consider the statistical model $S = \{f(x|(\boldsymbol{\lambda}, \beta))\}$, which is a mixture model with two components

$$f(x|(\boldsymbol{\lambda}, \beta)) = (1 - p(\boldsymbol{\lambda}, \beta))g(x|\boldsymbol{\lambda}_0) + p(\boldsymbol{\lambda}, \beta)g(x|\boldsymbol{\lambda}), \quad (50)$$

where

$$p(\boldsymbol{\lambda}, \beta) = \frac{\beta}{\sqrt{\|r_\lambda\|^2 - 1}}.$$

We can then refer to the parameter space $\Theta = \{\beta \in [0, \mathcal{B}_\lambda], \boldsymbol{\lambda} \in \Lambda\}$ where $\mathcal{B}_\lambda = \sqrt{\|r_\lambda\|^2 - 1}$.

Suppose we have an *iid* sample X_1, X_2, \dots, X_n generated by the true density

$g(x|\boldsymbol{\lambda}_0)$. For the hypothesis test

$$\begin{aligned} H_0 : X_1 &\sim g(X|\boldsymbol{\lambda}_0) \text{ with } \boldsymbol{\lambda}_0 \in \Lambda \text{ known, against} \\ H_1 : X_1 &\sim f(X|(\boldsymbol{\lambda}, \beta)) \text{ with } (\boldsymbol{\lambda}, \beta) \in \Theta = \Lambda \times \mathcal{B}_\lambda, \end{aligned}$$

defining $\Theta_0 = \{(\boldsymbol{\lambda}, \beta) \in \Theta | f = g\}$, we have the likelihood ratio

$$\sup_{\boldsymbol{\lambda} \in \Lambda, \beta \in \mathcal{B}_\lambda} L_n((\boldsymbol{\lambda}, \beta)) = \sup_{\boldsymbol{\lambda} \in \Lambda, \beta \in \mathcal{B}_\lambda} \sum_{i=1}^n \log \frac{f(X_i|(\boldsymbol{\lambda}, \beta))}{g(x|\boldsymbol{\lambda}_0)}. \quad (51)$$

From (50), we have

$$f(X_1|(\boldsymbol{\lambda}, \beta)) = \beta \frac{\left(\frac{g(\boldsymbol{\lambda})}{g_0} - 1\right)}{\left\|\frac{g(\boldsymbol{\lambda})}{g_0} - 1\right\|} g_0 + g_0, \quad (52)$$

where $g(\boldsymbol{\lambda})$ and g_0 represent $g(x|\boldsymbol{\lambda})$ and $g(x|\boldsymbol{\lambda}_0)$ respectively.

The conditions for S to be *locally conic* at $f_0 = g_0$ in the sense of Fukumizu (2003) are as follows

1. The parameter space Θ contains the set of true parameters $\Theta_0 = \Lambda \times \{0\}$, where $f(x|\boldsymbol{\lambda}, \beta) = g(x|\boldsymbol{\lambda}_0)$ [μ a.e.] $\iff \beta = 0$.
2. For each $\boldsymbol{\lambda} \in \Lambda$, the set $\Theta(\boldsymbol{\lambda}) = \{\beta \in \mathcal{B} | (\boldsymbol{\lambda}, \beta) \in \Theta\}$ is a closed interval with open interior.
3. $f(x; \boldsymbol{\lambda}, \beta)$ is differentiable on β (right differentiable at 0) for each $\boldsymbol{\lambda} \in \Lambda$ and $f_0\mu$ -a.e. x . For each $\boldsymbol{\lambda} \in \Lambda$ the Fisher information

$$\left\| \frac{\partial \log f(x|\boldsymbol{\lambda}, 0)}{\partial \beta} \right\| = 1.$$

We write the score function of $S_\lambda = \{f(x|\boldsymbol{\lambda}, \beta) | \beta \in \Theta(\boldsymbol{\lambda})\}$ at the origin $\beta = 0$

as

$$\begin{aligned}\nu_{\boldsymbol{\lambda}}(x) &= \frac{\partial \log f(x|\boldsymbol{\lambda}, 0)}{\partial \beta} \\ &= \frac{\left(\frac{g(\boldsymbol{\lambda})}{g_0} - 1\right)}{\left\|\frac{g(\boldsymbol{\lambda})}{g_0} - 1\right\|}.\end{aligned}\tag{53}$$

Writing $g(x|\beta) = f(x|\boldsymbol{\lambda}, \beta)$ with some $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ fixed, we present the conditions for *asymptotic normality* in the sense of Fukumizu (2003) to be as follows

1. For any $\beta \in \Theta(\boldsymbol{\lambda})$, the integral $\mathbb{E}_{g_0} [|g(x|\beta)|]$ is finite.
2. If $\Theta(\boldsymbol{\lambda}) = \mathbb{R}^+$, where $\mathbb{R}^+ = \{x; x \in \mathbb{R}, x \geq 0\}$, the function $H(x; t) = \sup_{\beta \geq t} \log g(x; \beta)$ satisfies $\lim_{t \rightarrow \infty} \mathbb{E}_{f_0 \mu} [H(x; t)] < \infty$ and there exists Δ such that $\int_{\Delta} f_0(x) d\mu > 0$ and $\lim_{t \rightarrow \infty} H(x; t) = -\infty$ for all $x \in \Delta$.
3. $\lim_{\rho \downarrow 0} \mathbb{E}_{f_0 \mu} [\sup_{|\beta' - \beta| \leq \rho} \log g(x|\beta')] < \infty$ for all $\beta \in \Theta(\boldsymbol{\lambda})$.
4. The density $g(x|\beta)$ is three times differentiable on β for all z and

$$\begin{aligned}\lim_{\rho \downarrow 0} \int \sup_{0 \leq \beta \leq \rho} \left| \frac{\partial^\nu g(x|\beta)}{\partial \beta^\nu} \right| d\mu < \infty \text{ for } \nu = 1, 2, \\ \lim_{\rho \downarrow 0} \mathbb{E}_{f_0 \mu} \left[\sup_{0 \leq \beta \leq \rho} \left| \frac{\partial^3 \log g(x|\beta)}{\partial \beta^3} \right| \right] < \infty.\end{aligned}$$

The conditions for S to be locally conic at g_0 and the conditions for asymptotic normality for each submodel $S_{\boldsymbol{\lambda}} = \{f(x|(\boldsymbol{\lambda}, \beta)|\beta)\}$ will be satisfied in the case of the general 2-component mixture model (50) due to (52), if we have the following:

$$\text{The ratio } r_{\boldsymbol{\lambda}} = \frac{g_{\boldsymbol{\lambda}}}{g_0} \text{ is well defined for all } \boldsymbol{\lambda}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda};\tag{54}$$

$$\|r_{\boldsymbol{\lambda}}\| = 1 \iff \boldsymbol{\lambda} = \boldsymbol{\lambda}_0; \text{ and}\tag{55}$$

$$\int |\log g_{\boldsymbol{\lambda}}(x)| g_0(x) dx < \infty \text{ for all } \boldsymbol{\lambda} \in \boldsymbol{\Lambda}.\tag{56}$$

We now present a version of Theorem 1 from Fukumizu (2003), with simplified regularity conditions, for the case of the mixture model (50).

Theorem 4.2. *Let $S = \{f(x|(\boldsymbol{\lambda}, \beta))\}$ be a statistical model given by (50), that satisfies (54), (55) and (56). Let $C = \{\nu_{\boldsymbol{\lambda}} | \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ be the family of score functions as in (53). If there exists a sequence of score functions $\{\nu_i\}_{i=1}^{\infty}$ in C such that $\nu_i \rightarrow 0$ in probability, then, for arbitrary $M > 0$, we have*

$$\lim_{n \rightarrow \infty} P \left(\sup_{(\boldsymbol{\lambda}, \beta)} L_n(\boldsymbol{\lambda}, \beta) \leq M \right) = 0. \quad (57)$$

Proof. Here, we show that a mixture model S as in (50) that satisfies (54), (55) and (56) satisfies the conditions for S to be locally conic at g_0 as well as the conditions for asymptotic normality.

The parameter space for β is $[0, \mathcal{B}_{\boldsymbol{\lambda}}]$ where $\mathcal{B}_{\boldsymbol{\lambda}} = \sqrt{\|r_{\boldsymbol{\lambda}}\|^2 - 1}$. Condition (54) implies that β is always well-defined. We see through the form of (52) that $\beta = 0$ implies $f(x; (\boldsymbol{\lambda}, \beta)) = g(x; \boldsymbol{\lambda}_0)$ [μ a.e]. If we let $f(x; (\boldsymbol{\lambda}, \beta)) = g(x; \boldsymbol{\lambda}_0)$ [μ a.e] then condition (54) implies that $\|r_{\boldsymbol{\lambda}}\| = 1$. Then, the form of $\mathcal{B}_{\boldsymbol{\lambda}}$ implies that $\beta = 0$. This gives us the first locally conic condition.

The second locally conic condition is easily satisfied since for each fixed $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, condition (54) implies that $\mathcal{B}_{\boldsymbol{\lambda}} = \sqrt{\|r_{\boldsymbol{\lambda}}\|^2 - 1}$ is given by a closed interval with open interior.

The third locally conic condition is trivial, due to the form of (52), which is linear in β for each $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and $f_0 \mu$ -a.e. x and calculating the Fisher information yields

$$\left\| \frac{\partial \log f(x; \boldsymbol{\lambda}, 0)}{\partial \beta} \right\| = \frac{\left\| \frac{g(\boldsymbol{\lambda})}{g_0} - 1 \right\|}{\left\| \frac{g(\boldsymbol{\lambda})}{g_0} - 1 \right\|} = 1.$$

Thus, the conditions (54) and (55) imply the conditions for S to be locally conic

at g_0 .

From condition (56) we have

$$\begin{aligned}
& \int |\log g_{\lambda}(x)| g_0(x) dx < \infty \text{ for all } \lambda \in \Lambda \\
\implies & \int (\log g_{\lambda}(x)) g_0(x) dx > -\infty \\
\implies & (1-p) \int (\log g_0(x)) g_0(x) dx + p \int (\log g_{\lambda}(x)) g_0(x) dx > -\infty \\
\implies & \int (\log [(1-p)g_0(x) + pg_{\lambda}(x)]) g_0(x) > -\infty \tag{58}
\end{aligned}$$

due to Jensen's inequality. We note also that condition (56) yields

$$\begin{aligned}
& \int |\log g_{\lambda}(x)| g_0(x) dx < \infty \text{ for all } \lambda \in \Lambda \\
\implies & \int (\log g_{\lambda}(x)) g_0(x) dx < \infty,
\end{aligned}$$

which gives the third condition for asymptotic normality. This property also implies

$$\begin{aligned}
\int (\log [(1-p)g_0(x) + pg_{\lambda}(x)]) g_0(x) dx &= \int (\log g_0) g_0 dx \\
&+ \int \log \left(1 + p \left(\frac{g_{\lambda}}{g_0} - 1 \right) \right) g_0 dx \\
&\leq \int (\log g_0) g_0 dx + \int p \left(\frac{g_{\lambda}}{g_0} - 1 \right) g_0 dx \\
&\quad \text{since } \log(1+x) \leq x \text{ for } x > 0 \\
&\leq \int (\log g_0(x)) g_0(x) dx < \infty. \tag{59}
\end{aligned}$$

Combining (58) and (59) yields

$$\int |\log [(1-p)g_0(x) + pg_{\lambda}(x)]| g_0(x) < \infty,$$

which gives the first condition for asymptotic normality.

The second condition for asymptotic normality does not apply for our case since $\Theta(\boldsymbol{\lambda}) \neq \mathbb{R}^+$. Finally, the fourth condition for asymptotic normality is trivial due to the form of (52), which is linear in β so the first two partial derivatives in β are finite constants and the supremum for $0 \leq \beta \leq \rho$ in

$$\frac{\partial^3 \log g(x; \beta)}{\partial \beta^3} = \left(\frac{\frac{g_{\boldsymbol{\lambda}}}{g_0} - 1}{\|g_{\boldsymbol{\lambda}} g_0 - 1\|} \right)^3 \left(\beta \frac{\frac{g_{\boldsymbol{\lambda}}}{g_0} - 1}{\|g_{\boldsymbol{\lambda}} g_0 - 1\|} + g_0 \right)^{-3}$$

is attained when $\beta = 0$, which gives a finite constant.

Therefore, a mixture model S as in (50) that satisfies (54), (55) and (56) satisfies the conditions for S to be locally conic at g_0 as well as the conditions for asymptotic normality.

We have satisfied all of the regularity conditions for Theorem 1 from Fukumizu (2003), which can now be applied to achieve our result. \square

Fukumizu (2003) proves a version of the above theorem for the general case, then shows that the example of the Gaussian mixture model with two components satisfies the conditions of the theorem quite easily, thus showing an alternate proof for Hartigan (1985).

We will also show that for general 2-component mixture models, we can have a simplified sufficient condition for divergence of the likelihood ratio test statistic (51). That is the following theorem:

Theorem 4.3. *Let a statistical model $S = \{f(x|(\boldsymbol{\lambda}, \beta))\}$ be a mixture model as in (50) that satisfies (54), (55) and (56). Then if there exists a sequence $\{\boldsymbol{\lambda}_i\}_{i=1}^{\infty}$ such that $\|r_{\boldsymbol{\lambda}_i}\| \rightarrow \infty$ then for any $0 < M < \infty$, we have*

$$\lim_{n \rightarrow \infty} P \left(\sup_{(\boldsymbol{\lambda}, \beta)} L_n(\boldsymbol{\lambda}, \beta) \leq M \right) = 0.$$

Proof. Here, we show how our sufficient condition for Theorem 4.3, that $\|r_{\lambda_i}\| \rightarrow \infty$, implies the sufficient condition in Theorem 4.2, where the sequence of score functions $\nu_i \rightarrow 0$ in probability. This simplification is easy to verify for general 2-component mixture problems.

Suppose $X \sim f_0$ and that a sequence $\{\lambda_i\}$ is given. Write $r_i = r_{\lambda_i}$. We now show that if $\|r_i\| \rightarrow \infty$ then

$$s_i(X) = \frac{r_i(X) - 1}{\sqrt{\|r_i\|^2 - 1}} \xrightarrow{p} 0.$$

Fix $\epsilon > 0$. There exists $0 < N_0 < \infty$ such that for all $i > N_0$ we have

$$\frac{1}{\sqrt{\|r_i\|^2 - 1}} \leq \epsilon.$$

Then for $i > N_0$,

$$\begin{aligned} s_i(X) &= \frac{r_i(X) - 1}{\sqrt{\|r_i\|^2 - 1}} \\ &\geq \frac{-1}{\sqrt{\|r_i\|^2 - 1}} \\ &\geq -\epsilon. \end{aligned}$$

Therefore, for such i ,

$$\begin{aligned} P(|s_i(X)| > \epsilon) &= P(s_i(X) > \epsilon) \\ &\leq P\left(r_i(X) > \epsilon\sqrt{\|r_i\|^2 - 1}\right) \\ &\leq \frac{1}{\epsilon\sqrt{\|r_i\|^2 - 1}} \rightarrow 0 \text{ as } i \rightarrow \infty, \end{aligned}$$

by Markov's inequality, since $\mathbb{E}[r_i(X)] \equiv 1$ for all i .

This gives us the sufficient condition for Theorem 4.2, so that for arbitrary

$M > 0$, we have

$$\lim_{n \rightarrow \infty} P \left(\sup_{(\boldsymbol{\lambda}, \beta)} L_n(\boldsymbol{\lambda}, \beta) \leq M \right) = 0.$$

as required. □

Suppose we have a sample X_1, \dots, X_n of n *iid* observations from a two-component mixture of Markov chains parameterised as in (35) with density

$$f(x(\cdot) | \mathbf{d}, \mathbf{s}, \boldsymbol{\lambda}, J) = d_{x(0)} \{ (1 - s_{x(0)}) g(x(\cdot) | \boldsymbol{\lambda}_0, J) + s_{x(0)} g(x(\cdot) | \boldsymbol{\lambda}, J) \},$$

for some $\boldsymbol{\lambda} = \Gamma \boldsymbol{\lambda}_0$ with $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_w)$ such that $\gamma_j < \infty$ for $j \in 1, \dots, w$ and known $\boldsymbol{\lambda}_0$. Write $g(x(\cdot) | \boldsymbol{\lambda}, J)$ above as $f_{\boldsymbol{\lambda}}$ and define the density ratio $r_{\boldsymbol{\lambda}}(\nu)$ as

$$\begin{aligned} r_{\boldsymbol{\lambda}}(\nu) &= \frac{g_{\Gamma \boldsymbol{\lambda}_0}(\nu)}{g_{\boldsymbol{\lambda}_0}(\nu)} \\ &= \frac{\prod_{i=1}^w e^{-\gamma_i \lambda_{0,i} \tau_i} \prod_{j \neq i} (\gamma_i \lambda_{0,i})^{n_{ij}} J_{ij}^{n_{ij}}}{\prod_{i=1}^w e^{-\lambda_{0,i} \tau_i} \prod_{j \neq i} \lambda_{0,i}^{n_{ij}} J_{ij}^{n_{ij}}}. \end{aligned} \quad (60)$$

Note that this is well defined for all $\boldsymbol{\lambda}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}$, thus (54) is satisfied. If we set $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, it is clear that $\|r_{\boldsymbol{\lambda}}\|^2 = \int \left(\frac{g_{\boldsymbol{\lambda}}}{g_0} \right) d\mu = 1$. Also, if $\|r_{\boldsymbol{\lambda}}\| = 1$, then we must have that $\frac{g_{\boldsymbol{\lambda}}}{g_0} = 1$. This implies that $\gamma_i = 1$ for $i \in 1, \dots, w$, which in turn implies that (55) is satisfied. We write

$$\begin{aligned} \int |\log g_{\boldsymbol{\lambda}}(x) | g_0(x) dx &= \int \left| \sum_{i=1}^w (-\gamma_i \lambda_{0,i} \tau_i) \sum_{j \neq i} n_{ij} \log(\gamma_i \lambda_{0,i} J_{ij}) \right| g_0(x) dx \\ &< \infty \end{aligned}$$

as each $\tau_i \leq T < \infty$ and each $n_{ij} \leq \tilde{n} < \infty$ for all $i, j \in 1, 2, \dots, w$ and all $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$.

Thus, (56) is satisfied. We also wish to show that we can find a sequence $\{\lambda_i\}_{i=1}^{\infty}$ such that

$$\begin{aligned}\|r_{\lambda_i}\|^2 &= \int_{\mathcal{W}} r_{\lambda_i}^2(\nu) g_{\lambda_0}(\nu) d\sigma(\nu) \\ &= \int_{\mathcal{W}} \frac{g_{\Gamma_i \lambda_0}^2(\nu)}{g_{\lambda_0}(\nu)} d\sigma(\nu) \\ &\rightarrow \infty\end{aligned}$$

as $i \rightarrow \infty$. It suffices to consider each element of Γ to be the same, that is

$$\Gamma_i = \gamma_i I_w \text{ for } \gamma_i < \infty \text{ for all } i,$$

where I_w is the $w \times w$ identity matrix. Then for $\nu \in \mathcal{W}$, we have

$$\begin{aligned}\frac{g_{\Gamma \lambda_0}^2(\nu)}{g_{\lambda_0}(\nu)} &= \frac{\left(\prod_{i=1}^w e^{-\gamma \lambda_{0,i} \tau_i} \prod_{j \neq i} (\gamma \lambda_{0,i})^{n_{ij}} J_{ij}^{n_{ij}}\right)^2}{\prod_{i=1}^w e^{-\lambda_{0,i} \tau_i} \prod_{j \neq i} \lambda_{0,i}^{n_{ij}} J_{ij}^{n_{ij}}} \\ &= \prod_{i=1}^w e^{-(2\gamma-1)\lambda_{0,i} \tau_i} \prod_{j \neq i} \left(\frac{\gamma^2}{2\gamma-1} (2\gamma-1) \lambda_{0,i}\right)^{n_{ij}} J_{ij}^{n_{ij}} \\ &\geq \frac{1}{4} \left(\frac{1}{2\gamma-1}\right) g_{(2\gamma-1)\lambda_0}(\nu), \text{ for all cases } \tilde{n} = 0, 1, \dots\end{aligned}$$

Here we have

$$\|r_{\gamma \lambda_0}\|^2 \geq \frac{1}{4} \left(\frac{1}{2\gamma-1}\right).$$

Thus, for any sequence $\{\frac{1}{2} < \gamma_i < \infty\}$ such that $\gamma_i \rightarrow \frac{1}{2}$ we have $\|r_{\Gamma_i \lambda_0}\|^2 \rightarrow \infty$ as required. We can now apply Theorem 4.3 to prove that Proposition 4.1 holds true.

4.3.4 A special case with 2 states

Frydman (2005) presents mixtures of Markov chains that migrate among the discrete states $1, \dots, w$, where the w^{th} state is an absorbing state. We consider the simplest case where we have $w = 2$ states in total. In the context of modelling credit rating migrations, the first state represents “non-default”, the other represents “default”. As in Frydman (2005) (and widely in practice), we assume the second, “default” state is an absorbing state. The time until a transition from “non-default” to “default” is exponentially distributed. However, since a firm is only observed for a fixed time period then the default time may or may not occur in the observation period. Thus, a sample of independent observations of n such firms is *iid* with a censored exponential distribution.

Now, we consider a sample X_1, X_2, \dots, X_n of *iid* observations from a 2-component mixture of 2-state continuous-time Markov chains $X_0(t)$ and $X_1(t)$, observed from time 0 to T , starting in state 1 and with state 2 being an absorbing state. In this case, we do not require the parameter \mathbf{d} since all of our firms begin in the non-default state 1 (i.e. $\mathbf{d} = (1, 0)$) and our parameter \mathbf{s} , which is represented as

$$s_{i,m} = \frac{\pi_m d_{i,m}}{\sum_{m=1}^N \pi_m d_{i,m}} \text{ for } i = 1, 2 \text{ and } m = 1, 2$$

can be represented by a simple scalar p where

$$\begin{aligned} p &= P(X(t) = X_1(t) | X(0) = 1) \\ &= P(X(t) = X_1(t)) \text{ and} \\ 1 - p &= P(X(t) = X_0(t)). \end{aligned}$$

Note that we do not have multiple jumps for a particular observation. It either stays in state 1 for the entire observation window $[0, T]$ or it jumps once to the

absorbing state 2 within the observation window and remains there until the end of the observation window T . Thus, we can represent an observation $x(\cdot)$ using the time until the first jump out of the initial state with

$$x = \begin{cases} T_0 & \text{if } T_0 < T \text{ and} \\ T & \text{otherwise.} \end{cases}$$

This is (in effect) the amount of time that a firm is observed in the “non-default” state. Our density for a particular observation $x(\cdot) = x$ is then

$$f(x) = (1 - p)g(x|\lambda_0) + pg(x|\lambda_1) \quad (61)$$

where, $g(x|\lambda)$ is given by

$$g(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 < x < T \\ e^{-\lambda T} & \text{if } x = T \\ 0 & \text{otherwise.} \end{cases} \quad (62)$$

This is a censored exponential distribution, with rate parameter λ observed within a finite time window $(0, T]$.

For the hypothesis test, which tests between 1 and 2 mixture components for the two-state case

$$\begin{aligned} H_0 : X(\cdot) &\sim g(x|\lambda_0), \\ H_1 : X(\cdot) &\sim (1 - p)g(x|\lambda_1) + pg(x|\lambda_2) \end{aligned} \quad (63)$$

we define the log-likelihood ratio for n *iid* observations to be

$$\Lambda_n = \sum_{i=1}^n \left\{ \sup_{\lambda_1, \lambda_2, p} \log [(1 - p)g(x|\lambda_1) + pg(x|\lambda_2)] - \sup_{\lambda_0} \log [g(x|\lambda_0)] \right\}.$$

From (48), we see that in the test between 1 and 2 components we have

$$\Lambda_n \geq \Lambda_1 + \mathcal{O}_p(1), \quad (64)$$

where

$$\Lambda_1 = \sup_{\lambda_1, p} \sum_{i=1}^n \log \left[\frac{(1-p)g(x|\lambda_0) + pg(x|\lambda_1)}{g(x|\lambda_0)} \right]. \quad (65)$$

If Λ_1 diverges as the sample size $n \rightarrow \infty$ then this implies Λ_n also diverges. We have shown that Λ_1 does in fact diverge, using Theorem 4.3. Understanding the rate of divergence of Λ_1 will give us a substantial insight into the rate of divergence for Λ_n . From (64), we see that if Λ_1 diverges at a rate $R_{1,n}$, then Λ_n diverges at a rate $R_{2,n} \geq R_{1,n}$. For the case where $w = 2$ we can go beyond the rate to find the exact limiting distribution of Λ_1 .

Without loss of generality, we will also assume that the true value under the null hypothesis for λ_0 is 1. Then we can present the following theorem, with proof to be provided in the following chapter:

Theorem 4.4. *If we let*

$$\Lambda = \sup_{\lambda > 0, 0 \leq p \leq 1} \sum_{k=1}^n \{ \log [(1-p)g(x_k|1) + pg(x_k|\lambda)] - \log g(x_k|1) \},$$

where $g(x|\lambda)$ is given by (62) above, then

$$\lim_{n \rightarrow \infty} P \{ 2\Lambda - 2 \log \log n + \log(16\pi^2) \leq x \} = e^{-e^{-x/2}}.$$

5 Censored exponential mixture detection

In this chapter, we continue with the motivating application of the previous chapter, where we are faced with the problem of modelling the non-homogeneous dynamics of credit rating migrations of firms. We are focussed on different regimes applying to different segments of the population, rather than different regimes over time. The theoretical developments in this chapter contribute towards establishing a proof of Theorem 4.4, which states the exact limiting distribution of the log-likelihood ratio test statistic for the test between 1 and 2 component mixtures of Markov chains, which each have 2 states, with the second state being an absorbing state. The challenges of testing between 1 and 2 component mixtures using the likelihood ratio test were explored for location mixtures of normal distributions in Hartigan (1985), which proves that the log-likelihood ratio diverges in probability and conjectures that the rate of divergence is of the order $\log \log n$ where n is the sample size. This conjecture was later proven in Bickel and Chernoff (1993) and Liu and Shao (2004). The problem of finding the limiting distribution was addressed for location mixtures of normal distributions in Garel (2005) and for mixtures of gamma distributions in Liu et al. (2003). Although there have been some studies that work towards a general solution, under particular regularity conditions, such as Dacunha-Castelle and Gassiat (1997) and Liu and Shao (2003), there remains a gap in the theory for our specific problem of testing between 1 and 2 Markov chain mixture components with 2 states, one of which is an absorbing state. It is motivated by a simple case of our practical example from Frydman (2005) and our key result is that we go beyond our findings in the previous chapter, where we proved that the log-likelihood ratio test statistic diverges to infinity as the sample size $n \rightarrow \infty$, to successfully derive its rate of divergence and exact limiting distribution. We find that this problem can be reframed as a test between a 1 and 2 component mixture of censored exponentials and so is more broadly applicable

than just to our Markov chain context.

We follow a similar strategy to Liu et al. (2003) and solve some key theoretical challenges that arise from the fact that our practical application requires that we have censoring (due to the finite observation window on our data). Liu et al. (2003) derive the limiting distribution of the log-likelihood ratio test statistic for testing between 1 and 2 components in a scale mixture of gamma distributions, with the constraint that the scale parameter of the second unknown component is greater than the scale parameter of the first known component. Technical difficulties prevent them from dealing with the two-sided version of the test. The log-likelihood ratio test statistic is shown in Liu et al. (2003) to be asymptotically equivalent to the square of the maximum of a stationary Gaussian process over an interval whose length increases as the logarithm of the sample size. The stationarity of the Gaussian process is crucial to their derivation of the limiting distribution of the statistic. The corresponding process in the censored case is no longer stationary and so in order to use the same general strategy of Liu et al. (2003) some new tools are required. Such tools are provided by the locally stationary Gaussian process extreme value theory developed by Hüsler (1990). One obstacle to the use of these tools is the potentially difficult verification that a given Gaussian process is indeed in the *locally stationary* class. Our Lemma 5.9 achieves this for the Gaussian process we consider by showing that certain higher-order derivatives of its correlation function are uniformly controlled.

A happy consequence of the censoring is that we are able to consider the two-sided version of the testing problem. We are able to elegantly extend the methods of Liu et al. (2003) to analyse the maximum of the log-likelihood ratio statistic over this extended range, thus removing the rather restrictive one-sided constraint that Liu et al. (2003) are forced to adhere to in the uncensored version of the problem. We then use this result to derive the exact limiting distribution of the log-likelihood

ratio test statistic, thus solving the outstanding practical problem from Frydman (2005). After providing an overview of the testing problem in Section 5.1, we work in Section 5.2 to establish our key results. We then provide the detailed proofs of these results in Section 5.3.

5.1 An overview of the testing problem

Censored exponentials are widely used in practice for modelling time-to-event data where events occur with a constant underlying rate over a given finite time window $(0, T]$. In the previous chapter we studied a problem that was motivated by the application of modelling credit rating migration dynamics of firms, which involved testing for mixtures of discrete-state Markov chains with an absorbing state observed continuously over a finite time period. In the simplest case when the Markov chain has 2 states, the time to absorption has a censored exponential distribution. We consider the problem of testing for the existence of a mixture of censored exponentials. Specifically, we study the asymptotics of the log-likelihood ratio test statistic for testing between 1 and 2 mixture components and show that it diverges in probability at a rate of $\log \log n$, where n is the sample size.

Let X_1, X_2, \dots, X_n be an independent and exponentially distributed sample with rate parameter λ . Since we are only observing the data from time 0 to T , we define $Y_i = \min(X_i, T)$, so that Y_1, Y_2, \dots, Y_n is an *iid* sample from a censored exponential distribution. We thus have the cumulative distribution function

$$G_\lambda(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - e^{-\lambda y} & \text{if } 0 \leq y < T \\ 1 & \text{if } y \geq T. \end{cases}$$

This distribution has a density

$$g_\lambda(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } 0 < y < T \\ e^{-\lambda T} & \text{if } y = T \\ 0 & \text{otherwise,} \end{cases}$$

with respect to a dominating measure given by the sum of Lebesgue measure on $[0, \infty)$ and counting measure on $\{T\}$. The expectation operator with respect to this density is given by

$$\mathbb{E}[f(Y_1)] = \int f g_\lambda d\mu = \int_0^T f(y) \lambda e^{-\lambda y} dy + f(T) e^{-\lambda T} \quad (66)$$

where $\mu(A) = \mathcal{L}(A) + \mathbb{1}\{T \in A\}$ with $\mathcal{L}(\cdot)$ the Lebesgue measure.

The log-likelihood of a series of observations $\mathbf{y} = (y_1, y_2, \dots, y_n)$ can thus be written as

$$L_n^{(1)}(\lambda | \mathbf{y}, T) = \sum_{i=1}^n \log(\lambda e^{-\lambda y_i} \mathbb{1}\{y_i < T\} + e^{-\lambda T} \mathbb{1}\{y_i = T\}). \quad (67)$$

The corresponding 2-component mixture distribution, where each observation y has density $(1-p)g(y|\lambda_0, T) + pg(y|\lambda, T)$, yields a log-likelihood for n *iid* observations $\mathbf{y} = (y_1, y_2, \dots, y_n)$ as follows

$$L_n^{(2)}(p, \lambda_0, \lambda | \mathbf{y}, T) = \sum_{i=1}^n \log([(1-p)\lambda_0 e^{-\lambda_0 y_i} + p\lambda e^{-\lambda y_i}] \mathbb{1}\{y_i < T\} + [(1-p)e^{-\lambda_0 T} + p e^{-\lambda T}] \mathbb{1}\{y_i = T\}). \quad (68)$$

We are interested in the testing problem

$$\begin{aligned} H_0 : Y_1 &\sim G_{\lambda_0}, \text{ for } \lambda_0 > 0 \text{ known, against} \\ H_1 : Y_1 &\sim (1-p)G_{\lambda_0} + pG_\lambda \text{ for } p \in (0, 1], \lambda > 0 \text{ both unknown,} \end{aligned} \quad (69)$$

where without loss of generality, we may take $\lambda_0 = 1$. For convenience, we write $g = g_1$. We write the log-likelihood ratio test statistic as

$$\begin{aligned}
\Lambda_n &= \sup_{p,\lambda} L_n(p, \lambda) = \sup_{p,\lambda} \{L_n^{(2)}(p, 1, \lambda|\mathbf{Y}, T) - L_n^{(1)}(1|\mathbf{Y}, T)\} \\
&= \sup_{p,\lambda} \sum_{i=1}^n \log \left[\frac{(1-p)g(Y_i) + pg_\lambda(Y_i)}{g(Y_i)} \right] \\
&= \sup_{p,\lambda} \sum_{i=1}^n \log [1 + pZ_i(\lambda)], \tag{70}
\end{aligned}$$

where

$$\begin{aligned}
Z_i(\lambda) &= \frac{g_\lambda(Y_i)}{g(Y_i)} - 1 \\
&= \{\lambda e^{-(\lambda-1)Y_i} \mathbb{1}\{Y_i < T\} + e^{-(\lambda-1)T} \mathbb{1}\{Y_i = T\}\} - 1. \tag{71}
\end{aligned}$$

From (66), we calculate the expected value and variance of $Z_1(\lambda)$ under the single component density as

$$\begin{aligned}
\mathbb{E}\{Z_1(\lambda)\} &= \mathbb{E} \left\{ \frac{g_\lambda(Y_1)}{g(Y_1)} - 1 \right\} = 0 \quad \text{and} \\
\mathbb{V}\text{ar}\{Z_1(\lambda)\} &= \mathbb{E} \left[(\lambda e^{-(\lambda-1)Y_1} \mathbb{1}\{Y_1 < T\} + e^{-(\lambda-1)T} \mathbb{1}\{Y_1 = T\} - 1)^2 \right] \\
&= \int_0^T \lambda^2 e^{-(2\lambda-1)y} dy + e^{-(2\lambda-1)T} - 1 \\
&= \left(\frac{\lambda^2}{2\lambda-1} - 1 \right) (1 - e^{-(2\lambda-1)T}). \tag{72}
\end{aligned}$$

5.2 Testing homogeneity in censored exponential mixture models

In Liu et al. (2003), the asymptotic distribution for the log-likelihood ratio test statistic for a one-sided test between a 1 and 2 component scale mixture of gamma

distributions is derived. The general approach of the paper follows that of Bickel and Chernoff (1993) and Liu and Shao (2003) for the analogous normal location mixture problem. Specifically, the profile log-likelihood, obtained by maximising only over p , is firstly approximated by the square of a standardised *score process*. The asymptotic distribution of the maximum of the score process with respect to λ is found to be the same as the maximum of a stationary Gaussian process over an interval of length $\log n$. The square of such a maximum can be represented as $G_n + \log \log n$, where G_n has an asymptotic Gumbel distribution. This approximation is shown to be suitably accurate so that the log-likelihood ratio statistic inherits the same limiting distribution. Note that this strategy can only hope to be successful for the one-sided version of the gamma scale mixture problem they consider. This is because for $\lambda < \frac{1}{2}$, the variance of the score process is infinite and the convergence to a Gaussian process fails. Taking $\kappa = 1$ in Liu et al. (2003) results in a one-sided uncensored version of our problem. We use the same general strategy as Liu et al. (2003) however several of their steps need new tools for application to our case, as foreshadowed in the introduction. We will establish our notation and present our key results in this section before providing details of original proofs in the next section.

Suppose Y_1, \dots, Y_n are *iid* random variables from a censored exponential distribution with rate parameter $\lambda = 1$, with density $g(y)$. We interpret $Y_i = G^{-1}(U_i)$ for uniform random variables U_1, \dots, U_n , where

$$G^{-1}(u) = \begin{cases} -\log(1 - u) & \text{for } 0 \leq u \leq 1 - e^{-T} \\ T & \text{for } 1 - e^{-T} < u \leq 1, \end{cases}$$

is the inverse cumulative distribution function of $G(\cdot)$. The test at (69) becomes

$$\begin{aligned} H_0 : Y_1 &\sim G \text{ against} \\ H_1 : Y_1 &\sim (1-p)G + pG_\lambda. \end{aligned} \tag{73}$$

Recall the definition of the norm $\|\cdot\|^2$ in (49). We write the likelihood ratio as $l_\lambda(y) = g_\lambda(y)/g(y)$ and define the standardised score process

$$\begin{aligned} S_n(\lambda) &= n^{-\frac{1}{2}} \sum_{i=1}^n \frac{[\lambda e^{-Y_i(\lambda-1)} \mathbb{I}\{Y_i < T\} + e^{-(\lambda-1)T} \mathbb{I}\{y_i \geq T\} - 1]}{\left[\left(\frac{\lambda^2}{2\lambda-1} - 1\right)(1 - e^{-(2\lambda-1)T})\right]^{\frac{1}{2}}} \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \frac{l_\lambda(Y_i) - 1}{\sqrt{\|l_\lambda\|^2 - 1}} \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n s_\lambda(Y_i). \end{aligned} \tag{74}$$

We will show that Λ_n has the same asymptotic distribution as $\frac{1}{2}M_n^2$, where Λ_n is the log-likelihood ratio test statistic in (70) and $M_n = S_n(\hat{\lambda})$ with $\hat{\lambda} = \operatorname{argmax}_{\lambda>0} S_n(\lambda)$.

Let $F_n(u)$ be the empirical cumulative distribution function of U_1, \dots, U_n . It is possible to define these on a suitable probability space together with a sequence of Weiner processes $\{W_n(u)\}$ so that the corresponding empirical process $\alpha_n(u) = \sqrt{n}[F_n(u) - u]$ is well approximated by the Brownian bridge $B_n(u) = W_n(u) - W_n(1)u$ (see (102) where this is made more precise).

For any function $g(\cdot)$ on $[0, 1]$ of bounded variation, we may, using an integration-by-parts formula, define the stochastic integral

$$\begin{aligned} \int_0^1 g(u) dB_n(u) &= - \int_0^1 B_n(u) dg(u) \\ &= - \int_0^1 [W_n(u) - W_n(1)u] dg(u) \\ &= W_n(1)g(1) - \int_0^1 W_n(u) dg(u) \end{aligned}$$

in terms of ordinary Riemann-Stieltjes integrals. We show in Appendix A that $\int_0^1 g(u)dB_n(u)$ is mean zero Gaussian for each $g(\cdot)$ and for any other $h(\cdot)$ of bounded variation we have

$$\mathbb{E} \left\{ \int_0^1 g(u)dB_n(u) \int_0^1 h(u)dB_n(u) \right\} = \int_0^1 g(u)h(u)du. \quad (75)$$

We may write our standardised score process from (74) as

$$S_n(\lambda) = \int_0^1 s_\lambda(G^{-1}(u))d\alpha_n(u) \quad (76)$$

$$= H_n(\lambda) + R_n(\lambda), \quad (77)$$

where

$$H_n(\lambda) = \int_0^1 s_\lambda(G^{-1}(u))dB_n(u) \text{ and} \quad (78)$$

$$R_n(\lambda) = \int_0^1 s_\lambda(G^{-1}(u))d[\alpha_n(u) - B_n(u)]. \quad (79)$$

In Liu et al. (2003), the score process is approximated by a Gaussian process, which after a certain transformation becomes stationary. In the censored case that we study here, the same transformation may be used; however, the approximating Gaussian process $\{H_n(e^s + \frac{1}{2}), -\log 2 \leq s < \infty\}$ is not stationary. It is however *locally stationary* in the sense of Berman (1985) and Hüsler (1990) (this is verified in Lemma 5.9). Theorem 4.2 of Hüsler (1990) then yields the following lemma. Using the same general strategy of Liu et al. (2003) we have developed analogues of the Lemmas and Theorems in their paper. It should be noted however that several of our proofs differ substantially from their analogues in Liu et al. (2003).

Lemma 5.1. *The Gaussian process $\{H_n(e^s + \frac{1}{2}), -\log 2 \leq s < \infty\}$ satisfies*

$$\lim_{C \rightarrow \infty} P \left\{ A_C \left[\sup_{-\log 2 \leq s \leq C} H_n(e^s + \frac{1}{2}) - A_C \right] + \log(4\pi) \leq y \right\} = e^{-e^{-y}}, \quad (80)$$

where $A_C = (2 \log C)^{\frac{1}{2}}$.

We show that, within the range of λ where the maximum of $H_n(\lambda)$ is attained, with probability tending to 1, the supremum of S_n is asymptotically equivalent to the supremum of $H_n(\lambda)$. We split up the parameter space into separate intervals to prove this with the following lemma. Let us write $\log_{(2)} n = \log \log n$ and $\log_{(3)} n = \log \log \log n$ for large enough n .

Lemma 5.2. *In a suitable probability space,*

$$\begin{aligned} \sup_{\lambda \in [1, \log n] \cup [n(\log n)^{-4}, \infty)} S_n(\lambda) \vee 0 &= \mathcal{O}_p(1) (\log_{(3)} n)^{\frac{1}{2}} \\ \sup_{\lambda \in [\log n, n(\log n)^{-4}]} |S_n(\lambda) - H_n(\lambda)| &= \mathcal{O}_p(1) (\log n)^{-1}. \end{aligned} \quad (81)$$

Then, the asymptotic distribution for $M_n = S_n(\hat{\lambda})$ is derived using Lemma 5.1 and Lemma 5.2 with the following theorem.

Theorem 5.3. *Under the null hypothesis for the test (73), $M_n = \sup_{\lambda \geq 0} S_n(\lambda)$ satisfies*

$$\lim_{n \rightarrow \infty} P \left\{ \sqrt{2 \log_{(2)} n} \left(M_n - \sqrt{2 \log_{(2)} n} \right) + \log(4\pi) \leq y \right\} = e^{-e^{-y}}.$$

Moreover, the asymptotic distribution of M_n can also be expressed as

$$\lim_{n \rightarrow \infty} P \left\{ M_n^2 - 2 \log_{(2)} n + \log(16\pi^2) \leq y \right\} = e^{-e^{-y/2}}.$$

We now wish to show that the log-likelihood ratio test statistic in (70) has the

same asymptotic distribution as $\frac{1}{2}M_n^2$. The log-likelihood ratio test statistic

$$\Lambda_n = \sup_{p,\lambda} \sum_{i=1}^n \log \{1 + pZ_i(\lambda)\} = \sup_{p,\lambda} \sum_{i=1}^n \log \{1 + p\|Z_i(\lambda)\|s_\lambda(Y_i)\}. \quad (82)$$

We examine the standardised score process $S_n(\lambda)$ over the region $0 < \lambda \leq 1$ and then over the region $1 < \lambda < \lambda^*$ for some constant $1 < \lambda^* < \infty$. Here,

$$\begin{aligned} s_\lambda(y) &= \frac{\lambda^{\mathbb{I}\{y < T\}} e^{-y(\lambda-1)} - 1}{\sqrt{\frac{(\lambda-1)^2}{2\lambda-1} (1 - e^{-T(2\lambda-1)})}} \\ &= \frac{\lambda^{\mathbb{I}\{y < T\}} e^{-y(\lambda-1)} - 1}{|1 - \lambda|} \sqrt{\frac{2\lambda - 1}{1 - e^{-T(2\lambda-1)}}}. \end{aligned} \quad (83)$$

Now the second factor in (83)

$$\sqrt{\frac{2\lambda - 1}{1 - e^{-T(2\lambda-1)}}} \rightarrow \frac{1}{\sqrt{1 - e^{-T}}}$$

as $\lambda \uparrow 1$. Writing $\lambda = 1 - \epsilon$ (for $\epsilon > 0$), the first factor in (83) becomes

$$\frac{e^{y\epsilon} - 1 - \epsilon e^{y\epsilon}}{\epsilon} = \frac{e^{y\epsilon} - 1}{\epsilon} - e^{y\epsilon} \rightarrow y - 1$$

for $y < T$ and

$$\frac{e^{T\epsilon} - 1}{\epsilon} \rightarrow T$$

for $y = T$ as $\epsilon \downarrow 0$. Thus

$$\lim_{\lambda \uparrow 1} s_\lambda(y) = \frac{y - \mathbb{I}\{y < T\}}{\sqrt{1 - e^{-T}}}.$$

As $\lambda \downarrow 0$,

$$s_\lambda(y) \rightarrow \begin{cases} -\frac{1}{\sqrt{e^T-1}} & \text{for } y < T, \\ -\frac{e^T-1}{\sqrt{e^T-1}} & \text{for } y = T. \end{cases}$$

We thus define $s_0(y)$ and $s_1(y)$ accordingly. Moreover, for all $0 \leq \lambda \leq 1$, $s_\lambda(y)$ is a non-decreasing function over $0 < y < T$. Thus

$$\begin{aligned} \inf_{0 < y \leq T} s_\lambda(y) &= \lim_{y \downarrow 0} s_\lambda(y) \\ &= -\sqrt{\frac{2\lambda-1}{1-e^{-T(2\lambda-1)}}} \\ &\geq -\frac{1}{\sqrt{1-e^{-T}}} \end{aligned} \tag{84}$$

since (84) is minimised at $\lambda = 1$. Also

$$\begin{aligned} \sup_{0 < y \leq T} s_\lambda(y) &= s_\lambda(T) \\ &= \frac{e^{T(1-\lambda)}-1}{1-\lambda} \sqrt{\frac{2\lambda-1}{1-e^{-T(2\lambda-1)}}} \\ &\leq \sqrt{e^T-1} \end{aligned} \tag{85}$$

since (85) is maximised at $\lambda = 0$. We have thus shown that the set of functions $\{s_\lambda(y) : 0 < y \leq T, 0 \leq \lambda \leq 1\}$ are monotone and all take values in the fixed closed interval $\left[-\frac{1}{\sqrt{1-e^{-T}}}, \sqrt{e^T-1}\right]$.

Now, for the case where $1 < \lambda < \lambda^*$, writing $\lambda = 1 + \epsilon$ (for $\epsilon > 0$), the first factor of (83) becomes

$$\frac{e^{-y\epsilon} - 1 + \epsilon e^{-y\epsilon}}{\epsilon} = \frac{e^{-y\epsilon} - 1}{\epsilon} + e^{-y\epsilon} \rightarrow -y + 1$$

as $\epsilon \rightarrow 0$ for $y < T$, and

$$\frac{e^{-T\epsilon} - 1}{\epsilon} \rightarrow -T \text{ as } \epsilon \rightarrow 0 \text{ for } y = T.$$

Thus,

$$\lim_{\lambda \downarrow 1} s_\lambda(y) = \frac{\mathbb{I}\{y < T\} - y}{\sqrt{1 - e^{-T}}} = -\lim_{\lambda \uparrow 1} s_\lambda(y).$$

As $\lambda \uparrow \lambda^*$,

$$s_\lambda(y) \rightarrow \begin{cases} \frac{\lambda^* e^{-y(\lambda^*-1)} - 1}{\lambda^* - 1} \sqrt{\frac{2\lambda^* - 1}{1 - e^{-T(2\lambda^*-1)}}} & \text{for } y < T, \\ \frac{e^{-T(\lambda^*-1)} - 1}{\lambda^* - 1} \sqrt{\frac{2\lambda^* - 1}{1 - e^{-T(2\lambda^*-1)}}} & \text{for } y = T. \end{cases}$$

For all $1 < \lambda < \lambda^*$, $s_\lambda(y)$ is a non-increasing function over $0 < y < T$. Thus,

$$\begin{aligned} \inf_{0 < y \leq T} s_\lambda(y) &= \lim_{y \uparrow T} s_\lambda(y) \\ &= \frac{e^{-T(\lambda-1)} - 1}{\lambda - 1} \sqrt{\frac{2\lambda - 1}{1 - e^{-T(2\lambda-1)}}} \end{aligned} \quad (86)$$

$$\geq -T \sqrt{\frac{1}{1 - e^{-T}}}, \quad (87)$$

since (86) is minimised at $\lambda = 1$. Also,

$$\begin{aligned} \sup_{0 < y \leq T} s_\lambda(y) &= \lim_{y \downarrow 0} s_\lambda(y) \\ &= \sqrt{\frac{2\lambda - 1}{1 - e^{-T(2\lambda-1)}}} \end{aligned} \quad (88)$$

$$\leq \sqrt{\frac{2\lambda^* - 1}{1 - e^{-T(2\lambda^*-1)}}}, \quad (89)$$

since (88) is maximised at $\lambda = \lambda^*$. We have thus shown that the set of functions $\{s_\lambda(y) : 0 < y \leq T, 1 \leq \lambda \leq \lambda^*\}$ are monotone and all take values in the fixed closed interval $\left[-T \sqrt{\frac{1}{1 - e^{-T}}}, \sqrt{\frac{2\lambda^* - 1}{1 - e^{-T(2\lambda^*-1)}}}\right]$. Thus by Example 19.11 in van der

Vaart (1998) they form a (universal) Donsker class so that the corresponding empirical process $\{S_n(\lambda) : 0 \leq \lambda \leq 1\}$ given by

$$S_n(\lambda) = n^{-\frac{1}{2}} \sum_{i=1}^n s_\lambda(Y_i)$$

converges (in the space of bounded functions on $[0, 1]$ under the uniform norm) to a tight Gaussian process. In particular

$$\sup_{0 \leq \lambda \leq 1} S_n(\lambda) = \mathcal{O}_p(1). \quad (90)$$

Similarly, we have

$$\sup_{1 \leq \lambda \leq \lambda^*} S_n(\lambda) = \mathcal{O}_p(1). \quad (91)$$

Since the functions $\{s_\lambda(y); 0 < y \leq T, 0 \leq \lambda \leq \lambda^*\}$ are uniformly bounded, they trivially have a square integrable envelope function; this, along with the fact that they form a Donsker class, implies by Theorem 3.1 from Liu and Shao (2003) that

$$\sup_{0 \leq \lambda < \lambda^*} \Lambda_n = \mathcal{O}_p(1). \quad (92)$$

Let us write our vector of random variables Y_1, \dots, Y_n in ascending order to form the order statistics $Y_{1,n}, \dots, Y_{n,n}$. Taking the partial derivative of the log-likelihood ratio yields

$$\frac{\partial}{\partial \lambda} L_n(p, \lambda) = \sum_{i=1}^n \frac{(1 - \lambda y) e^{-(\lambda-1)y} \mathbb{I}\{y < T\} - T e^{-(\lambda-1)T} \mathbb{I}\{y = T\}}{1 + p Z_i(\lambda)}.$$

When $\lambda > \frac{1}{Y_{1,n}}$, we have $\frac{\partial}{\partial \lambda} L_n(p, \lambda) < 0$. Since we know, from Theorem 4.3 in the

previous chapter, that the log-likelihood ratio test statistic diverges to infinity in probability, it suffices to maximise $L_n(p, \lambda)$ for $\lambda \in (0, \frac{1}{Y_{1,n}}]$.

Let $I_n = [\log n, n(\log n)^{-4}]$ and $I_n^* = [\lambda^*, \log n] \cup [n(\log n)^{-4}, \frac{1}{Y_{1,n}}]$. Then we have the following lemmas

Lemma 5.4. *Under the constraint that $L_n(p, \lambda) > 0$,*

$$\sup_{\lambda \in I_n \cup I_n^*} p\lambda = \mathcal{O}(1).$$

Lemma 5.5. *Letting $P_n s_\lambda^2 = n^{-1} \sum_{i=1}^n s_\lambda^2(Y_i)$, we have*

$$\sup_{\lambda \in I_n \cup I_n^*} \frac{1}{P_n s_\lambda^2} = \mathcal{O}_p(1).$$

Moreover, when $\lambda \in I_n$, $P_n s_\lambda^2 = 1 + \mathcal{O}_p((\log n)^{-\frac{1}{2}})$.

Now, by (90), Lemma 5.1 and Lemma 5.2, the supremum of $S_n(\lambda)$ is found when $\lambda \in I_n$. Therefore, we prove $2\Lambda_n = M_n^2 + o_p(1)$. Using (92) and Theorem 5.3, we have the following theorem on the asymptotic distribution of the log-likelihood ratio test statistic Λ_n .

Theorem 5.6. *The log-likelihood ratio test statistic for the test (73) satisfies $2\Lambda_n = M_n^2 + o_p(1)$ and*

$$\lim_{n \rightarrow \infty} P \{ 2\Lambda_n - 2 \log_{(2)} n + \log(16\pi^2) \leq y \} = e^{-e^{-y/2}}.$$

5.3 Details of original proofs

Lemma 5.7. *Suppose we have a function $D(s, t)$ which is symmetric in its arguments $s, t \geq k$ for some constant $-\infty < k < \infty$ and that we define the function*

$$\rho(s, t) = \frac{D(s, t)}{\sqrt{D(s, s)D(t, t)}} \quad (93)$$

with derivatives denoted by

$$D_{ij}(s, t) = \frac{\partial^{i+j} D(s, t)}{\partial s^i \partial t^j}.$$

Then if

1. the $(i, j)^{\text{th}}$ derivative exists and is continuous for all integers $i, j \geq 0$ and $i + j \leq 3$;
2. $\sup_{s, t \geq k} |D_{ij}(s, t)| < \infty$ for $i, j \geq 0$ such that $1 \leq i, j \leq 3$;
3. for two constants $0 < a < b < \infty$, $a \leq D(s, t) \leq b$ for all $k \leq s, t < \infty$;

for $t, t + \Delta \geq k$ we have the representation

$$\rho(t, t + \Delta) = 1 - \frac{V(t)}{2} \Delta^2 + \Delta^2 R_n(t, \Delta) \quad (94)$$

where

$$V(t) = \frac{D_{11}(t, t)}{D(t, t)} - \left[\frac{D_{01}(t, t)}{D(t, t)} \right]^2$$

satisfies

$$\sup_{t \geq k} |V(t)| < \infty \quad (95)$$

and

$$\lim_{\Delta \rightarrow 0} \sup_{k \leq t, t+\Delta < \infty} |R_n(t, \Delta)| = 0. \quad (96)$$

That is to say as $\Delta \rightarrow 0$, $R_n(t, \Delta) = o(1)$ uniformly in $t \geq k$.

Proof. By assumption 1 for some $0 \leq \alpha_1 \leq 1$ we have

$$\begin{aligned} D(t, t + \Delta) &= D(t, t) + \Delta D_{01}(t, t) + \frac{\Delta^2}{2} D_{02}(t, t + \alpha_1 \Delta) \\ &= D(t, t) + \Delta D_{01}(t, t) + \frac{\Delta^2}{2} \{D_{02}(t, t) + [D_{02}(t, t + \alpha_1 \Delta) - D_{02}(t, t)]\} \\ &= D(t, t) + \Delta D_{01}(t, t) + \frac{\Delta^2}{2} D_{02}(t, t) + \Delta^2 R_{n1}(t, \Delta) \end{aligned}$$

where $R_{n1}(t, \Delta) = o(1)$ uniformly in $t \geq k$ by assumption 2. By assumption 3, we can then also say

$$\frac{D(t, t + \Delta)}{D(t, t)} = 1 + \Delta \frac{D_{01}(t, t)}{D(t, t)} + \frac{\Delta^2}{2} \frac{D_{02}(t, t)}{D(t, t)} + \Delta^2 R_{n2}(t, \Delta) \quad (97)$$

where $R_{n2}(t, \Delta) = o(1)$ uniformly in $t \geq k$. Similarly we also have

$$\begin{aligned} D(t + \Delta, t + \Delta) &= D(t, t) + \Delta [D_{01}(t, t) + D_{10}(t, t)] \\ &\quad + \frac{\Delta^2}{2} [D_{02}(t, t) + 2D_{11}(t, t) + D_{20}(t, t)] + \Delta^2 R_{n3}(t, \Delta) \\ &= D(t, t) + 2\Delta D_{01}(t, t) + \Delta^2 [D_{02}(t, t) + D_{11}(t, t)] + \Delta^2 R_{n3}(t, \Delta) \end{aligned}$$

due to the symmetrical property of $D(s, t)$ and

$$\begin{aligned} \frac{D(t + \Delta, t + \Delta)}{D(t, t)} &= 1 + 2\Delta \frac{D_{01}(t, t)}{D(t, t)} + \Delta^2 \frac{D_{02}(t, t) + D_{11}(t, t)}{D(t, t)} + \Delta^2 R_{n4}(t, \Delta) \\ &= 1 + \Delta Q_{n1}(t, \Delta) \end{aligned}$$

where uniformly in $t \geq k$, $Q_{n1}(t, \Delta) = \mathcal{O}(1)$ and $R_{nj}(t, \Delta) = o(1)$ for both $j = 3, 4$.

For small enough Δ , applying a Taylor expansion we have (for some $0 \leq \alpha_2 \leq 1$)

$$\begin{aligned}
& \left[\frac{D(t + \Delta, t + \Delta)}{D(t, t)} \right]^{-\frac{1}{2}} \\
= & [1 + \Delta Q_{n1}(t, \Delta)]^{-\frac{1}{2}} \\
= & 1 - \frac{\Delta Q_{n1}(t, \Delta)}{2} + \frac{\Delta^2 Q_{n1}(t, \Delta)^2}{8} [1 + \alpha_2 \Delta Q_{n1}(t, \Delta)]^{-\frac{5}{2}} \\
= & 1 - \frac{\Delta Q_{n1}(t, \Delta)}{2} + \frac{\Delta^2 Q_{n1}(t, \Delta)^2}{8} [1 + \Delta Q_{n2}(t, \Delta)] \\
= & 1 - \frac{\Delta Q_{n1}(t, \Delta)}{2} + \frac{\Delta^2 Q_{n1}(t, \Delta)^2}{8} + \Delta^2 R_{n5}(t, \Delta) \\
= & 1 - \Delta \frac{D_{01}(t, t)}{D(t, t)} - \frac{\Delta^2}{2} \frac{D_{02}(t, t) + D_{11}(t, t)}{D(t, t)} - \frac{\Delta^2}{2} R_{n4}(t, \Delta) \\
& + \frac{1}{8} \left\{ 2\Delta \frac{D_{01}(t, t)}{D(t, t)} + \Delta^2 \frac{D_{02}(t, t) + D_{11}(t, t)}{D(t, t)} + \Delta^2 R_{n4}(t, \Delta) \right\}^2 + \Delta^2 R_{n5}(t, \Delta) \\
= & 1 - \Delta \frac{D_{01}(t, t)}{D(t, t)} - \frac{\Delta^2}{2} \left\{ \frac{D_{02}(t, t) + D_{11}(t, t)}{D(t, t)} - \left[\frac{D_{01}(t, t)}{D(t, t)} \right]^2 \right\} + \Delta^2 R_{n6}(t, \Delta)
\end{aligned}$$

where uniformly in $t \geq k$, $Q_{n2}(t, \Delta) = \mathcal{O}(1)$ and $R_{nj}(t, \Delta) = o(1)$ for $j = 5, 6$. Multiplying this by (97), the boundedness of the ratios $\frac{D_{ij}(t, t)}{D(t, t)}$ gives the result of (94) satisfying (95) and (96). □

Corollary 5.8. *If a function $\rho(s, t) = \rho_1(s, t)\rho_2(s, t)$ is the product of two functions admitting a representation of the form (94) satisfying (95) and (96), so that*

$$\rho_j(t, t + \Delta) = 1 - \frac{V_j(t)}{2} \Delta^2 + o(\Delta^2)$$

with $o(\Delta^2)$ uniform in $t \geq k$ for both $j = 1, 2$, then

$$\rho(t, t + \Delta) = 1 - \frac{[V_1(t) + V_2(t)]}{2} \Delta^2 + o(\Delta^2)$$

again with $o(\Delta^2)$ uniform in $t \geq k$.

Lemma 5.9. *The correlation function of the Gaussian process $\{H_n(e^s + \frac{1}{2}), -\log 2 \leq s < \infty\}$ is of the form (94) satisfying (95) and (96) with $k = -\log 2$.*

Proof. Since our standardised score function $s_\lambda(\cdot)$ is of bounded variation, we may use our result in (75) for our Gaussian process

$$H_n(\lambda) = \int_0^1 s_\lambda(G^{-1}(u))dB_n(u) \text{ from (78)}$$

to derive the correlation function

$$\begin{aligned} \mathbb{E}[H_n(\lambda_1)H_n(\lambda_2)] &= \int_0^1 s_{\lambda_1}(G^{-1}(u))s_{\lambda_2}(G^{-1}(u))du \\ &= \int_{-\infty}^{\infty} s_{\lambda_1}(y)s_{\lambda_2}(y)dG(y) = \mathbb{E}[S_n(\lambda_1)S_n(\lambda_2)] \\ &= \frac{\int_{-\infty}^{\infty} (l_{\lambda_1} - 1)(l_{\lambda_2} - 1)dG(y)}{\sqrt{\|l_{\lambda_1}\|^2 - 1}\sqrt{\|l_{\lambda_2}\|^2 - 1}} \\ &= \frac{\sqrt{(2\lambda_1 - 1)(2\lambda_2 - 1)}}{\lambda_1 + \lambda_2 - 1} \frac{[1 - e^{-T(\lambda_1 + \lambda_2 - 1)}]}{\sqrt{(1 - e^{-T(2\lambda_1 - 1)})(1 - e^{-T(2\lambda_2 - 1)})}}. \end{aligned}$$

The correlation function of the re-scaled process $\{H_n(e^s + \frac{1}{2}), -\log 2 \leq s < \infty\}$ becomes

$$\begin{aligned} \rho(t, t + \Delta) &= \mathbb{E}\left[H_n\left(e^t + \frac{1}{2}\right)H_n\left(e^{t+\Delta} + \frac{1}{2}\right)\right] \\ &= \frac{2}{e^{\frac{\Delta}{2}} + e^{-\frac{\Delta}{2}}} \frac{1 - e^{-T(e^t + e^{t+\Delta})}}{\sqrt{(1 - e^{-2Te^t})(1 - e^{-2Te^{t+\Delta}})}}. \end{aligned} \quad (98)$$

This is of product form as in Corollary 5.8 above. When $\Delta \rightarrow 0$ the first factor can be written as

$$\frac{2}{e^{\frac{\Delta}{2}} + e^{-\frac{\Delta}{2}}} = 1 - \frac{\Delta^2}{8} + o(\Delta^2),$$

where the remainder does not depend on t and thus trivially satisfies (94), (95)

and (96) with $V(t) \equiv \frac{1}{4}$.

The second factor is of the form (93) with

$$D(s, t) = 1 - e^{-T(e^s + e^t)}.$$

Note firstly that for all $s, t \geq k$,

$$1 - e^{2e^k T} \leq D(s, t) \leq 1$$

and thus assumption 3 of Lemma 5.7 is satisfied. The first derivative is

$$\begin{aligned} D_{01}(s, t) &= T e^t e^{-T(e^s + e^t)} \\ &= (T e^t) e^{-T e^t} e^{-T e^s} \\ &\leq (T e^t) e^{-(T e^t)} \\ &\leq \sup_{x \geq 0} x e^{-x} \\ &= e^{-1}. \end{aligned}$$

The second order partial derivatives satisfy

$$\begin{aligned} 0 \geq D_{11}(s, t) &= -T^2 e^{s+t} e^{-T(e^s + e^t)} \\ &= -[(T e^s) e^{-(T e^s)}] [(T e^t) e^{-(T e^t)}] \\ &\geq -e^{-2} \\ D_{02}(s, t) &= (T e^t - T^2 e^{2t}) e^{-T(e^s + e^t)} \\ D_{12}(s, t) &= -T e^s (T e^t - T^2 e^{2t}) e^{-T(e^s + e^t)} \\ D_{22}(s, t) &= (T e^t - T^2 e^{2t}) (T^2 e^{2s} - T e^s) e^{-T(e^s + e^t)}. \end{aligned}$$

Finally, the third order partial derivatives satisfy

$$\begin{aligned}
D_{03}(s, t) &= (T^3 e^{3t} - T^2 e^{2t} + T e^t) e^{-T(e^s + e^t)} \\
D_{13}(s, t) &= -T e^s (T^3 e^{3t} - T^2 e^{2t} + T e^t) e^{-T(e^s + e^t)} \\
D_{23}(s, t) &= (T^2 e^{2s} - T e^s) (T^3 e^{3t} - T^2 e^{2t} + T e^t) e^{-T(e^s + e^t)} \\
D_{33}(s, t) &= -(T^3 e^{3t} - T^2 e^{2t} + T e^t) (T^3 e^{3t} - T^2 e^{2t} + T e^t) e^{-T(e^s + e^t)}.
\end{aligned}$$

Each of these expressions is also uniformly bounded (since $x^j e^{-x}$ is uniformly bounded for each $0 \leq j \leq 3$). The conditions 1 and 2 of Lemma 5.7 are therefore both satisfied. Thus for this second factor, (94), (95) and (96) hold with

$$\begin{aligned}
V(t) &= \frac{D_{11}(t, t)}{D(t, t)} - \left[\frac{D_{01}(t, t)}{D(t, t)} \right]^2 \\
&= \frac{T^2 e^{2t} e^{-2Te^t}}{1 - e^{-2Te^t}} - \left[\frac{T e^t e^{-Te^t}}{1 - e^{-2Te^t}} \right]^2 \\
&= -\frac{T^2 e^{2t} e^{-2Te^t}}{1 - e^{-2Te^t}} \left(1 - \frac{e^{-2Te^t}}{1 - e^{-2Te^t}} \right) \\
&= -\left(T e^t e^{-Te^t} \right)^2 \frac{1 - 2e^{-2Te^t}}{(1 - e^{-2Te^t})^2} \\
&= -\left(T e^t e^{-Te^t} \right)^2 \frac{1 - 2e^{-2Te^t}}{1 - 2e^{-2Te^t} + e^{-4Te^t}}.
\end{aligned}$$

Thus the product $\rho(t, t + \Delta)$ also satisfies (94), (95) and (96) with

$$\begin{aligned}
V(t) &= \frac{1}{4} - \left(T e^t e^{-Te^t} \right)^2 \frac{1 - 2e^{-2Te^t}}{(1 - e^{-2Te^t})^2} \\
&= \frac{1}{4} - \left(T e^t e^{-Te^t} \right)^2 \frac{1 - 2e^{-2Te^t}}{1 - 2e^{-2Te^t} + e^{-4Te^t}}. \tag{99}
\end{aligned}$$

□

5.3.1 Proof of Lemma 5.1

Hüsler (1990) builds on the work of Leadbetter et al. (1983) and Berman (1985) to consider large values of locally stationary Gaussian processes, which satisfy Berman's condition of long range dependence. We are able to utilise the results of Hüsler (1990) if we are able to verify that a Gaussian process having the correlation function $\rho(s, t)$ from (98) is locally stationary. To do this, in addition to (94), (95) and (96), we need to verify

$$0 < \inf_{t \geq k} V(t) \leq \sup_{t \geq k} V(t) < \infty. \quad (100)$$

There are two cases to deal with here. Firstly, if $e^{-2Te^t} < \frac{1}{2}$ then according to the form (99) we can see that the first factor in the second term satisfies

$$0 \leq \left(Te^t e^{-Te^t} \right)^2 \leq e^{-2}$$

while the second factor of the second term in (99) satisfies

$$0 \leq \frac{1 - 2e^{-2Te^t}}{1 - 2e^{-2Te^t} + e^{-4Te^t}} \leq 1$$

and so for such t, T ,

$$\frac{1}{4} \geq V(t) \geq \frac{1}{4} - e^{-2} \approx 0.114.$$

Now, if $\frac{1}{2} \leq e^{-2Te^t} < 1$ then we have

$$-1 \leq 1 - 2e^{-2Te^t} \leq 0$$

and since for $0 \leq x < 1$,

$$\frac{x}{2} \leq x - \frac{x^2}{2} \leq 1 - e^{-x} \leq x \leq 1$$

we also have

$$(Te^t)^2 \leq (1 - e^{-2Te^t})^2 \leq 1$$

and so for such t, T ,

$$\begin{aligned} \frac{1}{4} \leq V(t) &= \frac{1}{4} - \frac{(Te^t e^{-Te^t})^2}{(1 - e^{-2Te^t})^2} (1 - 2e^{-2Te^t}) \\ &\leq \frac{1}{4} + \frac{(Te^t e^{-Te^t})^2}{(Te^t)^2} \\ &\leq \frac{1}{4} + e^{-2Te^t} \\ &\leq \frac{5}{4}. \end{aligned}$$

We have thus established that for all $t \geq -\log 2$,

$$\frac{1}{4} - e^{-2} \approx 0.114 \leq V(t) \leq \frac{5}{4}.$$

Thus condition (100) holds. This implies that our Gaussian process $\{H_n(e^s + \frac{1}{2})\}$ with correlation function $\rho(s, t)$ given in (98) is locally stationary.

In order to utilise the results of Theorem 4.2 of Hüsler (1990), we must verify the long range condition

$$\sup_{t, t+\Delta \geq k} \rho(t, t + \Delta) = o\left(\frac{1}{\log(\Delta)}\right) \text{ as } \Delta \rightarrow \infty, \quad (101)$$

so that conditions (5), (6) and (10) of Hüsler (1990) are satisfied. This is straight-

forward, since the first factor in (98) satisfies

$$\frac{2}{e^{\frac{\Delta}{2}} + e^{-\frac{\Delta}{2}}} = 2e^{-\frac{\Delta}{2}} [1 + o(1)]$$

as $\Delta \rightarrow \infty$ which decays much faster than the rate in condition (101) above (i.e. $e^{-\frac{\Delta}{2}} \log \Delta \rightarrow 0$ as $\Delta \rightarrow \infty$). Thus, by Lemma 5.9 and by Theorem 4.2 of Hüsler (1990) we complete the proof of Lemma 5.1.

5.3.2 Proof of Lemma 5.2

Liu et al. (2003, equation (10), page 234) show that on a suitable probability space there exist versions of $\{\alpha_n(u) : 0 \leq u \leq 1\}$ and $\{B_n(u) : 0 \leq u \leq 1\}$ satisfying

$$\begin{aligned} \sup_{U_{1,n} \leq u \leq U_{n,n}} n^{\frac{1}{4}} \left| \frac{\alpha_n(u) - B_n(u)}{[u(1-u)]^{\frac{1}{4}}} \right| &= \mathcal{O}_p(1) \\ \implies |\alpha_n(u) - B_n(u)| &= \mathcal{O}_p(1) [u(1-u)]^{\frac{1}{4}} n^{-\frac{1}{4}}. \end{aligned} \quad (102)$$

Case (1) When $\lambda \in [\log n, n(\log n)^{-4}]$.

Letting $v(\lambda) = \sqrt{\frac{(\lambda-1)^2}{2\lambda-1} (1 - e^{-T(2\lambda-1)})}$, the remainder term $R_n(\lambda)$ from (77) can be written as

$$\begin{aligned} R_n(\lambda) &= \int_0^{1-e^{-T}} \frac{\lambda(1-u)^{\lambda-1}}{v(\lambda)} d[\alpha_n(u) - B_n(u)] + \int_{1-e^{-T}}^1 \frac{e^{-T(\lambda-1)}}{v(\lambda)} d[\alpha_n(u) - B_n(u)] \\ &= [\alpha_n(1 - e^{-T}) - B_n(1 - e^{-T})] \left[\frac{\lambda e^{-T(\lambda-1)} - e^{-T(\lambda-1)}}{v(\lambda)} \right] \\ &\quad + \int_0^{1-e^{-T}} [\alpha_n(u) - B_n(u)] d \left(\frac{-\lambda(1-u)^{\lambda-1}}{v(\lambda)} \right) \\ &= [\alpha_n(1 - e^{-T}) - B_n(1 - e^{-T})] \left[\frac{\lambda e^{-T(\lambda-1)} - e^{-T(\lambda-1)}}{v(\lambda)} \right] \\ &\quad + \mathcal{O}_p(1) n^{-\frac{1}{2}} \log n \left[\frac{\lambda(1 - e^{-T(\lambda-1)})}{v(\lambda)} \right] \quad \text{using (102)} \\ &= \mathcal{O}_p(1) n^{-\frac{1}{2}} \log n \lambda^{\frac{1}{2}} = \mathcal{O}_p(1) (\log n)^{-1}. \end{aligned} \quad (103)$$

Case (2) When $\lambda \in [1, \log n]$.

Similarly to Case (1), we have

$$\begin{aligned} |R_n(\lambda)| &= |S_n(\lambda) - H_n(\lambda)| \\ &= \mathcal{O}_p(1) n^{-\frac{1}{2}} \log n \lambda^{\frac{1}{2}} = \mathcal{O}_p(1) n^{-\frac{1}{2}} (\log n)^{\frac{3}{2}}. \end{aligned}$$

Lemma 5.1 with $C = \log_{(2)} n$ yields

$$\sup_{\lambda \in [1, \log n]} H_n(\lambda) = \mathcal{O}_p(1) (\log_{(3)} n)^{\frac{1}{2}}.$$

Thus, $\sup_{\lambda \in [1, \log n]} S_n(\lambda) \vee 0 = \mathcal{O}_p(1) (\log_{(3)} n)^{\frac{1}{2}}$.

Case (3) When $\lambda \in [n(\log n)^{-4}, n]$.

Lemma 5.1, with $C = \log_{(2)} n$ yields

$$\sup_{\lambda \in [n(\log n)^{-4}, n]} H_n(\lambda) = \mathcal{O}_p(1) (\log_{(3)} n)^{\frac{1}{2}}.$$

From (103), we can write

$$R_n(\lambda) = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4,$$

where

$$\begin{aligned}\Delta_1 &= [\alpha_n(1 - e^{-T}) - B_n(1 - e^{-T})] \left[\frac{(\lambda - 1)e^{-T(\lambda-1)}}{v(\lambda)} \right], \\ \Delta_2 &= \int_{U_{1,n}}^{1-e^{-T}} [\alpha_n(u) - B_n(u)] d \left(\frac{-\lambda(1-u)^{-(\lambda-1)}}{v(\lambda)} \right), \\ \Delta_3 &= \int_0^{U_{1,n}} \sqrt{n} [F_n(u) - u] d \left(\frac{-\lambda(1-u)^{-(\lambda-1)}}{v(\lambda)} \right) \text{ and} \\ \Delta_4 &= \int_0^{U_{1,n}} B_n(u) d \left(\frac{\lambda(1-u)^{-(\lambda-1)}}{v(\lambda)} \right).\end{aligned}$$

From (102),

$$\Delta_1 = \mathcal{O}_p(1) n^{-\frac{1}{4}} \left[\frac{(\lambda - 1)e^{-T(\lambda-1)}}{v(\lambda)} \right] = o_p(1)$$

due to the fact that $\lambda \geq n(\log n)^{-4} \rightarrow \infty$. Note that (102) implies $[\alpha_n(u) - B_n(u)] = \mathcal{O}_p(1) n^{-\frac{1}{4}} (u(1-u))^{\frac{1}{4}}$ and also that $u = G(y) = \mathcal{O}(1) y$. We can show the asymptotic order of Δ_2, Δ_3 , and Δ_4 using the same reasoning as Liu et al. (2003), so that we have

$$\begin{aligned}\Delta_2 &= \mathcal{O}_p(1) n^{-\frac{1}{4}} \int_{Y_{1,n}}^T \frac{y^{\frac{1}{4}} \lambda (\lambda - 1) e^{-y(\lambda-1)}}{v(\lambda)} dy \\ &= \mathcal{O}_p(1) n^{-\frac{1}{4}} \lambda^{\frac{1}{4}} = \mathcal{O}_p(1).\end{aligned}$$

When $0 \leq u \leq U_{1,n}$, we have $F_n(u) = 0$ and $\left(\frac{\lambda(\lambda-1)e^{-y(\lambda-1)}}{v(\lambda)} \right)$ is a decreasing function of y . Thus, $\Delta_3 \leq 0$.

Finally, when $0 \leq y \leq Y_{1,n}$, we have $B_n(G(y)) = \mathcal{O}_p(1) G^{\frac{1}{2}}(Y_{1,n}) = \mathcal{O}_p(1) Y_{1,n}^{\frac{1}{2}}$.

We note that $Y_{1,n} = \mathcal{O}_p(1) \frac{1}{n}$. Then we have

$$\begin{aligned}
|\Delta_4| &= \int_0^{U_{1,n}} B_n(u) d\left(\frac{\lambda(\lambda-1)e^{-y(\lambda-1)}}{v(\lambda)}\right) \\
&= \mathcal{O}_p(1) \int_0^{Y_{1,n}} \frac{Y_{1,n}^{\frac{1}{2}} \lambda(\lambda-1)e^{-y(\lambda-1)}}{v(\lambda)} dy \\
&= \mathcal{O}_p(1) Y_{1,n}^{\frac{1}{2}} \lambda^{\frac{1}{2}} = \mathcal{O}_p(1) n^{\frac{1}{2}} Y_{1,n}^{\frac{1}{2}} = \mathcal{O}_p(1).
\end{aligned}$$

Therefore,

$$\sup_{\lambda \in [n(\log n)^{-4}, n]} S_n(\lambda) \vee 0 = \mathcal{O}_p(1) (\log_{(3)} n)^{\frac{1}{2}}.$$

Case (4) When $\lambda \geq n$.

Note firstly that $G(y) = \mathcal{O}(1)y$, $F_n(y) = \mathcal{O}_p(1)y$ and that $v(\lambda) = \mathcal{O}(1)\lambda^{\frac{1}{2}}$ for large λ . Then from (76) we have

$$\begin{aligned}
S_n(\lambda) &= \int_0^1 s_\lambda(G^{-1}(u)) d\alpha_n(u) \\
&= n^{\frac{1}{2}} \int_0^{1-e^{-T}} \frac{\lambda(1-u)^{(\lambda-1)}}{v(\lambda)} d[F_n(u) - u] + n^{\frac{1}{2}} \int_{1-e^{-T}}^1 \frac{e^{-T(\lambda-1)}}{v(\lambda)} d[F_n(u) - u] \\
&= n^{\frac{1}{2}} \int_0^{1-e^{-T}} \frac{(F_n(u) - u) \lambda(\lambda-1)(1-u)^{\lambda-2}}{v(\lambda)} du \\
&\quad + n^{\frac{1}{2}} [F_n(1-e^{-T}) - (1-e^{-T})] \left(\frac{(\lambda-1)e^{-T(\lambda-1)}}{v(\lambda)}\right) \\
&= n^{\frac{1}{2}} \frac{1}{v(\lambda)} \lim_{t \uparrow T, t < T} \int_0^t \frac{(F_n(G(y)) - G(y))}{y} y \lambda(\lambda-1) e^{-y(\lambda-1)} dy + o_p(1) \\
&= \mathcal{O}_p(1) n^{\frac{1}{2}} \lambda^{-\frac{1}{2}} \lim_{t \uparrow T, t < T} \int_0^t \lambda^2 y e^{-y\lambda} dy \\
&= \mathcal{O}_p(1).
\end{aligned}$$

This completes the proof of Lemma 5.2.

5.3.3 Proof of Theorem 5.3

Lemma 5.2 yields that

$$\begin{aligned} M_n &= \sup_{\lambda > 1} S_n(\lambda) = \sup_{\lambda \in [\log n, n(\log n)^{-4}]} S_n(\lambda) \\ &= \sup_{\lambda \in [\log n, n(\log n)^{-4}]} H_n(\lambda) + \mathcal{O}_p(1) (\log n)^{-1}. \end{aligned}$$

If we write $\lambda = e^s + \frac{1}{2}$, then we have $s = \log(\lambda - \frac{1}{2})$. Applying this transformation to the upper bound of s in Lemma 5.1 yields

$$\begin{aligned} \lambda &\leq n (\log n)^{-4} \\ \implies s &\leq \log \left(n (\log n)^{-4} - \frac{1}{2} \right) \\ &= \log \left\{ n (\log n)^{-4} \left[1 - \frac{1}{2n (\log n)^{-4}} \right] \right\} \\ &= \log n - 4 \log \log n + \log \left[1 - \frac{1}{2n (\log n)^{-4}} \right] \\ &= \log n - 4 \log \log n + \mathcal{O}_p(1) \frac{1}{n (\log n)^{-4}} \\ &= C_n. \end{aligned}$$

As in Lemma 5.1, we define

$$\begin{aligned}
A_{C_n} &= \sqrt{2 \log C_n} \\
&= \sqrt{2 \log \left[\log n - 4 \log \log n + \mathcal{O}_p(1) \frac{1}{n (\log n)^{-4}} \right]} \\
&= \sqrt{2 \log \left\{ \log n \left[1 - \frac{4 \log \log n}{\log n} + \mathcal{O}_p(1) \frac{(\log n)^3}{n} \right] \right\}} \\
&= \sqrt{2 \left\{ \log \log n + \log \left[1 - \frac{4 \log \log n}{\log n} + \mathcal{O}_p(1) \frac{(\log n)^3}{n} \right] \right\}} \\
&= \sqrt{2 \log \log n + \mathcal{O}_p(1) \frac{\log \log n}{\log n}} \\
&= \sqrt{2 \log \log n} \left[1 + \mathcal{O}_p(1) \frac{1}{\log n} \right]^{\frac{1}{2}} \\
&= \sqrt{2 \log \log n} \left[1 + \mathcal{O}_p(1) \frac{1}{\log n} \right]. \tag{104}
\end{aligned}$$

Lemma 5.1 yields

$$A_{C_n} \left[M_n - \mathcal{O}_p(1) (\log n)^{-1} - A_{C_n} \right] + \log(4\pi) = G_n + o_p(1),$$

where G_n converges to a Gumbel distribution as $n \rightarrow \infty$. From (104) we have $A_{C_n} \rightarrow \infty$ and letting $K = \log(4\pi)$, we can write

$$\begin{aligned}
M_n &= A_{C_n} + \frac{G_n - K + o_p(1)}{A_{C_n}} + o_p(1) \quad \text{so that} \\
M_n^2 &= A_{C_n}^2 + 2[G_n - K] + o_p(1), \quad \text{which implies} \\
\frac{M_n^2 - A_{C_n}^2}{2} + K &= G_n + o_p(1).
\end{aligned}$$

Therefore, we have

$$\lim_{n \rightarrow \infty} P \left\{ \frac{M_n^2 - A_{C_n}^2}{2} + K \leq w \right\} = e^{-e^{-w}}$$

and substituting $w = \frac{y}{2}$ yields

$$\lim_{n \rightarrow \infty} P \{ M_n^2 - A_{C_n}^2 + 2K \leq y \} = e^{-e^{-\frac{y}{2}}} \quad (105)$$

as required.

5.3.4 Proof of Lemma 5.4

If $p\lambda \leq 2$, then we would have $p\lambda = \mathcal{O}(1)$. Thus, we can assume here that $p\lambda > 2$.

Similar to Liu et al. (2003) we define $y_0(p, \lambda) = \log(p\lambda)/(\lambda - 1)$, which we will simply denote by y_0 .

Case 1 ($y_0 \geq T$).

If $y_0 \geq T$ then

$$\log(p\lambda)/(\lambda - 1) \geq T \implies \lambda \geq p\lambda \geq e^{T(\lambda-1)} \implies p\lambda = \mathcal{O}(1). \quad (106)$$

Case 2 ($y_0 < T$).

We note that $G(y)$ and $F_n(G(y))$ are identical to the functions for the gamma distribution with $\kappa = 1$ from Liu et al. (2003) for $y < T$. Thus we have

$$\begin{aligned} n^{-1}L_n(p, \lambda) &= \left(\int_{y_0}^T + \int_0^{y_0} \right) \log(1 + p(l(y, \lambda) - 1)) dF_n(G(y)) + \frac{1}{n} \sum_{i=1}^n e^{-T(\lambda-1)} \mathbb{I}\{Y_i = T\} \\ &\leq \int_{y_0}^T (e^{-(y-y_0)(\lambda-1)} - p) dF_n(G(y)) + [1 - F_n(G(T))] e^{-T(\lambda-1)} \\ &\quad + \int_0^{y_0} [\log 2 + (y_0 - y)(\lambda - 1)] dF_n(G(y)). \end{aligned} \quad (107)$$

Note that $F_n(G(y)) = \mathcal{O}_p(1) G(y)$ and also that $e^{-T(\lambda-1)} = O(1)$, otherwise $\lambda = \mathcal{O}(1)$ and hence $p\lambda = \mathcal{O}(1)$.

Now, the first part of (107) yields

$$\begin{aligned}
& \int_{y_0}^T (e^{-(y-y_0)(\lambda-1)} - p) dF_n(G(y)) + [1 - F_n(G(T))] e^{-T(\lambda-1)} \\
= & [(e^{-(T-y_0)(\lambda-1)} - p) F_n(G(T)) - (1-p) F_n(G(y_0))] \\
& + \int_{y_0}^T F_n(G(y)) d(-e^{-(y-y_0)(\lambda-1)}) + [1 - F_n(G(T))] e^{-T(\lambda-1)} \\
\leq & -p F_n(G(T)) + (e^{-(T-y_0)(\lambda-1)} - e^{-T(\lambda-1)}) F_n(G(T)) \\
& + \mathcal{O}_p(1) \int_{y_0}^T G(y) d(-e^{-(y-y_0)(\lambda-1)}) + e^{-T(\lambda-1)} \\
= & -p F_n(G(T)) + (e^{-(T-y_0)(\lambda-1)} - e^{-T(\lambda-1)}) F_n(G(T)) \\
& + \mathcal{O}_p(1) \left(-G(T) e^{-(T-y_0)(\lambda-1)} + G(y_0) + \int_{y_0}^T e^{-(y-y_0)(\lambda-1)} dG(y) + e^{-T(\lambda-1)} \right) \\
\leq & -p F_n(G(T)) + \mathcal{O}_p(1) ((F_n(G(T)) - G(T)) e^{-(T-y_0)(\lambda-1)} + e^{-T(\lambda-1)} (1 - F_n(G(T)))) \\
& + \mathcal{O}_p(1) (G(y_0) + p(1 - G_\lambda(y_0))).
\end{aligned}$$

Note that Equation 12 from Liu et al. (2003) yields

$$\sup_{U_{1,n} \leq u \leq U_{n,n}} |\alpha_n(u)| = \mathcal{O}_p(1) (\log_{(2)} n)^{\frac{1}{2}} (u(1-u))^{\frac{1}{2}}, \quad (108)$$

which is applied such that

$$\begin{aligned}
\mathcal{O}_p(1) ((F_n(G(T)) - G(T)) e^{-(T-y_0)(\lambda-1)}) &= \mathcal{O}_p(1) \left(\frac{\log_{(2)} n}{n} \right)^{\frac{1}{2}} (e^{-T}(1 - e^{-T}))^{\frac{1}{2}} \\
&= o_p(1) \text{ for } p\lambda \rightarrow \infty.
\end{aligned}$$

The second part of (107) can be bounded by $\mathcal{O}_p(1) G(y_0) \log(p\lambda)$, thus, follow-

ing in the same way as Liu et al. (2003),

$$n^{-1}L_n(p, \lambda) \leq p \{-F_n(G(T)) + \mathcal{O}_p(1) [h(p, \lambda)]\} \quad (109)$$

where the right hand side of (109) is decreasing when $p\lambda$ is large and $h(p, \lambda) \rightarrow 0$ as $p\lambda \rightarrow \infty$. Therefore, when $L_n(p, \lambda) > 0$, we have $p\lambda = \mathcal{O}(1)$.

This completes the proof of Lemma 5.4.

5.3.5 Proof of Lemma 5.5

We define

$$\begin{aligned} P_n s_\lambda^2 &= \frac{1}{n} \sum_{i=1}^n s_\lambda^2(Y_i) \\ &= \int_0^1 \frac{[\lambda^{\mathbb{I}\{G^{-1}(u) < T\}} e^{-G^{-1}(u)(\lambda-1)} - 1]^2}{v^2(\lambda)} du \\ &\quad + \int_0^1 \frac{[\lambda^{\mathbb{I}\{G^{-1}(u) < T\}} e^{-G^{-1}(u)(\lambda-1)} - 1]^2}{v^2(\lambda)} d(F_n(u) - u) \\ &= 1 + \int_0^{1-e^{-T}} \frac{(\lambda(1-u)^{(\lambda-1)} - 1)^2}{v^2(\lambda)} d(F_n(u) - u) \\ &\quad + \int_{1-e^{-T}}^1 \frac{(e^{-T(\lambda-1)} - 1)^2}{v^2(\lambda)} d(F_n(u) - u) \\ &= 1 + \frac{(e^{-T(\lambda-1)} - 1)^2}{v^2(\lambda)} [1 - e^{-T} - F_n(1 - e^{-T})] \\ &\quad + \frac{F_n(1 - e^{-T}) - (1 - e^{-T})}{v^2(\lambda)} \{\lambda^2 e^{-2T(\lambda-1)} - 2\lambda e^{-T(\lambda-1)} + 1\} \\ &\quad - \int_0^{1-e^{-T}} [F_n(u) - u] \frac{d}{du} \left[\frac{(\lambda(1-u)^{\lambda-1} - 1)^2}{v^2(\lambda)} \right] du \\ &= 1 + \Delta_5 + \Delta_6 + \Delta_7, \end{aligned}$$

where

$$\begin{aligned}\Delta_5 &= \int_0^{Y_{1,n}} (F_n(G(y)) - G(y)) d(-s_\lambda^2(y)), \\ \Delta_6 &= \lim_{t \uparrow T, t < T} \int_{Y_{1,n}}^t (F_n(G(y)) - G(y)) d(-s_\lambda^2(y)) \text{ and} \\ \Delta_7 &= \frac{[F_n(1 - e^{-T}) - (1 - e^{-T})]}{v^2(\lambda)} \{(\lambda^2 - 1)e^{-2T(\lambda-1)} - (2\lambda - 2)e^{-T(\lambda-1)}\}.\end{aligned}$$

Case (1) When $\lambda \in [\lambda^*, n(\log n)^{-4}]$.

Since $F_n(G(y))$ and $G(y)$ are identical to the case in Liu et al. (2003) with $\kappa = 1$ for $y < T$, using the same steps as their's (although ignoring the second term in their Δ_5), we have

$$\begin{aligned}|\Delta_5| &= \mathcal{O}_p((\log n)^{-1}) \text{ and} \\ |\Delta_6| &= \mathcal{O}_p((\log n)^{-1}).\end{aligned}$$

We can also see that

$$|\Delta_7| \leq \frac{[F_n(1 - e^{-T}) - (1 - e^{-T})]}{4v^2(\lambda)}$$

since $\lambda^k e^{-2T(\lambda-1)}$ is uniformly bounded in λ for $k = 1, 2$. When $\lambda \in [\log n, n(\log n)^{-4}]$, we have $v^2(\lambda) = \mathcal{O}(1)$ so that from (108) we can see that $|\Delta_7| = o_p(1)$.

Case (2) When $\lambda \in [n(\log n)^{-4}, \frac{1}{Y_{1,n}}]$.

We define

$$\mathcal{X}^2(x, \lambda) = (2\lambda - 1)(e^{-x(\lambda-1)} - \lambda^{-1})^2 \tag{110}$$

to be the same as the square of the centered, standardised score function for the

case of $\kappa = 1$ in Liu et al. (2003). Note that Lemma 4 of Liu et al. (2003) yields

$$\frac{1}{\frac{1}{n} \sum_{i=1}^n \mathcal{X}^2(x_i, \lambda)} = \mathcal{O}_p(1) \quad (111)$$

Note that,

$$\begin{aligned} \frac{(1 - e^{-(\lambda-1)T})^2}{1 - e^{-(2\lambda-1)T}} &= 1 + \frac{e^{-(2\lambda-1)T} - 2e^{-(\lambda-1)T} + e^{-(2\lambda-2)T}}{1 - e^{-(2\lambda-1)T}} \\ &= 1 + o(1) \text{ as } n \rightarrow \infty. \end{aligned} \quad (112)$$

Then using (112) we can show directly that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n s_{\lambda}^2(Y_i) &= \frac{1}{n} \sum_{i=1}^n \mathcal{X}^2(x_i, \lambda) \left[\frac{(1 - e^{-(\lambda-1)T})^2}{(1 - \lambda e^{-(\lambda-1)x_i})^2} \right]^{\mathbb{I}\{x_i \geq T\}} \frac{1}{1 - e^{-T(2\lambda-1)}} \\ &\geq \frac{1}{n} \sum_{i=1}^n \mathcal{X}^2(x_i, \lambda) \left[\frac{(1 - e^{-(\lambda-1)T})^2}{1 - e^{-(2\lambda-1)T}} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathcal{X}^2(x_i, \lambda) [1 + o(1)] \end{aligned} \quad (113)$$

Then (111) yields that

$$\frac{1}{\frac{1}{n} \sum_{i=1}^n s_{\lambda}^2(Y_i)} = \mathcal{O}_p(1). \quad (114)$$

This completes the proof of Lemma 5.5.

5.3.6 Proof of Theorem 5.6

The proof of Theorem 5.6 is identical, *mutatis mutandis*, to the proof of Theorem 2 in Liu et al. (2003).

6 Conclusion

In this work, we have explored the theory and applications of various multi-regime models involving Markov chains. We have addressed a series of problems involving non-homogeneous data, where Markov chains are used in the parameter estimation procedure as well as in the model itself. We saw in Chapter 2 how a multi-regime model can be applied to a single discrete time series. Specifically, we applied a Poisson change-point model to the history of quarterly driver fatality counts in the state of Victoria, Australia. This approach is the first of its kind on this data and provides some useful insights into when change-points occurred in the data and what the magnitude of the changes were. We gained a deeper understanding of the properties of the Markov chain used in the estimation procedure. That is, we proved that the Gibbs sampler for the Poisson change-point model is geometrically ergodic. This result is of great importance to practitioners using Poisson change-point models in a Bayesian framework for many different types of data. Thus, given a specific convergence level for the distribution, the minimum number of iterations required can be calculated. Although we have identified a key quality of the convergence rate of the sampler, the calculation of the specific rate of convergence is left for further research. It would also be of interest to see if the bounding technique of Section 2.4 can be used to prove geometric ergodicity of MCMC algorithms for other models.

The Gibbs sampler is again used in Chapter 3. Here, we are applying a double chain Markov model to multiple discrete time series of differing lengths. Convergence of the Gibbs sampler is improved by adding an additional step, where the hidden data labels are randomly permuted. The nature of the data provides a challenge when specifying the exact steps for parameter estimation. We derive these steps and apply the model to credit rating migration data that are driven by Markov chains that are selected from a Markov chain of hidden regimes. When

we overlay the regimes selected by the model on the data with historical economic data, we are able to see that a remarkable pattern emerges, where credit migration dynamics switch in positive and negative market conditions. We also show that our model is more effective than other existing double chain Markov models, using a simulation study. It would be of interest to further the theoretical work on this problem so that we can develop more specific results about the convergence rate of the modified Haar PX-DA algorithm and to compare it to the convergence rate of the regular DA without the extra permutation step.

A similar dataset on credit rating migrations, albeit with multiple observations over continuous time, is involved in Chapter 4. The multiple regimes are across the population rather than over time, thus we are studying a Markov chain mixture model for the data. We address the problem of testing for the number of mixture components using the log-likelihood ratio. We adapt the results of Fukumizu (2003) that show the divergence of the log-likelihood ratio under certain conditions, to show that the log-likelihood ratio for our model also diverges to infinity. This is contrary to the claims of Frydman (2005) and we provide evidence for our claim through a parametric bootstrap procedure. We then look at a simplified version of the mixture problem, where each Markov chain mixture component has only 2 states, one of which is the absorbing *default* state, which is equivalent to a mixture of censored exponentials problem.

In Chapter 5, we analyse this particular problem and derive the exact limiting distribution of the log-likelihood ratio test statistic, so that we are able to test for the presence of a mixture. Each of the problems we explore enable us to gain a greater understanding of the nature of multi-regime models that involve Markov chains in the parameter estimation procedures or in the models themselves. These significant insights are gained through the application of these models to address practical problems that do not have a clear solution.

It would be of interest to develop a general theorem, similar to Fukumizu (2003), which applies to a greater number of Markov chain components than the 1 vs. 2 component test we explored. The results of Chapter 5 could also be extended to tests for the number of mixture components of N -state Markov chains with $N > 2$, to derive the limiting distribution of the log-likelihood ratio test statistic. This would enable the theory to address all of the problems involving tests for the number of mixture components in the class of models discussed in Frydman (2005). We leave these problems to be explored with further research.

A Covariance calculation for stochastic integrals

The left hand side of (75) on page 109 can be written as follows

$$\begin{aligned}
& \mathbb{E} \left\{ \left[W_n(1)g(1) - \int_0^1 W_n(u)dg(u) \right] \left[W_n(1)h(1) - \int_0^1 W_n(u)dh(u) \right] \right\} \\
&= \mathbb{E} \left\{ g(1)h(1) - g(1) \int_0^1 W_n(1)W_n(u)dh(u) \right. \\
&\quad \left. - h(1) \int_0^1 W_n(1)W_n(u)dg(u) + \int_0^1 \int_0^1 W_n(u)W_n(v)dg(u)dh(v) \right\}. \\
&= g(1)h(1) - g(1) \int_0^1 u dh(u) - h(1) \int_0^1 u dg(u) \\
&\quad + \int_0^1 \int_0^1 (u \wedge v) dg(u)dh(v). \tag{115}
\end{aligned}$$

The fact that we may take expectations inside these integrals follows e.g. from the representation of the Wiener process as a series $W(u) = \sum_j A_j(u)Z_j$ for independent standard normal random variables $\{Z_j\}$ and a countable class of continuous functions $\{A_j(\cdot)\}$ satisfying $\sum_j A_j(u)A_j(v) = u \wedge v$ for all $0 \leq u, v \leq 1$ for our counting index $j = 1, 2, \dots$ and with $u \wedge v$ denoting the minimum of u and v (see McKean (1969) for further details).

The last integral at (115) in turn can be written as

$$\begin{aligned}
& \int_0^1 \int_0^1 (u \wedge v) dg(u) dh(v) \\
&= \int_{u=0}^1 \left(\left[\int_{v=0}^u v + \int_{v=u}^1 u \right] dh(v) \right) dg(u) \\
&= \int_0^1 uh(1)dg(u) - \int_{u=0}^1 \int_{v=0}^u h(v)dv dg(u) \\
&= h(1) \left[g(1) - \int_0^1 g(u)du \right] - \int_{v=0}^1 \int_{u=v}^1 dg(u)h(v)dv \\
&= g(1)h(1) - h(1) \int_0^1 g(u)du - \int_{v=0}^1 [g(1) - g(v)] h(v)dv \\
&= \int_0^1 g(u)h(u)du - g(1)h(1) + g(1) \int_0^1 udh(u) + h(1) \int_0^1 udg(u).
\end{aligned}$$

Inserting this into (115) gives the result.

References

- ALBERT, A. (1962). Estimating the infinitesimal generator of a continuous time, finite state Markov process. *Annals of Mathematical Statistics* **33** 727–753.
- AZZALINI, A. and BOWMAN, A. W. (1990). A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society: Series C* **39** 357–365.
- BANGIA, A., DIEBOLD, F. X., KRONIMUS, A., SCHAGEN, C. and SCHUERMANN, T. (2002). Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking & Finance* **26** 445–474.
- BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* **37** 1554–1563.
- BERCHTOLD, A. (1999). The double chain Markov model. *Communications in Statistics. Theory and Methods* **28** 2569–2589.
- BERCHTOLD, A. (2002). High-order extensions of the double chain Markov model. *Stochastic Models* **18** 193–227.
- BERMAN, S. M. (1985). An asymptotic formula for the distribution of the maximum of a Gaussian process with stationary increments. *Journal of Applied Probability* **22** 454–460.
- BICKEL, P. and CHERNOFF, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non-regular problem. *Statistics and Probability: A Raghuraj Bahadur Festschrift* .
- BOYS, R. J. and HENDERSON, D. A. (2002). On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientific Programming* **10** 795–809.

- CAPPÉ, O. (2001). Ten years of HMMs (online bibliography 1989 - 2000).
URL <http://www.tsi.enst.fr/~cappe/docs/hmbib.html>
- CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in hidden Markov models*. Springer, New York.
- CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society, Series C* **41** 389–405.
- CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25** 573–578.
- CHIB, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75** 79–97.
- CHURCHILL, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51** 79–94.
- CONWAY, J. B. (1990). *A Course in Functional Analysis*. 2nd ed. Springer, New York.
- DACUNHA-CASTELLE, D. and GASSIAT, É. (1997). Testing in locally conic models, and application to mixture models. *European Series in Applied and Industrial Mathematics. Probability and Statistics* **1** 285–317.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B.* **39** 1–38.
- DOOB, J. L. (1953). *Stochastic processes*. Wiley, New York.

- EISENKOPF, A. (2008). The real nature of credit rating transitions. *Available at SSRN 968311* .
- FITZPATRICK, M. (2014). Geometric ergodicity of the Gibbs sampler for the Poisson change-point model. *Statistics & Probability Letters* **91** 55–61.
- FITZPATRICK, M. and MARCHEV, D. (2013). Efficient Bayesian estimation of the multivariate double chain Markov model. *Statistics and Computing* **23** 467–480.
- FORNEY, G. D., JR. (1973). The Viterbi algorithm. *Proceedings of the IEEE* **61** 268–278.
- FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96** 194–209.
- FRYDMAN, H. (2005). Estimation in the mixture of Markov chains moving with different speeds. *Journal of the American Statistical Association* **100** 1046–1053.
- FUKUMIZU, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics* **31** 833–851.
- GAREL, B. (2005). Asymptotic theory of the likelihood ratio test for the identification of a mixture. *Journal of statistical planning and inference* **131** 271–296.
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 721–741.
- GHOSH, J. K. and SEN, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, California, 1983)*. Wadsworth, Belmont, CA.

- GIAMPIERI, G., DAVIS, M. and CROWDER, M. (2005). Analysis of default data using hidden Markov models. *Quantitative Finance* **5** 27–34.
- HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, California, 1983)*. Wadsworth, Belmont, CA.
- HOBERT, J. P. and MARCHEV, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics* **36** 532–554.
- HOBERT, J. P., ROY, V. and ROBERT, C. P. (2011). Improving the convergence properties of the data augmentation algorithm with an application to Bayesian mixture modeling. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* **26** 332–351.
- HUGHES, J. P., GUTTORP, P. and CHARLES, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society. Series C* **48** 15–30.
- HÜSLER, J. (1990). Extreme values and high boundary crossings of locally stationary Gaussian processes. *Annals of Probability* **18** 1141–1158.
- JARRETT, R. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66** 191–193.
- JARROW, R. A., LANDO, D. and TURNBULL, S. (1997). A Markov model for the term structure of credit risk spreads. *Review of Financial Studies* **10** 481–523.
- KENNY, P., LENNIG, M. and MERMELSTEIN, P. (1990). A linear predictive HMM for vector-valued observations with applications to speech recognition. *Acoustics, Speech & Signal Processing* **38** 220–225.

- KHARE, K. and HOBERT, J. P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. *Annals of Statistics* **39** 2585–2606.
- KIRSHNER, S. (2005). *Modeling of multivariate time series using hidden Markov models*. Ph.D. thesis, University of California, Irvine.
- KOROLKIEWICZ, M. W. and ELLIOTT, R. J. (2008). A hidden Markov model of credit quality. *Journal of Economic Dynamics & Control* **32** 3807 – 3819.
- LANCHANTIN, P., LAPUYADE-LAHORGUE, J. and PIECZYNSKI, W. (2008). Un-supervised segmentation of triplet Markov chains hidden with long-memory noise. *Signal Processing* **88** 1134–1151.
- LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics, Springer-Verlag, New York-Berlin.
- LIU, X., PASARICA, C. and SHAO, Y. (2003). Testing homogeneity in gamma mixture models. *Scandinavian Journal of Statistics. Theory and Applications* **30** 227–239.
- LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics* **31** 807–832.
- LIU, X. and SHAO, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference* **123** 61–81.
- McKEAN, H. P., JR. (1969). *Stochastic integrals*. Probability and Mathematical Statistics, No. 5, Academic Press, New York-London.

- MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- PALIWAL, K. (1993). Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer. *Acoustics, Speech and Signal Processing* **2** 215–218.
- PIECZYNSKI, W. (2007). Multisensor triplet Markov chains and theory of evidence. *International Journal of Approximate Reasoning* **45** 1–16.
- PIECZYNSKI, W. and DESBOUVRIES, F. (2005). On triplet Markov chains. In *International Symposium on Applied Stochastic Models and Data Analysis, (ASMDA 2005), Brest, France*. Citeseer.
- PORITZ, A. B. (1982). Linear predictive hidden Markov models and the speech signal. *Acoustics, Speech & Signal Processing* **7** 1291–1294.
- RAFTERY, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B.* **47** 528–539.
- RAFTERY, A. E. and AKMAN, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73** 85–89.
- ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2** 13–25.
- ROY, V. (2012). Spectral analytic comparisons for data augmentation. *Statistics & Probability Letters* **82** 103–108.
- STEPHENS, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B.* **62** 795–809.

TAC (2011). *Transport Accident Commission, Victoria, Online Crash Database.*

URL <http://www.tac.vic.gov.au/road-safety/statistics/online-crash-database>

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82** 528–550.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

WELLEKENS, C. (1987). Explicit time correlation in hidden Markov models for speech recognition. *Acoustics, Speech & Signal Processing* **12** 384–386.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9** 60–62.