Bayesian Modelling of Dependent Data



Daniel Moss

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

To Zsóf, my beloved wife.

Acknowledgments

First and foremost, I wish to thank my outstanding supervisor, Judith Rousseau, whose support and guidance over the past four years has been invaluable, and with whom it has been a great pleasure to collaborate on these projects. I appreciate immensely all of the time you have taken to help me understand the complexities and nuances of the subject.

Next, I wish to thank all members of the StatML CDT, and also thank my friends in the Department of Statistics. I give special thanks to Sahra, who has always been willing to take a walk with me and share her advice, relating to the DPhil but also beyond. I give special thanks also to Matteo, with whom I have had many interesting conversations both statistical and otherwise, and who offered his hospitality so generously at his home in Oxford; for this I also extend thanks to Silvia for being so accommodating.

I thank Neil, with whom I shared my undergraduate years at Emmanuel College in Cambridge, and now my DPhil journey in closely related fields. You have been a true friend to me throughout these eight years, and your support both mathematically and personally is deeply appreciated.

I thank my parents for their continued encouragement, and all of the opportunities with which they have provided me. I thank my elder sister, Rebecca, for her love and kindness towards me throughout my life. I thank my sister-in-law, Nikolette, for her good humour and company, especially during the lockdowns. I thank also my secondary school mathematics teacher, Richard, for his encouragement throughout my teenage years to stretch myself mathematically.

Finally, I thank my wife, Zsóf, to whom I dedicate this thesis. You have always brought out the best in me, and encouraged me to strive for excellence in everything I do, both personally and academically. It is a great pleasure and a privilege to have such a supportive partner, and as I look forward to our future together, I am filled with happiness.

Abstract

Bayesian approaches to statistical modelling allow practitioners to coherently update prior beliefs based on observed data, and encode these updated beliefs in the *posterior distribution*. The resulting inferences are then supported by a meaningful measure of uncertainty, and as such the Bayesian statistical paradigm is attractive in a number of applications.

In a variety of these applications, however, the assumption that data is conditionally independent, given a common data generating process, is simply not well-founded. As such, it is essential that Bayesian approaches, which nevertheless retain their rigour, coherence, and practical appeal, be formulated and studied in these contexts.

In this thesis, we examine the Bayesian modelling of dependent data, both in theory and in practice. We place particular focus on semiparametric Hidden Markov models. These models are based upon the idea that many processes may possess a simple, Markovian, latent structure underlying dependence in data. This latent structure cannot be observed directly, and the relationship between this latent structure, and the observations through which our posterior distribution is formed, may be very complex and therefore not easily captured by finite dimensional models.

We study the theoretical properties of such models from a frequentist perspective, which may equally be viewed from the fully subjective Bayesian perspective as a notion of robustness to prior choice. In doing so, we propose the application of a modular Bayesian approach, the cut posterior, to semiparametric inference more broadly, showing its specific benefit in the model we study.

Alongside this line of work, we pursue a second, thematically related, line of work relating to Bayesian modelling of conflict data, which exhibits dependencies both in space and time. Much like the HMMs we study, such models can be interpreted as capturing an underlying structure in the data, and we will see that inference in such models provides us with interpretable and meaningful insights into conflict events.

Contents

1 Introduction **1**

1.1 Bayesian Modelling of Dependent Data 1

1.2 HMMs, Mixtures and Identifiability 4

1.2.1 Hidden Markov Models 4

1.2.2 Identifiability of Mixtures and HMMs 8

1.3 Efficiency, Bernstein-von Mises and Semiparametric Estimation 13

1.3.1 Efficient Estimation 14

1.3.2 Bernstein von-Mises 18

1.3.3 Efficiency in semiparametric models 20

1.4 Cut Posterior 25

1.4.1 Multiple data sources 26

1.4.2 Multiple prior models 27

1.5 Concentration Rates 28

1.5.1 Framework for concentration in Bayesian nonparametrics 29

1.5.2 Inversion Inequalities 31

1.6 Bayesian Model Selection 32

1.7 Bayesian Analysis of Conflict Risk 33

1.7.1 Discrete-time Hawkes processes 33

1.7.2 Statistical Modelling of Conflict and the ACLED Dataset 34

1.8 Summary of Contributions 35

2 Efficient Bayesian Estimation and use of Cut Posterior in Semiparametric HMMs. **37**

2.1 Introduction 37

2.2 Inference in semiparametric finite state space HMMs 40

2.2.1 Model and notation 41

2.2.2 Cut posterior inference: A general strategy for joint inference on Q and $\mathbf{f}$ . . . 43
2.3 Semi - parametric estimation of Q: Bernstein-von Mises property and efficient estimation . . . 45
2.3.1 Asymptotic normality and Bernstein-von Mises . . . 46
2.3.2 Efficient estimation . . . 49
2.4 Cut posterior contraction . . . 56
2.4.1 Concentration of marginal densities $g^{(\ell)}_{Q,\mathbf{f}}$ . . . 56
2.4.2 Concentration of emission distributions . . . 58
2.4.3 Concentration of smoothing distributions . . . 58
2.4.4 Example: Dirichlet process mixtures of Gaussians . . . 59
2.5 Practical considerations and simulation study . . . 60
2.5.1 MCMC algorithm for Q . . . 61
2.5.2 MCMC algorithm for $\{f_r\}_{r=1}^R$ . . . 63
2.6 Proofs of main results . . . 71
2.6.1 Proof of Proposition 2.5: Convergence of the scores . . . 71
2.6.2 Proof of Theorem 2.10 . . . 79
2.6.3 Proof of Theorem 2.13 . . . 81
2.7 Conclusion and discussion . . . 87
2.8 Further Proofs of main results . . . 88
2.8.1 Proof of Theorem 2.6 . . . 88
2.8.2 Proof of Proposition 2.7 . . . 90
2.9 General theorem for cut posterior contraction . . . 91
2.10 Proofs of technical results . . . 96
2.10.1 Non-singularity of Fisher Information . . . 97
2.10.2 Fisher Information and asymptotic lower bound . . . 108
2.10.3 Technical results for the deconvolution argument . . . 109
2.10.4 Admissible partitions . . . 114
2.10.5 Technical results for the proofs of Theorems 2.10 and 2.13 . . . 115
2.11 Assumptions required for the application of Theorem 2.12 to Dirichlet process mixtures . . . 119
2.12 Key results from literature . . . 120
2.12.1 Forgetting of the hidden Markov chain . . . 120
2.12.2 Convolution Theorem and Efficiency . . . 121

2.13 Details on MCMC algorithms used in Section 2.5 . . . 127
2.13.1 MCMC Algorithm for Π_1 . . . 127
2.13.2 MCMC Algorithm for Π_2 . . . 127
2.13.3 MCMC Algorithm for Π' . . . 128

3 Bayesian Model Selection in Semiparametric Mixtures and HMMs 129

3.1 Introduction . . . 129
3.2 Models and Notation . . . 131
3.3 Estimation of R^* . . . 133
3.3.1 Consistency of Penalised Sequential Marginal Likelihood Approach . 133
3.3.2 Regarding the penalisation . . . 136
3.4 Experiments . . . 138
3.4.1 Algorithm Description . . . 138
3.4.2 Categorical Simulations . . . 139
3.4.3 Semiparametric Simulations . . . 145
3.5 Conclusion . . . 150
3.6 Proofs . . . 150
3.6.1 Proof of Theorems 3.1 and 3.2 . . . 150
3.6.2 Proof of Proposition 3.4 . . . 153

4 Bayesian Analysis of Conflict Risk 158

4.1 Introduction . . . 158
4.2 Existing methodology for modelling the ACLED dataset . . . 160
4.2.1 Data and Analysis from ACLED . . . 160
4.2.2 Discrete time Hawkes processes and spatial dependency . . . 161
4.3 Structure of proposed models . . . 164
4.3.1 Multivariate model . . . 164
4.3.2 Likelihood models . . . 167
4.3.3 Implementation of Bayesian Approach . . . 170
4.4 Empirical Analysis and Model Application . . . 173
4.4.1 Simulated Data . . . 173
4.4.2 Application to Bangladesh data from ACLED . . . 181
4.5 Conclusion . . . 193

5 Conclusion **195**

5.1 Semiparametric Hidden Markov Models 195

5.2 Bayesian Modelling of Conflict Data 196

1. Introduction

1.1 Bayesian Modelling of Dependent Data

Given a collection of data, the task of statistical inference is to make statements about the characteristics of the random process giving rise to this data. If we have n data points, denoted $X_1, X_2, \ldots, X_n$, and satisfying $X_i \in \mathcal{X}$ for each $i = 1, \ldots, n$, this task can be described mathematically as saying that the collection $(X_1, X_2, \ldots, X_n)$ are a particular realisation of some underlying *distribution* P_n over the $n-$fold product $\mathcal{X}^n$, which we imagine as describing the universe of possible datasets of size n which we might have observed. The idea is that our inference should aim not only to summarise the particular data set which we have observed, but should also seek to generalise to the underlying distribution, so as to understand hypothetical collections of data in which the randomness present in P_n manifests in a different way.

In practice, we have only one realisation of P_n, and so in order to make meaningful generalisations to other hypothetical datasets, we must impose some structure on the form of P_n. The imposition of this structure is referred to as *statistical modelling*, and amounts to the specification of a (possibly large) set $\mathcal{P}_n$, called a *model*, for which we assume $P_n \in \mathcal{P}_n$.

A typical structural assumption for the model $\mathcal{P}_n$ is that the distributions P_n over $\mathcal{X}^n$ factorise as a product. This means that

$$P_n(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P^{(1)}(X_1 \in A_1)P^{(2)}(X_2 \in A_2)\ldots P^{(n)}(X_n \in A_n) \tag{1.1}$$

for some subsets $A_1, A_2, \ldots, A_n \subset \mathcal{X}$, and some distributions $P^{(1)}, P^{(2)}, \ldots, P^{(n)}$ each over $\mathcal{X}$. This describes a model under which the randomness present in each observation is *independent* of the others.

A simple example of such a model is the *many normal means* model, in which each $P^{(i)}$

is the distribution of a one-dimensional Gaussian centred at mean $\mu_i \in \mathbb{R}$. Regression may also be considered in this framework by choosing the P_i to have some characteristics varying according to some exogenous covariates C_i associated to the measurement X_i.

Often, one assumes simply that all the $P^{(i)}$ are equal to a common distribution P, so that the data X_i are *independent and identically distributed* (iid). A simple example of such a model would be the repeated tossing of a biased coin, whose probability p of landing on heads is unknown. Intuitively, once n is sufficiently large, we will understand very well the process by which further data would be generated, and in this way we can extrapolate from our original data to make statements about hypothetical datasets. It is precisely the imposition of the structure on $\mathcal{P}_n$ which makes this step of extrapolation possible; if a new coin of unknown bias were minted before each coin toss, the data from the first n coin tosses would give no insight about subsequent results.

However, there are many situations where the assumption of independence is inappropriate, in the sense that one would not reasonably expect the requirement that $P_n \in \mathcal{P}_n$ to hold. One such instance is in *time series* data, in which the index i counts units of time, and the variable X_i gives the state of some system at the time i. Such data arises in a broad range of fields, with examples including finance, climate modelling, seismology and speech recognition [94]. In such cases, we must seek alternative structures for the model $\mathcal{P}_n$ which allow for dependence between data, while still having sufficient structure to permit meaningful inference.

Once we fix a model $\mathcal{P}_n$, and have at hand data $X_1, X_2, \ldots, X_n$ the problem of inference can be formulated as attempting to find the element $P_n \in \mathcal{P}_n$ which generated the data. This task is often referred to as *point estimation*. In practice, the finiteness of the data means that we will never know precisely the generating distribution P_n, and so the task of statistical inference also extends to *uncertainty quantification*, in which we try to make statements about subsets of $\mathcal{P}_n$ which are, in some sense, more or less likely. There exist two statistical paradigms, the *frequentist* and *Bayesian* approaches, which differ in their approach to these two problems, of point estimation and uncertainty quantification.

The frequentist statistician asserts the existence of some fixed 'true' distribution P_n^*, from which our data is generated. A classical frequentist procedure for obtaining point estimates is maximum likelihood estimation, where we estimate P_n^* with $\hat{p}_n d\mu$, with $\hat{p}_n$ a $\mu-$density given by

$$\hat{p}_n = \max_{p_n d\mu \in \mathcal{P}_n} p_n(X_1, X_2, \ldots, X_n), \tag{1.2}$$

where $p_n(X_1, X_2, \ldots, X_n)$ is the *likelihood* and denotes the probability density[1] of having observing the data $X_1, X_2, \ldots, X_n$, had it been generated according to P_n.

The uncertainty in our estimation procedure is based entirely on the randomness present in our data, and this is reflected in the frequentist's approach to uncertainty quantification, which typically involves the use of *confidence sets*. A confidence set (of coverage $1 - \alpha$) is simply a (random) subset $C_n(X_1, \ldots, X_n) \subset \mathcal{P}_n$ with the property that, under repeated generations of datasets $(X_1, \ldots, X_n)$ from the distribution P_n^*, and corresponding computation of the set $C_n = C_n(X_1, \ldots, X_n)$, we would have $P_n^* \in C_n$ a proportion $1 - \alpha$ of the time.

Contrarywise to the frequentist, the Bayesian statistician must specify not only the model $\mathcal{P}_n$ of possible data generating processes for the data, but also a *prior distribution* Π_n over the model $\mathcal{P}_n$, which reflects the statistician's epistemic uncertainty about the data generating process. The Bayesian statistician then 'updates' these prior beliefs about $\mathcal{P}_n$ once the data is observed, and so views the *a posteriori* probability distribution of the data generating process P_n as being given by

$$d\Pi(P_n | X_1, \ldots X_n) \propto d\Pi(P_n) P_n(X_1, \ldots, X_n) \tag{1.3}$$

with the normalising constant chosen so that the *posterior distribution* $\Pi(\cdot | X_1, \ldots X_n)$ satisfies $\Pi(\mathcal{P}_n | X_1, \ldots X_n) = 1$. The posterior distribution then naturally encodes uncertainty about our parameter space $\mathcal{P}_n$, with regions C_n of posterior probability $1 - \alpha$ being called *credible sets* (of coverage $1 - \alpha$). We may secondarily produce point estimates, typically by taking the mean or mode of the posterior distribution, but the entire process of inference is downstream of the posterior distribution.

The statistical model $\mathcal{P}_n$ is often most naturally expressed by writing

$$\mathcal{P}_n = \{P_n^{(\theta)} : \theta \in \Theta\} \tag{1.4}$$

for Θ some *index set* and $P_n^{(\theta)}$ some distribution associated to the particular value, or

[1] We assume here, and throughout this thesis, that each $P \in \mathcal{P}$ indeed admits a density p with respect to some common dominating measure μ - that is, that $\mathcal{P}$ is a *dominated family*. This assumption is technically important, but in practice is very weak for the kinds of models we study.

parameter, $\theta \in \Theta$. We then call Θ the *parameter space*.

Provided that the map $\theta \mapsto P_n^{(\theta)}$ is injective, meaning that $P_n^{(\theta)} = P_n^{(\theta')} \implies \theta = \theta'$ for all $\theta, \theta' \in \Theta$, estimation of the data generating process $P_n^{(\theta)}$ is equivalent to estimating the unique, corresponding $\theta \in \Theta$. The requirement that $\theta \mapsto P_n^{(\theta)}$ is injective is called *identifiability* of the parameter θ, or more generally of the set Θ. When taking a Bayesian approach, we can then replace Equation (1.3) by

$$d\Pi(\theta|X_1, \ldots, X_n) = \frac{d\Pi(\theta)P_n^{(\theta)}(X_1, \ldots, X_n)}{m(X_1, \ldots, X_n)}, \tag{1.5}$$

where $m(X_1, \ldots, X_n) := \int_\Theta P_n^{(\theta)}(X_1, \ldots, X_n)d\Pi(\theta)$ is the *marginal likelihood*.

In this thesis, *Bayesian Modelling of Dependent Data*, we focus on Bayesian approaches to inference in models $\mathcal{P}_n$ which permit dependency between observations. Although we focus on Bayesian inference, these procedures are analysed under the frequentist assumption of a true distribution P_n^* (or parameter θ^*), and we assume the existence of this throughout. The thesis encompasses two distinct lines of work: we consider Bayesian inference in HMMs in Chapters 2 and 3; the Bayesian estimation of point processes in then considered in Chapter 4 in the context of an application to conflict data. While each consider the use of a different family of models $\mathcal{P}_n$, we are in each case interested in the use of Bayesian models to model dependent data, and also we are interested in uncovering some latent structure of the process we model.

1.2 HMMs, Mixtures and Identifiability

1.2.1 Hidden Markov Models

In Chapters 2 and 3, we focus on finite state space, discrete time, semiparametric Hidden Markov Models. In what follows, we give a brief overview of these models.

Hidden Markov Models (HMMs) are families of models $\mathcal{P}_n$ for dependent data, whose dependence structure arises from the dynamics of some underlying (or hidden) *Markov chain* which we do not observe directly, but which influences the distribution of the data. One can view this unobserved variable as being the 'state' of a system which is not directly observable. HMMs earned widespread popularity for their application in speech recognition [40], but have since enjoyed broader use in a number of applications, including modelling of weather [3,60], circadian rhythms [59], animal behaviour [29,64], finance [75], information

retrieval [79, 99], biomolecular dynamics [56], genomics [107] and speech recognition [88]..

The problem of inference in HMMs primarily involves the understanding of the evolution of the unobserved state of the system, and the link between the unobserved state and the observable data. These two aspects provide a complete description of the data generating process; inference on such quantities as the distribution over the states given the data can then be considered downstream of these inferences.

While there is considerable literature on *parametric* HMMs, in which the model $\mathcal{P}_n$ is a finite dimensional subspace of $\mathbb{R}^d$ for some $d < \infty$, the literature for *nonparametric* HMMs, in which $\mathcal{P}_n$ is of infinite dimension, is comparatively sparse. In the following sections, we provide the background to the material of Chapters 2 and 3, each of which tackle key problems in the Bayesian inference of nonparametric Hidden Markov Models.

1.2.1.1 Markov Chains

Let $(X_t)_{t\geq 1}$ be a sequence of $\mathcal{X}-$valued random variables for which $X_1 \sim \nu$ for some probability measure ν on $\mathcal{X}$, and such that, for each $t > 1$, the distribution

$$\Pr(X_t \in dx | X_1, \ldots, X_{t-1}) = Q(X_{t-1}, dx) \tag{1.6}$$

for some *transition kernel* $Q(X_{t-1}, dx)$. The transition kernel has the property that, for each $x' \in \mathcal{X}$, $Q(x', dx)$ is a probability measure over $\mathcal{X}$.

Of particular note is that the distribution defined in (1.6) depends only on the immediately preceding value of the process $(X_t)_{t\geq 0}$; this property is the *Markov property* and the process $(X_t)_{t\geq 0}$ is correspondingly called a *Markov chain*.

In this thesis, we consider only the case in which $\mathcal{X}$ is finite, and henceforth we assume $\mathcal{X} = \{1, 2, \ldots, R\}$ for some $R < \infty$. In this case, the transition kernels $Q(X_{t-1}, dx)$ can be described entirely by a *transition matrix* $Q \in \mathbb{R}^{R\times R}$ whose entries satisfy

$$Q_{ij} = \Pr(X_t = j | X_{t-1} = i) \quad i, j \in \{1, \ldots, R\}. \tag{1.7}$$

As a consequence of the definition (1.7), the matrix Q has unit row sums, i.e.

$$\sum_j Q_{ij} = 1 \quad i \in \{1, \ldots, R\}. \tag{1.8}$$

The matrix Q of Equation (1.7) is obtained from the kernels Q described in Equation

(1.6) by stacking the probability mass functions of the discrete distributions $Q(x', dx)$ in rows, for each $x' \in \{1, \ldots, R\}$. Henceforth we shall always use Q to refer to the matrix definition.

Identifying the distribution ν with its probability mass function, valued in the $(R-1)$ dimensional simplex embedded into $\mathbb{R}^R$, we call a distribution ν *invariant* for the transition matrix Q if it satisfies $\nu Q = Q$. This ensures that, if $X_1 \sim \nu$, then $X_t \sim \nu$ also for all $t > 1$. In this way, the $(X_t)_{t\geq 1}$ are *identically distributed* (but are not independent).

For a full background of the theory of discrete time Markov chains, see Chapter 6 of [51].

1.2.1.2 Hidden Markov Models

Let $(X_t)_{t\geq 1}$ be a Markov chain valued in $[R] = \{1, \ldots, R\}$, with initial distribution ν and transition matrix Q. Let $(Y_t)_{t\geq 1}$ be a sequence of $\mathcal{Y}-$valued random variables with the property that

$$Y_t | (X_s)_{s\geq 1} \overset{indep.}{\sim} F_{X_t} \quad t \geq 1. \tag{1.9}$$

That is, conditionally on the chain $(X_s)_{s\geq 1}$, the Y_t are independent draws from the distribution F_{X_t}. Since $X_t \in [R]$, specification of the data generating process of Y_t given the X_t requires specification of the set of *emission distributions* $\mathbf{F} = \{F_1, F_2, \ldots, F_R\}$.

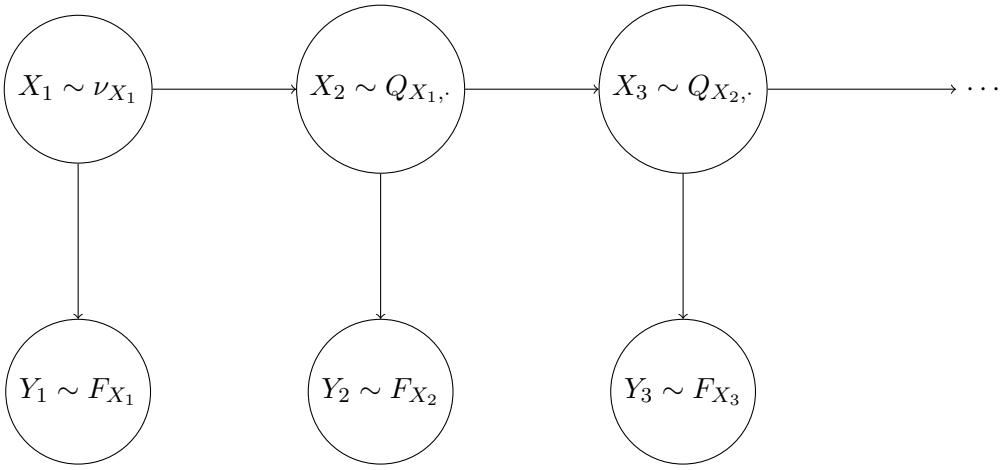


Figure 1.1: Visual representation of data generating process of HMM

Suppose now our data is $(Y_t)_{1\leq t\leq n}$, and the corresponding $(X_t)_{1\leq t\leq n}$ are unobserved. Then the $(Y_t)_{1\leq n}$ are no longer independent, and we say that the process $(Y_t)_{t\geq 1}$ is generated according to a *Hidden Markov Model* or *HMM*. Writing $\mathbb{P}_{1:n}^{(\nu,Q,\mathbf{F})}$ for the distribution of $(Y_t)_{1\leq t\leq n}$ under a specified choice of ν, Q and $\mathbf{F}$, the statistical model $\mathcal{P}_n$ then takes the form

$$\mathcal{P}_n = \{\mathbb{P}_{1:n}^{(\nu,Q,\mathbf{F})} : (\nu, Q, \mathbf{F}) \in \Delta_R \times \Delta_R^R \times \mathcal{P}(\mathcal{Y})^R\}, \tag{1.10}$$

where we write $\Delta_R = \{p \in \mathbb{R}^R : p_i \geq 0, \sum_{i=1}^R p_i = 1\}$ for the simplex of dimension $R-1$, and we write also $\mathcal{P}(\mathcal{Y})$ for the set of probability measures on $\mathcal{Y}$.

In Chapters 2 and 3, we will further restrict the parameter set to ensure the existence and uniqueness of a stationary distribution for Q. We then assume that the chain starts in equilibrium. This assumption may be justified practically in instances where the data arises from a process which is being measured from a particular point in time, but which has had ample time to reach equilibrium before measurement is made. From a theoretical point of view, the assumptions which ensure the existence and uniqueness of a stationary distribution provide also geometric ergodicity of the chain $(X_t)_{t\geq 1}$, and so asymptotic results are readily adapted to a differing initial distribution whose influence anyway decays exponentially in time. In light of this, we consider throughout models of the form

$$\mathcal{P}_n = \{\mathbb{P}_{1:n}^{(Q,\mathbf{F})} : (Q, \mathbf{F}) \in \Delta_R^R \times \mathcal{P}(\mathcal{Y})^R\}, \tag{1.11}$$

with the law $\mathbb{P}_{1:n}^{(Q,\mathbf{F})}$ being that of the process $(Y_t)_{1\leq t\leq n}$ when the hidden chain is started in its unique invariant distribution.

Technically we should restrict Δ_R^R to the subset of transition matrix which admit a unique invariant distribution, but in practice this is not necessary so long as we make the frequentist assumption that the true $(Q^*, \mathbf{F}^*)$ has the desired properties, and also the Bayesian assumption that the prior assigns zero probability to values in violation of this assumption.

Extension to indexing by $\mathbb{Z}$ Suppose $(X_t)_{t\geq 1}$ is a Markov chain started in its stationary distribution ν. Let

$$\tilde{Q}_{ij} = \Pr(X_t = i | X_{t+1} = j) = \frac{\nu_i Q_{ij}}{\nu_j} \quad i, j \in \{1, \ldots, R\}, t \geq 1 \tag{1.12}$$

Then $\tilde{Q}_{ij}$ are the *reversal* transition probabilities. We can then extend the chain $(X_t)_{t\geq 1}$ to $(X_t)_{t\in\mathbb{Z}}$ by defining $(X_t)_{t\leq 0}$ as arising from this reversal process, given X_1. Correspondingly, we can extend the process $(Y_t)_{t\geq 1}$ to $(Y_t)_{t\in\mathbb{Z}}$ by defining Y_t for each $t \in \mathbb{Z}$ as in (1.9), replacing $(X_s)_{s\geq 1}$ by $(X_s)_{s\in\mathbb{Z}}$.

From a statistical perspective, we then think of the 'hidden data' as being the full process $(X_t)_{t\in\mathbb{Z}}$, as well as the 'historical observables' $(Y_t)_{t\leq 0}$. We will see that this abstraction is necessary for the exposition of efficient estimation discussed in Section 1.3.

Relation to mixture models Suppose that we constrain the hidden states $(X_t)_{t\geq 1}$ to be independent and identically distributed random variables from a probability mass function $\pi \in \Delta_R$. Taking $(Y_t)_{1\leq t\leq n}$ again as in (1.9), we have that the observations are also iid, with the distribution of one observation given as

$$\tilde{\mathbb{P}}_1^{(\pi,\mathbf{F})} = \sum_{r=1}^{R} \pi_r F_r, \tag{1.13}$$

which is a *mixture* of the distributions $(F_r)_{1\leq r\leq R}$ according to the *weights* $(\pi_r)_{1\leq r\leq R}$. The resulting statistical model $\tilde{\mathcal{P}}_n$ is then a *mixture model*, with the distribution over $Y_{1:n}$ being simply the $n-$fold product of $\tilde{\mathbb{P}}_1$.

In Section 2.3 of Chapter 2, the problems we consider in the context of HMMs were previously addressed in the mixture setting [45]. Nevertheless, the proof techniques are substantially different owing to the additional dependency structure. In Chapter 3, we examine the mixture setting as an intermediate step towards the HMM setting.

1.2.2 Identifiability of Mixtures and HMMs

The theory of identifiability underpins our study of mixtures and HMMs in Chapters 2 and 3, and we introduce the relevant aspects of this body of literature in what follows.

1.2.2.1 Overview

In HMMs and mixtures, the requirement for identifiability of the parameters $(Q, \mathbf{F})$ or $(\pi, \mathbf{F})$ is not straightforward, as we require in particular for these parameters to be identified from the distribution over the observations $(Y_t)_{t\geq 1}$, rather than the distribution over the full collection $(X_t, Y_t)_{t\geq 1}$ of hidden states and observations (which we refer to as the *full data*). Since the distribution $\mathbb{P}_n^{(Q,\mathbf{F})}$ of $Y_{1:n}$ is simply the marginal distribution over the observations of the distribution over the full data, it is intuitively at least as difficult to recover the parameters from.

Example 1.1 (Failure of Identifiability in Mixtures)**.** Suppose $R = 3$ and $F_3 = qF_1 + (1-q)F_2$ with $0 < q < 1$, $F_1 \neq F_2$. Let $\pi = (\pi_1, \pi_2, \pi_3)$, and $\pi' = (\pi_1 + q\eta\pi_3, \pi_2 + (1-q)\eta\pi_3, (1-\eta)\pi_3)$ with $0 < \eta < 1$. Let $\mathbf{F} = (F_1, F_2, F_3)$ and $\mathbf{F}' = (F_1, F_2, F_3)$.

Then $\tilde{\mathbb{P}}_1^{(\pi,\mathbf{F})} = \tilde{\mathbb{P}}_1^{(\pi',\mathbf{F}')}$, and consequently $\tilde{\mathbb{P}}_n^{(\pi,\mathbf{F})} = \tilde{\mathbb{P}}_n^{(\pi',\mathbf{F}')}$ for all n by taking the $n-$fold product. But $(\pi, \mathbf{F}) \neq (\pi', \mathbf{F}')$ and so the mixture models $\tilde{\mathcal{P}}_n$ are not identifiable in full generality.

The non-identifiability in Example 1.1 is derived from the linear dependence between F_1, F_2, F_3. Notice that the parameters would be identified from the distribution over the full data, but are not identified from the marginal over the observations $Y_{1:n}$.

Example 1.2 (Failure of Identifiability in HMMs)**.** Supppose $R = 3$ and set

$$Q_{\eta_1} = \begin{pmatrix} 0.2 & 0.8\eta_1 & 0.8(1-\eta_1) \\ 0.1 & 0.9\eta_1 & 0.9(1-\eta_1) \\ 0.1 & 0.9\eta_1 & 0.9(1-\eta_1) \end{pmatrix}.$$

Let F_1, F_2, F_3 linearly independent probability measures, and set

$$\mathbf{F}_{\eta_2,\eta_3} = (F_1, \eta_2 F_2 + (1-\eta_2)F_3, (1-\eta_3)F_2 + \eta_3 F_3).$$

Here $0 < \eta_1, \eta_2, \eta_3 < 1$. Then given η_1, choosing η_2, η_3 such that $\eta_1 = \dfrac{1-\eta_3}{2-\eta_3-\eta_2}$, we obtain the same distribution over the observed data as if $\eta_2 = \eta_3 = 1$. As a concrete example, we could choose $\eta_1 = \frac{3}{7}$ and then both $(\eta_2, \eta_3) = (1, 1)$ and $(\eta_2', \eta_3') = (\frac{1}{3}, \frac{1}{2})$ give the same marginals over $Y_{1:n}$.

The non-identifiability in Example 1.2 arises from the singularity of the transition matrix Q_{η_1}. To see this, set $\tilde{X}_t = \mathbb{1}_{X_t=1} + 2\mathbb{1}_{X_t \in \{2,3\}}$. Set also $\tilde{F}_1 = F_1$ and $\tilde{F}_2 = \eta_1 F_2 + (1-\eta_1)F_3$. Then $\tilde{X}$ is a Markov chain with transition matrix

$$\tilde{Q} = \begin{pmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{pmatrix}$$

and the distribution of $Y_{1:t}$, when $Y_t | X_t \sim \tilde{F}_{\tilde{X}_t}$, coincides with the distribution under each set of parameters described in Example 1.2.

Examples 1.1 and 1.2 suggest that we should pay particular attention to the cases in which the emission distributions $\mathbf{F}$ are linearly dependent, and (in the HMM context) the cases in which Q is rank-deficient. In fact, there is always an inevitable lack of identifiability for *any* mixture model or HMM, as the next example shows.

Example 1.3 (Failure of Identifiability due to Label Swapping)**.** Suppose we have a parameter configuration $(\pi, \mathbf{F})$ for mixtures, or $(Q, \mathbf{F})$ in HMMs. Given a permutation $\tau \in \mathcal{S}_R$, ${}^{\tau}\mathbf{F} = (F_{\tau(1)}, \ldots, F_{\tau(R)})$, ${}^{\tau}\pi = (\pi_{\tau(1)}, \ldots, \pi_{\tau(R)})$, and $(({}^{\tau}Q)_{ij} : i, j \leq R) = (Q_{\tau} i), \tau(j) : i, j \leq R)$. Then $\tilde{\mathbb{P}}_1^{(\pi,\mathbf{F})} = \tilde{\mathbb{P}}_1^{({}^{\tau}\pi, {}^{\tau}\mathbf{F})}$, and also $\mathbb{P}^{(Q,\mathbf{F})} = \mathbb{P}^{({}^{\tau}Q, {}^{\tau}\mathbf{F})}$.

Example 1.3 shows that identifiability is only ever achievable up to a relabelling of the labels of the hidden states, which we call *identifiability up to label swapping* (or *label switching*). However, these state labels are fundamentally arbitrary so this is not a real problem. Going forward, we will re-appropriate the term 'identifiable' in the context of these models to mean 'identifiable up to label swapping.'

One way to guarantee identifiability in the mixtures setting is by restricting $\tilde{\mathcal{P}}_n$ so that the emission distributions $\mathbf{F}$ vary in some parametric (finite-dimensional) family.

Example 1.4 (Identifiability of Gaussian Mixtures)**.** Set

$$\tilde{\mathcal{P}}_n = \{\tilde{\mathbb{P}}_1^{(\pi,\mathbf{F}_\theta)} \underbrace{\otimes \cdots \otimes}_{n \text{ times}} \tilde{\mathbb{P}}_1^{(\pi,\mathbf{F}_\theta)} : \pi \in \Delta_R, \theta \in \mathbb{R}^R\}.$$

with the emission distributions $\mathbf{F}_\theta = (\mathcal{N}(\theta_1, 1), \ldots, \mathcal{N}(\theta_R, 1))$ each unit-variance Gaussian centred at some $\theta_r \in \mathbb{R}$. We assume also that the θ_r are distinct - the frequentist may place this assumption on the true data generating process, and the Bayesian may choose a prior for which this holds almost surely (which is straightforward). Then $\tilde{\mathcal{P}}_n$ is identifiable (up to label swapping) by considering the characteristic function and using linear independence of the resulting exponential functions.

1.2.2.2 Nonparametric Identifiability

In Example 1.4 we achieve identifiability by restricting the emission distributions to a parametric family, consequently $\tilde{P}_n$ is a *parametric model*. Similarly we would get identifiability in the HMM case with Gaussian emissions, and many other parametric families (such as Beta or Gamma) could be chosen to match the support of the data. As discussed in Section 4.1, we now pay a price in the flexibility of our model, which no longer contains data generating processes in which the observations are non-Gaussian.

In the following section, we examine a part of the literature on nonparametric identifiability in mixtures and HMMs. Further examples can be found in [41], but we focus on the results relevant to this thesis. The key results we present can be roughly described as follows:

- Nonparametric mixtures, where the observations are three dimensional and each dimension is conditionally independent given the hidden state, are identifiable (under minimal further assumptions)
- Hidden Markov models are identifiable under minimal assumptions

Of particular interest in this thesis is the latter result, which essentially states that the *only* causes of unidentifiability (modulo label-swapping) for HMMs are those presented in Examples 1.1 and 1.2. We are able to make do with fewer assumptions on the form of the emission densities because of the dependency structure of the HMM: essentially the joint densities over consecutive observations carry more information that the density of a single observation, which need not be so informative by itself. The origin of these results is the following algebraic result of Kruskal [63], which we simplify somewhat for our exposition.

Theorem 1.5 (Kruskal, 1977)**.** *Suppose a three-way tensor $T = (T_{ijk})$ satisfies*

$$T_{ijk} = \sum_{\ell=1}^{R} A_{i\ell} B_{j\ell} C_{k\ell} = \sum_{\ell=1}^{R} \tilde{A}_{i\ell} \tilde{B}_{j\ell} \tilde{C}_{k\ell}. \tag{1.14}$$

Suppose $A, B, C \in \mathbb{R}^{d \times R}$ are each of rank R. Then there exists a permutation matrix $P \in \mathbb{R}^{R \times R}$ and diagonal matrices $D_A, D_B, D_C \in \mathbb{R}^{R \times R}$ such that

$$\tilde{A} = APD_A, \quad \tilde{B} = BPD_P, \quad \tilde{C} = CPD_C, \quad D_A D_B D_C = I.$$

The result in some sense states that the matrices A, B, C are 'identified' by T up to a kind of permutation and rescaling. In our context, the rescaling is not possible as the matrices contain probabilities and are constrained in their row sums; the permutation P arises due to the possibility of label-swapping which remains present in our setting.

The result was then used by Allman et al. [5] to prove identifiability for three-dimensional mixtures with conditionally independent components, given the states. This conditional independence means that the distribution of an observation takes the form

$$G = \sum_{\ell=1}^{R} p_r F_r^{(1)} \otimes F_r^{(2)} \otimes F_r^{(3)}, \tag{1.15}$$

where R is the number of hidden states. By defining (for $c = 1, 2, 3$) matrices $\tilde{F}_{rj}^{(c)} = F_r(A_j^c)$ for partitions $(A_j^c)_{j=1}^d$ of the observation space, it is possible to make the link between Equation (1.15) and Equation (1.14).

Theorem 1.6 (Allman et al. 2009)**.** *Let G be a mixture of the form (1.15), such that for every $j \in \{1, 2, 3\}$, the measures $\{F_r^{(c)}\}$ over $\mathbb{R}^{d_c}$ are linearly independent. Then $(p_r, F_r^{(c)} : r = 1, \ldots, R; c = 1, 2, 3)$ are identifiable up to label-swapping.*

This result was then used in [4,44] to achieve identifiability in HMMs. In [42], the authors

show that the parameters are identified from three consecutive observations whenever the HMM has linearly independent emission distributions, and a transition matrix Q of full rank. Essentially, this is because three consecutive observations of a HMM can be viewed as a single observation from a three-dimensional mixture, with the associated hidden state of the single observation being the HMM state associated to the second observation.

The dependency structure of the HMM then means that the three consecutive observations are conditionally independent given this state; the result for mixture models then applies by writing the distribution of three consecutive observations in the form (1.15), with $p_R = \mathbb{P}^{(Q,\mathbf{F})}(X_2 = r)$, and $F_r^{(c)} = \mathbb{P}^{(Q,\mathbf{F})}(Y_c \in \cdot | X_2 = r)$ for $c = 1, 2, 3$. Linear independence of $\mathbf{F}$ together with the full-rank assumption on Q implies that the sets $\{F_r^{(c)} : r = 1, \ldots, R\}$ are linearly independent for each $c = 1, 2, 3$; Theorem 1.6 then allows us to identify the $\mathbf{F}^{(c)} = \{F_r^{(c)}\}_{r=1}^R$, from which $\mathbf{F}$ and Q can then be inferred.

Theorem 1.7 (Gassiat et al. 2015)**.** *Assume R is known, that the probability measures $F_1, \ldots, F_R$ are linearly independent, and that Q has full rank. Then the parameters Q and $\mathbf{F}$ are identifiable from the distribution of three consecutive observations Y_1, Y_2, Y_3, up to label-swapping.*

In [4], the requirement of linearly independent emissions is weakened, so that we require only the $F_1, \ldots, F_R$ are distinct, but in so doing we require the distribution over more consecutive observations in order to identify the parameter. Specifically, they consider the joint distribution over $2R + 1$ consecutive observations, with they split into the middle observation and a block of R observations on each side. The result of [5] is then applied with $(d_1, d_2, d_3) = (R, 1, R)$; the crux of the argument is that the conditional distributions over the blocks of R observations either side of the middle observation, given the middle state, are linearly independent as soon as the $F_1, \ldots, F_R$ are distinct.

Theorem 1.8 (Alexandrovich et al. 2015)**.** *Assume R is known, that the probability measures $F_1, \ldots, F_R$ are distinct, and that Q has full rank and is ergodic. Then the parameters Q and $\mathbf{F}$ are identifiable from the distribution of three consecutive observations $Y_1, Y_2, \ldots, Y_{2R+1}$, up to label-swapping.*

A subtlety in both results is that the assumptions on the true data generating process need only hold for one reference parameter configuration. This means that, so long as the distribution of the HMM arises from some configuration of parameters $Q, \mathbf{F}$) satisfying the restrictions of the previous results, it cannot also arise from a different set of parameters

$(\tilde{Q}, \tilde{\mathbf{F}})$, even if we do not require that the different parameters satisfy any of the same requirements in terms of rank and linear independence.

In [4], this observation is leveraged to further show that the number R of hidden states may itself be identified from the distribution of sufficiently many consecutive observations. Suppose the HMM parameters $(Q, \mathbf{F})$ satisfy the assumptions of Theorem 1.8, with R states. If the HMM was expressible using only $R' < R$ states, we could embed this R' state representation as a set of parameters $\tilde{Q}, \tilde{\mathbf{F}}$ over R states, where $R - R'$ of the states are not visited. This $\tilde{Q}$ would be non-ergodic and so different to Q, but this is prohibited by Theorem 1.8. Inverting this argument allows us to identify the minimal number R of states required to express the HMM, although now the identification is only *within the class* of parameters satisfying the stated assumptions (contrary to the earlier results which required only a reference parameter satisfying the assumptions).

We conclude with their final result

Corollary 1.9 (Alexandrovich et al. 2015)**.** *Within the class of parameters* $(Q, \mathbf{F})$ *for which* Q *is ergodic and full-rank, and* $\mathbf{F}$ *are distinct, the number of states* R *are identified from the distribution of the observed process* $(Y_t)_{t \in \mathbb{N}}$.

If we assume an a priori upper bound $R_{\max}$ on the possible number of states, we require only $R_{\max} = (2R_{\max}+1)(R_{\max}^2 - 2R_{\max}+2)+1$ observations. The reason for the number of observations being higher than before is that Corollary 1.9 relies on a version of Theorem 1.8 in which the initial distribution need not be the stationary one.

Corollary [4] paves the way for principled estimation of the *order* R of the HMM, being the minimal number of hidden states required to express the distribution of the observations. The estimation of this quantity is explored in Chapter 3.

1.3 Efficiency, Bernstein-von Mises and Semiparametric Estimation

In the following section, we introduce the concepts studied in the first half of Chapter 2. In Section 1.3.1 we outline basic concepts of efficient estimation in parametric, iid models. In Section 1.3.2 we detail the Bernstein-von Mises result in such models. In Section 1.3.3 we introduce the notion of semiparametric estimation in iid models. The extension of the iid concepts to the HMM context is deferred to the body of Chapter 2.

1.3.1 Efficient Estimation

Consider a parametric model $\mathcal{P}_n$ over iid data $X_1, \ldots, X_n$, indexed by a finite-dimensional parameter space $\theta \in \Theta$. Let θ^* be the true parameter, so that the data is generated as $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_{\theta^*}$. Suppose now we produce estimates $\hat{\theta}_n(X_1, \ldots, X_n)$ for θ^*, via some estimation procedure (which can be Bayesian, but need not be).

Asymptotic theory for estimation in parametric models seeks to understand the limiting behaviour of $\hat{\theta}_n(X_1, \ldots, X_n)$ as $n \to \infty$. A basic asymptotic guarantee, known as *consistency*, ensures that

$$\hat{\theta}_n(X_1, \ldots, X_n) \to^{\mathbb{P}_{\theta^*}} \theta^*, \tag{1.16}$$

with the convergence in some probabilistic sense. Going one step further, results may seek to understand the *rate* at which this convergence takes place. Such results typically take the form

$$R_n(\hat{\theta}_n - \theta^*) = O_{\mathbb{P}_{\theta^*}}(1), \tag{1.17}$$

for some suitably chosen $R_n \to \infty$[2].

Example 1.10 (Shifted Exponential)**.** Let $\Theta = \mathbb{R}$ and $\mathbb{P}_\theta$ admit the density

$$f_\theta(x) = e^{-(x-\theta)} \mathbb{1}_{x>\theta}$$

with respect to the Lebesgue measure. Then $X_1, \ldots, X_n$ are draws from a unit rate *shifted exponential* with true shift θ^*. Let $\hat{\theta}_n = \min_{1 \le i \le n} X_i$. Then it is straightforward to show that $n(\hat{\theta}_n - \theta^*)$ is a standard exponential random variable, and in particular is bounded in probability. Thus Equation (1.17) holds for the choice $R_n = n$, but not for any faster rates R_n.

The model of Example 1.10 is somewhat unusual in that it does not satisfy the assumptions of so-called *regular* parametric models. Such regularity assumptions typically involve conditions on the derivatives of the log-likelihood $\ell_\theta = \log f_\theta$, and in particular do not hold in the previous example. These regularity assumptions can be conveniently collected in the single (albeit slightly more complex) assumption of *quadratic mean differentiability*, as described in Section 7.2 of [100].

[2]We say a sequence of random variables Z_n is $O_{\mathbb{P}_{\theta^*}}(1)$ if, for all sequences $M_n \to \infty$, $\mathbb{P}_{\theta^*}(Z_n > M_n) \to 0$. Intuitively M_n can increase arbitrarily slowly, but Z_n is asymptotically bounded and so M_n always increases faster.

Definition 1.11. An iid model with data $X_i \in \mathcal{X}$ $(i = 1, \ldots, n)$ and parameter space $\Theta \subset \mathbb{R}^d$ and densities f_θ with respect to a dominating measure μ is *differentiable in quadratic mean* (QMD) at θ if the root density $\theta \mapsto \sqrt{f_\theta} \in L^2(\mathbb{R}^d, \mu)$ is differentiable as a functional in L^2, meaning that there exists a vector of measurable functions $\dot{\ell}_\theta = (\dot{\ell}_{\theta,1}, \ldots, \dot{\ell}_{\theta,d})$ such that

$$\sqrt{f_{\theta+h}} = \sqrt{f_\theta} + \frac{1}{2} h^T \sqrt{f_\theta} \dot{\ell}_\theta + o_{L^2}(\|h\|). \tag{1.18}$$

Equation (1.18) resembles a pointwise expansion

$$\sqrt{f_{\theta+h}(x)} = \sqrt{f_\theta(x)} + h^T g_x(\theta) + o(\|h\|) \quad x \in \mathbb{R}^d, \tag{1.19}$$

with $g_x(\theta) \in \mathbb{R}^d$ the usual gradient at θ of the function $\theta \mapsto f_\theta(x)$ when x is fixed. Indeed, under the conditions of Lemma 7.6 of [100], we have a matching between the functions $x \mapsto g_x(\theta)$ and $\frac{1}{2}\sqrt{f_\theta}\dot{\ell}_\theta$, and in this case $\dot{\ell}_\theta = \nabla_\theta \log f_\theta$, the gradient of the log-likelihood at θ. We call $\dot{\ell}$ the *score function*.

Example 1.12 (Gaussian Location model)**.** Take instead the model with (Lebesgue) densities

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\theta)^2}{2\sigma^2} \right\},$$

where σ is fixed and known. Let $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\sqrt{n}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, \sigma^2)$, and so Equation (1.17) holds for $R_n = \sqrt{n}$, but not for any faster rates R_n.

We can see in Example 1.12 that $\hat{\theta}_n$ achieves the $\sqrt{n}-$rate and, because we have access to its exact distribution, we know that this is the best R_n we can get for this $\hat{\theta}_n$.

It turns out that estimation of a parameter θ in regular (QMD) models is 'asymptotically equivalent' to estimation of the θ parameter in the Gaussian Location model. The precise sense in which this is true is detailed in Chapter 7 of [100], and involves a property of regular models known as *Local Asymptotic Normality* (LAN) which holds under QMD. Consequently, it turns out that for QMD models (such as Gaussian, Student-t, Beta etc.), the *best possible* rate in Equation (1.17) is $R_n = \sqrt{n}$.

Theorem 1.13 (LAN for iid models, Theorem 7.2 of [100])**.** *Suppose the model with densities $(f_\theta : \theta \in \Theta)$ is QMD. Then the function $\dot{\ell}_\theta : \mathcal{X} \longrightarrow \mathbb{R}^d$ (with θ fixed) has expectation zero under $\mathbb{P}_\theta$, i.e. $\mathbb{E}_\theta[\dot{\ell}_\theta] = 0$, and the expectation $I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^T]$ exists and is*

called the Fisher Information. Then for every sequence with $h_n \to h$ as $n \to \infty$, we have

$$\log \prod_{i=1}^{n} \frac{p_{\theta+h_n/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i) - \frac{1}{2} h^T I_\theta h + o_{\mathbb{P}_\theta}(1). \tag{1.20}$$

Given an iid model $\{f_\theta : \theta \in \Theta\}$, we can further seek to understand the limiting shape of the quantity $T_n = \sqrt{n}(\hat{\theta}_n - \theta^*)$ appearing in Equation (1.17). If we can exhibit a random variable T for which $T_n \Rightarrow^{\mathbb{P}_{\theta^*}} T$, then we can construct asymptotically valid confidence sets based on the quantiles of T. Here $\Rightarrow^{\mathbb{P}^*}$ denotes weak convergence under $\mathbb{P}^*$. If we can further choose $\hat{\theta}_n$ so that the corresponding T is as tightly distributed around θ as possible, then the corresponding confidence sets will be of optimal size. Such estimators are called *efficient* and are explored at length in Chapter 8 of [100]. The key results are the convolution theorems (Theorems 8.8 and 8.9) as well as the local asymptotic minimax result of Thereom 8.11.

We state the first of these results, which gives the optimal T when $\hat{\theta}_n$ is a *regular estimator*, intended to rule out 'pathological' examples such as the Hodges estimator (Example 8.1. of [100]).

Definition 1.14 (Regular Estimator)**.** An estimator $\hat{\theta}_n$ is called *regular* at θ if, for all fixed $h \in \mathbb{R}^d$,

$$\sqrt{n}\left(\hat{\theta} - (\theta + \frac{h}{\sqrt{n}})\right) \overset{\mathbb{P}_{\theta+h/\sqrt{n}}}{\Longrightarrow} T_\theta \tag{1.21}$$

to a distribution T_θ which does not depend on h (but may depend on θ).

Theorem 1.15 (Convolution Theorem, Theorem 8.8 of [100])**.** *Let the model $\{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be QMD at θ and let the Fisher Information I_θ be non-singular. Let $\hat{\theta}_n$ be a regular estimator at θ, with limit distribution T. Then there exists a probability measure M_θ such that*

$$T_\theta = M_\theta * \mathcal{N}_d(0, I_\theta^{-1}),$$

where $*$ denotes the convolution. In particular, the covariance matrix of L_θ satisfies

$$T_\theta - I_\theta^{-1} \succeq 0.$$

Convolutions of two distributions P_1, P_2 can be understood as the distribution of $Z_1 + Z_2$ where Z_i are independent random variables with distribution P_i. In this way, the limiting distribution T can be thought of as involving some unavoidable noise $Z_1 \sim \mathcal{N}_d(0, I_\theta^{-1})$, and some additional independent noise $Z_2 \sim M_\theta$. Ideally we would have $M_\theta = \delta_0$ so

$Z_2 = 0$, in this case our estimators are 'best possible' or *efficient*. We summarise this in the following definition.

Definition 1.16 (Efficient Estimator)**.** An estimator $\hat{\theta}$, regular at $\theta \in \Theta$, is *efficient* for estimation of θ if

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\mathbb{P}_\theta}{\Longrightarrow} \mathcal{N}_d(0, I_\theta^{-1}).$$

It turns out that the maximum likelihood estimator $\hat{\theta}$ is efficient in these models, and we outline roughly the arguments here to set the scene for what follows. Consider the maximum likelihood estimator $\hat{\theta}$ satisfying

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell_\theta(X_i)}_{=\ell_\theta^n}, \tag{1.22}$$

which exists by compactness of Θ. Now $\dot{\ell}_{\hat{\theta}}^n = 0$, hence a Taylor expansion yields

$$0 = \dot{\ell}_{\hat{\theta}}^n = \dot{\ell}_\theta^n + (\hat{\theta} - \theta)\ddot{\ell}_{\bar{\theta}}^n, \tag{1.23}$$

for $\bar{\theta}$ lying on the line segment connecting θ and $\hat{\theta}$, and $\ddot{\ell}_\theta$ the Hessian of ℓ_θ. Under regularity assumptions the Fisher information satisfies $\mathbb{E}_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T = -\mathbb{E}_\theta \ddot{\ell}_\theta$.

Consistency of $\hat{\theta}$ in $\mathbb{P}_\theta-$probability for estimating θ follows from straightforward compactness arguments and the fact that θ maximises the expectation $\mathbb{E}_\theta \ell_\theta$; this implies also consistency for $\bar{\theta}$. Applying then the law of large numbers to the expressions in Equation (1.23), assuming also a uniform law of large numbers for the Hessian in a neighbourhood of θ, we arrive at

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i) - \sqrt{n}(\hat{\theta} - \theta) I_\theta + o_{\mathbb{P}_\theta}(1). \tag{1.24}$$

Assuming invertibility of I_θ, we finally have

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{I_\theta^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i) + o_{\mathbb{P}_\theta}(1). \tag{1.25}$$

The asymptotic normality then follows from an application of the central limit theorem to the term $\Delta_{n,\theta} := \frac{I_\theta^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i)$, since each summand $\dot{\ell}_\theta(X_i)$ is an iid $\mathcal{N}_d(0, I_\theta)$ random variable.

It turns out that all efficient regular estimators satisfy Equation (1.25), this is shown

in Chapter 8 of [100]. In this way, we see that in some sense the maximum likelihood estimator is *the* efficient estimator for θ. If we set $\psi_\theta = I_\theta^{-1}\dot{\ell}_\theta$, we can further rewrite Equation (1.25) as

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(X_i) + o_{\mathbb{P}_\theta}(1). \tag{1.26}$$

Written in this form, we refer to ψ as the *influence function*.

1.3.2 Bernstein von-Mises

The Bernstein von-Mises theorem states that, in regular parametric models with iid data and frequentist true parameter θ^*, the posterior distribution over θ is asymptotically shaped as a Gaussian distribution of variance is $I_{\theta^*}^{-1}/n$, as soon as the prior has a positive density in a neighbourhood of θ^* and the Fisher information I_{θ^*} is non-singular. It states also that this Gaussian is centred at an efficient estimator $\hat{\theta}$, and an important consequence of this is that we get an asymptotic equivalence between Bayesian measures of uncertainty (i.e. credible sets) and frequentist measures of uncertainty (i.e. confidence sets) constructed using the asymptotic distribution given in Definition 1.16

The Bernstein-von Mises theorem, in its most general iid parametric form, is presented (along with its proof) in Chapter 10 of [100]. We state a version of it here for compact parameter sets Θ.

Theorem 1.17 (Bernstein-von Mises, Theorem 10.1 of [100])**.** *Let $\Theta \subset \mathbb{R}^d$ be compact, and let the model with $\mu-$densities $(f_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at θ^*. Assume also that I_{θ^*} is non-singular, and that the prior Π over Θ has a continuous and positive $\mu-$density π at θ^*. Then*

$$\|\Pi(\sqrt{n}(\theta - \hat{\theta}) \in \cdot | X_1, \ldots, X_n) - \mathcal{N}_d(0, I_{\theta^*}^{-1})\|_{TV} \overset{\mathbb{P}_{\theta^*}}{\to} 0, \tag{1.27}$$

where $\hat{\theta}$ is an efficient estimator, regular at θ^*, satisfying Definition 1.16.

Of note is that the centering of the asymptotic Gaussian is not at θ^*, but rather at the efficient estimator $\hat{\theta}$, and so centred credible sets of a given radius do not necessarily include θ^*, but rather have the appropriate frequentist coverage as confidence sets. The result also implies that Bayesian point estimators, such as the posterior mean, mode and median, are themselves efficient estimators.

The particular choice of efficient estimator $\hat{\theta}$ does not matter, as all such $\hat{\theta}$ agree up to neglible terms on the $\sqrt{n}-$scale, as discussed in the previous section and detailed in

Chapter 8 of [100]. It is perfectly valid to think of the $\hat{\theta}$ in Theorem 1.17 as the MLE in the regular parametric model $(f_\theta : \theta \in \Theta)$.

1.3.2.1 Estimation in parametric HMMs

Having introduced in sections 1.3.1 and 1.3.2 the notions of consistency, convergence rates, efficiency and Bernstein-von Mises, we now introduce the literature surrounding these notions in the parametric HMM context. The results introduced in this section are important prerequisites to the work on semiparametric HMMs covered in Chapter 2.

Theoretical properties of maximum likelihood estimation in parametric HMMs was initially considered in [9,86], who dealt with the case of the emission distributions being categorical random variables (i.e. supported on a finite number of points). In [68], the analysis was extended to general parametric HMMs, under mild regularity assumptions. Although consistency is weaker than asymptotic normality (and efficiency), proofs of those stronger results still require consistency first, essentially to ensure we are in a neighbourhood of the true parameter so that a Taylor expansion is valid.

Local asymptotic normality for parametric HMMs, along the lines of 1.13, was studied in [11]. Subsequently, [12] demonstrated asymptotic normality of the MLE. These results reference the Fisher Information matrix, which can be defined in the HMM context as follows. Let $p_n^{Q,\theta}(Y_1, \ldots, Y_n)$ be the density of $\mathbb{P}_n^{(Q, \mathbf{f}_\theta d\mu)}$, i.e. the likelihood of $Y_1, \ldots, Y_n$. Set $\ell_n^{Q,\theta} = \log p_n^{Q,\theta}$ the log-likelihood. Then it can be shown that $\frac{1}{\sqrt{n}} \nabla_{Q,\theta} \ell_n^{Q^*,\theta^*} \Rightarrow^* \mathcal{N}(0, I_*)$ for some positive semidefinite matrix I^*, under the assumption of Theorem 1.18 that Q^* is ergodic and the family $\{\mathbf{f}_\theta\}$ is 'regular' in a neighbourhood of θ^*. We call I_* the Fisher information matrix. For details of these regularity assumptions on the densities, we refer the reader to [12]. We state here the result of [12], which also requires these same regularity assumptions, in conjunction with other assumptions which we state here explicitly.

Theorem 1.18 (Bickel et al. 1998)**.** *Assume the transition matrix Q^* is ergodic and of full rank, the emission distributions admit μ-densities in the family $\{\mathbf{f}_\theta = (f_{\theta_1}, \ldots, f_{\theta_R}) : \theta \in \Theta\}$ indexed by a finite-dimensional parameter space Θ, and that the model is identifiable. Assume also that the MLE is consistent, and that the Fisher Information matrix I^* is invertible. Let $\theta^* \in \mathrm{Int}(\Theta)$ and assume the family of densities $\{\mathbf{f}_\theta\}$ is 'regular' uniformly over θ in a small neighbourhood of θ^*. Then the MLE $(\hat{Q}, \hat{\theta})$ satisfies*

$$\sqrt{n}((\hat{Q}, \hat{\theta}) - (Q^*, \theta^*)) \Rightarrow^* \mathcal{N}(0, I_*^{-1}). \tag{1.28}$$

In [32], this analysis is extended to autoregressive models with Markov regime, which generalises the class of parametric HMMs to a larger parametric family. Though this larger family is not of direct interest for our study of nonparametric HMMs, we adopt much of the notation of that work, and use several of their technical lemmas, when we study semiparametric efficiency in Chapter 2.

Regarding the Bernstein-von Mises phenomenon of Section 1.3.2, the adaptation to the parametric HMM context (under similar assumptions to the asymptotic normality result of [12]) was studied in [28]. We state their result here.

Theorem 1.19 (De Gunst & Shcherbakova, 2008)**.** *Grant the assumptions of Theorem 1.18. Let $\ell_n^{(Q,\theta)}$ be the log-likelihood and*

$$T_n = (Q^*, \theta^*) + \frac{1}{n} I_*^{-1} \nabla_{Q,\theta} \ell_n^{Q^*,\theta^*}. \tag{1.29}$$

Assume the prior Π admits a continuous (Lebesgue) density π for which $\pi(Q^*, \theta^*) > 0$. Then

$$\|\Pi(\sqrt{n}((Q,\theta) - T_n) \in \cdot | Y_1, \ldots, Y_n) - \mathcal{N}_d(0, I_{\theta^*}^{-1})\|_{TV} \overset{\mathbb{P}_{Q^*,\theta^*}}{\to} 0, \tag{1.30}$$

for $d = R(R-1) + \dim \Theta$.

Similarly to what was discussed in Sections 1.3.1 and 1.3.2, all efficient estimators agree up to negligible terms on the $\sqrt{n}-$scale, and so T_n represents the 'canonical' efficient estimator which may be equivalently replaced by the MLE.

1.3.3 Efficiency in semiparametric models

The problem of semiparametric estimation may be roughly formulated as:

Is it possible to estimate the finite-dimensional aspect of an infinite-dimensional model, in a way which resembles estimation in a parametric (finite-dimensional) model?

Suppose the model takes the form $\mathcal{P}_n = \{\mathbb{P}_n^{(\theta)} : \theta \in \Theta\}$, where Θ is infinite-dimensional. Then if $\psi : \Theta \longrightarrow \mathbb{R}^d$ be a $d-$dimensional functional, semiparametric estimation seeks to estimate $\psi(\theta)$ with guarantees similar to estimation of parametric models. Let θ^* be the true parameter and $\psi^* = \psi(\theta^*)$. Then semiparametric theory seeks to prove analogues of results like those discussed in Sections 1.3.1 and 1.3.2, but with $\hat{\theta}$ replaced by $\hat{\psi}$, an estimator of the functional $\psi(\theta)$.

Such results are non-trivial, and are an area of active research; we elaborate on this further in Chapter 2. Hence, our focus for this chapter is not on proving the existence of efficient estimators or Bernstein-von Mises type results, but rather on understanding the notion of efficiency in this context.

When the parameter $\theta \in \Theta$ splits naturally into a tuple (ϑ, η), where $\vartheta \in \mathbb{R}^d$ for some $d < \infty$, the functional ψ of interest is often $\psi(\theta) = \vartheta$. This happens for instance in mixture models, where weights vary in the simplex but the emissions may vary in some infinite-dimensional family. Such models, though infinite-dimensional and thus nonparametric, are often referred to as being *semiparametric*. In the remainder of this thesis, we concentrate specifically with semiparametric estimation in such models, for the case $\psi(\theta) = \vartheta$. The material of this section is covered in Chapter 25 of Section [100].

Inference with multiple finite-dimensional parameters To set the scene, consider first the case that we have a parametric model with densities $(f_{\vartheta,\eta} : \vartheta \in \mathbb{R}^{d_1}, \eta \in \mathbb{R}^{d_2})$, so $\Theta = \mathbb{R}^{d_1+d_2}$ is finite dimensional. Then the theory of efficient estimators developed in Section 1.3.1 implies that, assuming I_{θ^*} is invertible, efficient estimators $\hat{\theta} = (\hat{\vartheta}, \hat{\eta})$ satisfy

$$\sqrt{n}(\hat{\theta} - \theta^*) = \frac{I_{\theta^*}^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(X_i) + o_{\mathbb{P}_*}(1). \tag{1.31}$$

Now write

$$I_{\theta^*} = \begin{pmatrix} I_{\vartheta^*} & I_{\vartheta^*,\eta^*} \\ I_{\vartheta^*,\eta^*}^T & I_{\eta^*} \end{pmatrix} \quad \dot{\ell}_{\theta^*} = \begin{pmatrix} \dot{\ell}_{\vartheta^*} \\ \dot{\ell}_{\eta^*} \end{pmatrix} \tag{1.32}$$

where I_{ϑ^*} is the upper left $d_1 \times d_1$ block, I_{η^*} is the lower right $d_2 \times d_2$ block, and I_{ϑ^*,η^*} is the upper right block. We can see that I_{ϑ^*} is the Fisher Information for the model $(f_{\vartheta,\eta^*} : \vartheta \in \mathbb{R}^{d_1})$. Now inverting Equation (1.32) we obtain a matrix of the form

$$I_{\theta^*}^{-1} = \begin{pmatrix} \tilde{I}_{\vartheta^*}^{-1} & -\tilde{I}_{\vartheta^*}^{-1} I_{\vartheta^*,\eta^*} I_{\eta^*,\eta^*}^{-1} \\ * & * \end{pmatrix} \tag{1.33}$$

where

$$\tilde{I}_{\vartheta^*} = I_{\vartheta^*} - I_{\vartheta^*,\eta^*} I_{\eta^*,\eta^*}^{-1} I_{\vartheta^*,\eta^*}^T, \tag{1.34}$$

the expression for which arises by considering the Schur complement of the block matrix I_{θ^*}. Since efficient estimators $\hat{\theta}$ of θ satisfy Equation (1.31), efficient asymptotic estimators

of ϑ within the model with unknown η satisfy

$$\sqrt{n}(\hat{\vartheta} - \vartheta^*) = \frac{\tilde{I}_\vartheta^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{(\dot{\ell}_{\vartheta^*} - I_{\vartheta^*,\eta^*} I_{\eta^*,\eta^*}^{-1} \dot{\ell}_{\eta^*})}_{\tilde{\ell}_{\vartheta^*}}(X_i) + o_{\mathbb{P}_*}(1), \tag{1.35}$$

since otherwise we would obtain a better covariance matrix in (1.31) by replacing the ϑ coordinates of $\hat{\theta}$ by the improved ϑ, which would yield a contradiction to the conclusion of Theorem 1.15.

We note that $\tilde{\ell}$ replaces the usual score ℓ and $\tilde{I}_{\vartheta^*}$ now replaces the Fisher information I_θ in (1.35) as the matrix whose inverse is the optimal variance (on the $\sqrt{n}$ scale); we also have $\mathbb{E}_{\theta^*} \tilde{\ell}_{\vartheta^*} \tilde{\ell}_{\vartheta^*}^T = \tilde{I}_{\vartheta^*}$. Hence, we refer to $\tilde{I}_{\vartheta^*}$ as the *efficient Fisher information matrix* (or simply the *efficient information*), and $\tilde{\ell}_{\vartheta^*}$ as the *efficient score*.

The definitions are relative to both the parameter of interest ϑ and the additional unknown parameter η in the model, which we refer to as the *nuisance parameter*. The difference between the optimal variances in the $\eta = \eta^*$ fixed and the η unknown cases can be understood as the negative impact of not knowing η on the estimation of ϑ.

We make one further remark regarding Equation (1.35). We have

$$\mathbb{E}_\theta \tilde{\ell}_\vartheta \dot{\ell}_\eta^T = I_{\vartheta,\eta} - I_{\vartheta,\eta} I_{\eta,\eta}^{-1} I_{\eta,\eta} = 0, \tag{1.36}$$

and so the functions in the vector $\tilde{\ell}_\vartheta$ are orthogonal to the functions in the vector ℓ_η, in the $L^2(\mathbb{P}_\theta)$ sense. In fact, from the definition of $\tilde{\ell}_\theta$ in Equation (1.35), we can see the $\tilde{\ell}_\theta$ function as the orthogonal projection of each component onto the complement of the space spanned by $\dot{\ell}_\eta$. In this way, we see that part of the score function has been 'lost' when η is unknown, and so likelihood-based inference involving the score becomes more difficult.

One-dimensional submodels for η Suppose we have a model with densities $(f_{\vartheta,\eta} : \vartheta \in \mathcal{T}, \eta \in \mathcal{F})$, so $\Theta = \mathcal{T} \times \mathcal{F}$ with $\mathcal{T} \subset \mathbb{R}^d$. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} f_{\vartheta^*,\eta^*}$. Inference in the parameter space Θ is at least as difficult as inference in the parameter space $\bar{\Theta} \subset \Theta$ where $(\vartheta^*, \eta^*) \in \bar{\Theta}$, since all we have done is excluded some incorrect parameter configurations.

If we take $\bar{\Theta}$ to be finite-dimensional, we understand from the preceding discussion the optimal variance of efficient estimators in this model, and we know that the corresponding $\tilde{\ell}_\vartheta$ is the projection of $\dot{\ell}_\theta$ onto the complement of spaces spanned by the scores on η. The

idea is that the semiparametric efficient score must live in the orthogonal complement of the space spanned by *all* such scores on η, arising from different parametric submodels.

According to Section 25.3 of [100], it typically suffices to consider one-dimensional submodels where we vary the the nuisance parameter η along a path $t \mapsto \eta_t$ with $\eta_t = \eta^*$ at $t = 0$, of the form

$$\bar{\Theta}_{\mathcal{T}, t \mapsto \eta_t} = \{(\vartheta, \eta_t) : \vartheta \in \mathcal{T}, t \in (-\epsilon, \epsilon)\}. \tag{1.37}$$

We only require t to vary in an arbitrarily small neighbourhood of zero, since we are interested in the derivative at $t = 0$. Although $\bar{\Theta}$ is of dimension $d_1 + 1$ (since ϑ is also unknown), we are only really interested in $\bar{\Theta}$ in order to understand the span of the nuisance score arising from differentiation with respect to t. For this purpose, we can consider one-dimensional models purely on the nuisance parameter

$$\bar{\Theta}_{\vartheta^*, t \mapsto \eta_t} = \{(\vartheta^*, \eta_t) : t \in (-\epsilon, \epsilon)\}, \tag{1.38}$$

since the $\dot{\ell}_\vartheta$ for the model would just be the same in every case and can be separately evaluated. By varying the path $t \mapsto \eta_t$, we get different submodels $\bar{\Theta}_{\vartheta^*, t \mapsto \eta_t}$ with different nuisance scores. We write $\mathcal{S}$ be the space in $L^2(\mathbb{P}^*)$ spanned by these scores, and write P for the orthogonal projection onto this space. This finally leads us to define the *efficient score* function,

$$\tilde{\ell}_\vartheta = \dot{\ell}_\vartheta - P\dot{\ell}_\vartheta, \tag{1.39}$$

the *efficient Fisher information matrix*,

$$\tilde{I}_\vartheta = \mathbb{E}_{\vartheta, \eta^*} \tilde{\ell}_\vartheta \tilde{\ell}_\vartheta^T, \tag{1.40}$$

and (assuming non-singularity of the information matrix) the *efficient influence function*,

$$\tilde{\psi}_\vartheta = I_\vartheta^{-1} \tilde{\ell}_\vartheta. \tag{1.41}$$

Theorem 25.20 of [100] gives a semiparametric analogue of Theorem 1.15, the convolution theorem. Moreover, Lemma 25.23 of [100] yields that estimators $\hat{\vartheta}$ of ϑ^* with limiting

distribution $\mathcal{N}_{d_1}(0, I_{\vartheta^*}^{-1})$ take the familiar form

$$\sqrt{n}(\hat{\vartheta} - \vartheta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}(X_i) + o_{P_{\vartheta^*,\eta^*}}(1). \tag{1.42}$$

Are one-dimensional submodels sufficient? It is not immediately clear that one-dimensional models should indeed be sufficient. If we think of the information matrix we construct in this way as being relative to the choice of submodels, and if the class of one-dimensional models is insufficiently rich, them we may have an overly optimistic view on the true efficient variance (as would happen if, for instance, we only considered submodels where η is known). However, if we can then exhibit regular, asymptotically normal, estimators within the semiparametric model at hand, which attain this efficient variance, then we know that this variance, and these estimators, are indeed optimal. We can think of this as having a duality between variance of regular estimators (lower bounded by the optimal semiparametric variance) and inverse efficiency matrices relative to some class of submodels (upper bounded by the optimal semiparametric variance), and when two feasible solutions coincide, they must be optimal.

To provide some intuition as to why one-dimensional models should indeed suffice, consider the previous expression for $\tilde{\ell}_\theta$, where $\eta \in \mathbb{R}^{d_2}$

$$\tilde{\ell}_\theta = \dot{\ell}_\theta - I_{\theta,\eta} I_{\eta,\eta}^{-1} \dot{\ell}_\eta, \tag{1.43}$$

we can trivially rewrite

$$(I_{\theta,\eta} I_{\eta,\eta}^{-1} \dot{\ell}_\eta)_i = \sum_{j=1}^{d_2} \sum_{k=1}^{d_2} (I_{\theta,\eta})_{ik} (I_{\eta,\eta}^{-1})_{kj} \dot{\ell}_{j,\eta}. \tag{1.44}$$

Now if we can take d_2 one-dimensional submodels on the nuisance parameter η, for which the nuisance scores are an orthogonal set with equal span to the $\{\dot{\ell}_{j,\eta}\}_{j \leq d_2}$, then the terms $(I_{\eta,\eta}^{-1})_{kj} = 0$ for $k \neq j$. Then

$$(I_{\theta,\eta} I_{\eta,\eta}^{-1} \dot{\ell}_\eta)_i = \sum_{j=1}^{d_2} (I_{\theta,\eta})_{ij} (I_{\eta,\eta}^{-1})_{jj} \dot{\ell}_{j,\eta}, \tag{1.45}$$

from which it is clear that the coordinates $\tilde{\ell}_{\theta,i}$ of (1.43) can be obtained by sequentially projecting within one-dimensional submodels, which is equivalent to asking that we are in the orthogonal complement of the space spanned by all one-dimensional nuisance scores.

Adaptation to non-iid models The focus of what precedes is on models for iid data; importantly for this thesis, is the application of convolution theorems in non-iid settings, as laid out in Section 2 of [78]. We defer this technical discussion to Section 2.12.2 of Chapter 2

1.4 Cut Posterior

The *cut posterior* is a modification to the Bayesian procedure which allows us to infer two parameters (ϑ, η) of a Bayesian model jointly, in a way which modifies the usual flow of information between prior and posterior. The idea is that, in certain instances, we may wish to learn ϑ in a first 'module' which we believe to provide better inference for that parameter, before imputing into a second module from which we wish to learn η, but for which we still require a value of (or distribution over) ϑ.

A basic way to do this might be to estimate a value $\hat{\vartheta}$ in the first module, and then fix this value as the true, known parameter in the second module. Then our data in the latter module will certainly not effect our estimate for $\hat{\vartheta}$, but we are at a disadvantage as we have not propagated forward any uncertainty in ϑ, and unless $\hat{\vartheta}$ coincides exactly with the true parameter we are necessarily misspecified.

Instead, we wish to construct a joint distribution over (ϑ, η) which allows us to propagate uncertainty in ϑ into the η module, while *cutting* the flow of information back into the ϑ parameter. This notion of cutting the flow motivates the term *cut posterior*.

An early example of the use of such modular approaches is in the study of pharmacokinetics-pharmacodynamics models [10], with the notion of cuts also introduced in the WinBUGS software [74], see also [87] for an analysis of the associated MCMC scheme. For a subsequent discussion of cut posterior approaches, see [61]. For recent advances in modular inference, especially relating to generalisations of cut posterior and variational inference, we direct the reader to [19, 20, 71, 108].

In Section 1.4.1 we introduce mathematically the cut posterior in the classical setting of two data sources. In Section 1.4.2, we see that the procedure can be easily adapted to instead use two differing prior schemes. It is the latter notion which will play a role in the work of Chapter 2.

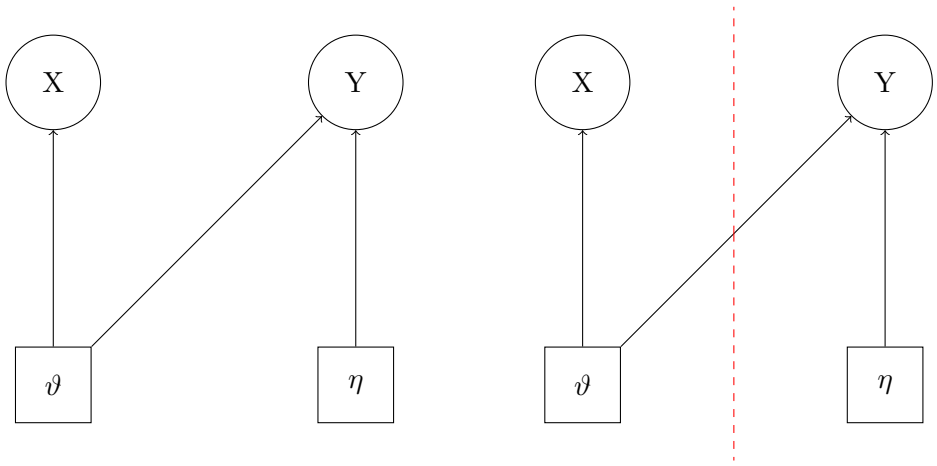


Figure 1.2: Bayesian model with parameters ϑ and η. The left-hand graph represents the usual Bayesian posterior $\pi(\cdot|X, Y)$, with information flowing back from Y to ϑ.
On the right-hand graph we cut the feedback from Y to ϑ so information cannot flow backward from the second module, representing the cut posterior $\pi_{\text{cut}}(\cdot|X, Y)$

1.4.1 Multiple data sources

Suppose we have a model built of two modules, which each model different datasets but share a common parameter. For the first module, our model $\mathcal{P}_n = \{\mathbb{P}_n^{(\vartheta)} : \vartheta \in \mathcal{T}\}$ for data $X_{1:n}$. We do not assume the model is iid, but for simplicity we assume the joint distributions $\mathbb{P}_n^{(\vartheta)}$ have (Lebesgue) densities f_n^{ϑ}. Suppose also we have a second model $\mathcal{Q}_n = \{\mathbb{Q}_n^{(\vartheta,\eta)} : \theta \in \mathcal{T}, \eta \in \mathcal{F}\}$ for data $Y_{1:n}$, and here the distributions $\mathbb{Q}_n^{(\vartheta,\eta)}$ have densities $g_n^{\vartheta,\eta}$. We assume the parameter ϑ is shared across the two modules. Conditionally on the parameters, we assume the two datasets are independent of one another.

Let Π be a prior over Θ with density $\pi(\vartheta, \eta) = \pi(\vartheta)\pi(\eta)$ factorising over the parameters. Then the usual Bayes posterior for the parameters given the data has density

$$\begin{aligned}\pi(\vartheta, \eta|X_{1:n}, Y_{1:n}) &\propto \pi(\vartheta)\pi(\eta)f_n^{\vartheta}(X_{1:n})g_n^{\vartheta,\eta}(Y_{1:n}) \\ &\propto \pi(\vartheta|X_{1:n})\pi(\eta)g_n^{\vartheta,\eta}(Y_{1:n}) \\ &\propto \pi(\vartheta|X_{1:n})\pi(\eta|Y_{1:n}, \vartheta)p(Y_{1:n}|\vartheta)\end{aligned} \tag{1.46}$$

Between the second and third lines, the normalising constant for the distribution $\pi(\vartheta|X_{1:n})$ is absorbed into the proportionality as there is no dependence on the parameter. We then go from the second to the third line by noting that $\pi(\eta)g_n^{\vartheta,\eta}(Y_{1:n})$ is proportional to the distribution $\pi(\eta|Y_{1:n}, \vartheta)$, which is the posterior in the second module where we impute ϑ, but the normalising constant now depends on ϑ and so we must track the extra term

$p(Y_{1:n}|\vartheta)$.

For the cut posterior, we delete this term in Equation (1.46) and define

$$\pi_{\text{cut}}(\vartheta, \eta|X_{1:n}, Y_{1:n}) \propto \pi(\vartheta|X_{1:n})\pi(\eta|Y_{1:n}, \vartheta), \tag{1.47}$$

which still gives a valid distribution over the parameters ϑ and η. The term $p(Y_{1:n}|\vartheta)$ can be thought of as the feedback from the data $Y_{1:n}$ on the parameter ϑ, and this flow of information is being cut so that we only learn ϑ through $X_{1:n}$. We summarise this graphically in Figure 1.4.

1.4.2 Multiple prior models

In Section 1.4.1, we considered the situation where we learn ϑ and η from separate data sources, by constructing a distribution over (ϑ, η) where the distribution over ϑ is estimated purely from the first data source in the $\vartheta-$module, and this distribution is 'plugged in' to the η-module without using that second data source to further inform ϑ.

This idea can be adapted quite naturally to using two separate prior distributions, one for each module. In the setup of this section, we imagine a common data source Y with parameters ϑ and η. To learn ϑ, we first place a prior $\pi_1(\vartheta, \eta)$ over ϑ and η, and infer ϑ via the marginal posterior $\pi_1(\vartheta|Y) = \int_{\eta\in\mathcal{F}} \pi_1(\vartheta, d\eta|Y)$. Following this, we propogate this ϑ into the second module to learn η, based on the data Y and the conditional prior $\pi_2(\eta|\vartheta)$.

In fact, since the two η appearing in the π_1 and π_2 are distinct random variables, we should distinguish them as η' and η, with η' being the version distributed according to π_1, and η as being distributed according to π_2 (and hence the parameter of interest). This process yields a joint over all three random variables

$$\pi'_{\text{cut}}(\eta', \vartheta, \eta) = \pi_1(\eta', \vartheta|Y)\pi_2(\eta|Y, \vartheta). \tag{1.48}$$

Taking the (ϑ, η) marginal in Equation (1.48) yields the cut distribution in this context

$$\pi_{\text{cut}}(\vartheta, \eta) = \pi_1(\vartheta|Y)\pi_2(\eta|Y, \vartheta), \tag{1.49}$$

whose form resembles that of Equation (1.47). Note however that, in the case that $\pi_1(\eta'|\vartheta) = \pi_2(\eta)$; in this sense there is no 'cut' in the model in the before sense. We

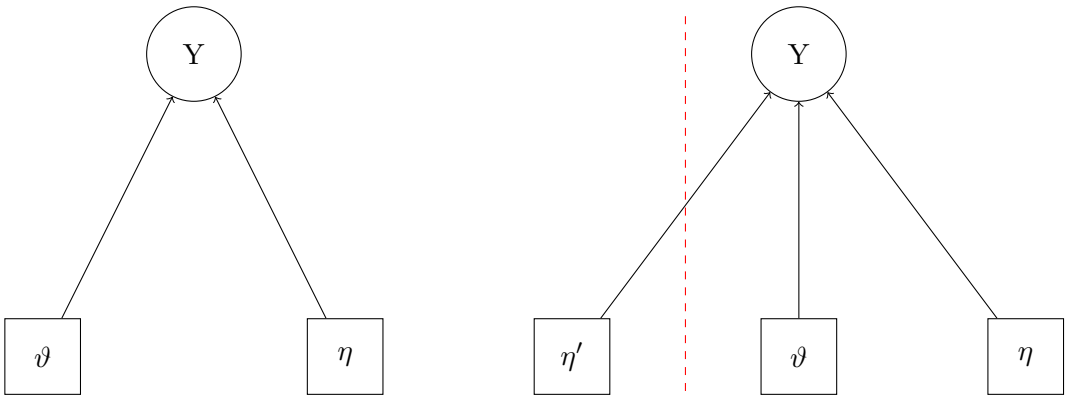


Figure 1.3: Comparison between Bayes posterior (left) and cut posterior (right) in the multiple prior case with a single data source. We use the η' variable with a joint prior scheme $\pi_1(\eta', \vartheta)$ for learning ϑ; this is then propagated to learn η via a separate prior on η alone. The red dashed line in this context refers to the fact that the η' variable is discarded.

nevertheless adopt the same terminology here.

1.5 Concentration Rates

Concentration rates provide an asymptotic guarantee on the rate at which the posterior concentrates around the true parameter θ^*, similar in spirit to the notion of rate of convergence of estimators as in Equation (1.17). Indeed, the Bernstein-von Mises Theorem 1.17 in the parametric setting is a particularly strong kind of concentration result, as it guarantees both the regular parametric $\sqrt{n}$-rate of convergence, while also describing the shape of the limiting distribution. As with the Bernstein-von Mises theorem, the study of posterior concentration rates seeks to provide a frequentist guarantee on a Bayesian method.

As mentioned in Section 1.3.3, proving Bernstein-von Mises like results in semiparametric settings is an active area of research; such results also exist in nonparametric contexts (for instance for the Dirichlet process [89] and the Pitman-Yor process [37]), with limiting distributions being Gaussian processes rather than finite-dimensional Gaussians. Regardless, these results are typically difficult to obtain, or limited in their scope, and so slightly weaker results which quantify the rate of convergence of the posterior, even if not the limiting shape, are themselves of interest.

In Section 1.5.1, we give an overview of the now-classical framework introduced by Ghosh, Ghosal and van der Vaart [47, 48]; the primary reference for this section is [49]. In Section 1.5.2, we discuss a further step required to prove posterior concentration in mixtures and

HMMs.

1.5.1 Framework for concentration in Bayesian nonparametrics

The material of the following section is covered in Chapter 8 of [49]. We begin by defining the notion of posterior concentration (also known as posterior contraction).

Definition 1.20 (Posterior concentration)**.** Consider a Bayesian model with parameter space Θ, data X and posterior $\Pi(d\theta|X)$. Assume $X \sim \mathbb{P}_{\theta^*}$ for some frequentist true value of θ^*. we say that the posterior *concentrates at rate* $\epsilon_n \to 0$, in the metric d, at the point θ^*, if for all sequences $M_n \to \infty$ we have

$$\Pi(\{\theta : d(\theta, \theta^*) > M_n \epsilon_n\}) \overset{\mathbb{P}_{\theta^*}}{\to} 0. \tag{1.50}$$

When Definition 1.20 holds for some sequence $\epsilon_n \to 0$, we refer to ϵ_n as a concentration rate. If ϵ_n is optimal in the sense that $\epsilon_n \lesssim \tilde{\epsilon}_n$ for all concentration rates $\tilde{\epsilon}_n$, we often refer to ϵ_n as *the* concentration rate.

Example 1.21 (Parametric models)**.** Let the assumptions of the Bernstein-von Mises Theorem 1.17 hold. Then the posterior concentrates at rate $\epsilon_n = \frac{1}{\sqrt{n}}$ in the usual Euclidean metric. To see this, we note that $\sqrt{n}(\hat{\theta} - \theta^*) = O_{\mathbb{P}_{\theta^*}}(1) = o_{\mathbb{P}_{\theta^*}}(M_n)$ for all sequences $M_n \to \infty$, and also that $\Pr(\|Z\| > M_n) \to 0$ for any fixed Gaussian random variable Z.

Before stating the concentration result, we introduce the notion of Kullback-Leibler neighbourhoods, and metric entropy, both used in the proof.

Definition 1.22 (Kullback-Leibler diverge/variation)**.** Given densities $p_1, p_2 \in \mathcal{P}$ on a set $\mathcal{X}$, with respect to a common dominating measure ν, the *Kullback-Leibler divergence* between the densities is

$$\mathcal{K}(p_1; p_2) = \mathbb{E}_{X \sim p_1} \log \frac{p_1(X)}{p_2(X)} = \int_{\mathcal{X}} \log \frac{p_1(x)}{p_2(x)} p_1(x) d\nu(x). \tag{1.51}$$

Formally the KL is defined this way only when p_1 is absolutely continuous with respect to p_2, otherwise it is equal to infinity. Similarly, we define the order k *Kullback-Leibler variation* as

$$V_k(p_1; p_2) = \mathbb{E}_{X \sim p_1} \left| \log \frac{p_1(X)}{p_2(X)} - \mathcal{K}(p_1; p_2) \right|^k. \tag{1.52}$$

We can then define the following neighborhoods used in the proof:

$$\begin{aligned} B_0(p,\epsilon) &= \{p' : \mathcal{K}(p;p') < \epsilon^2\}. \\ B_k(p,\epsilon) &= \{p' : \mathcal{K}(p;p') < \epsilon^2, V_k(p;p') < \epsilon^k\} \quad (k>0). \end{aligned} \tag{1.53}$$

Secondly, we introduce the notion of the covering number of a (parameter) set $\mathcal{P}$.

Definition 1.23 (Covering Number)**.** The *ϵ-covering number* of a (compact) set $\mathcal{P}$, with respect to a metric d, denoted $N(\mathcal{P}, d, \epsilon)$, is defined as

$$N(\mathcal{P}, d, \epsilon) = \inf\{N \in \mathbb{N} : \mathcal{P} \subset \bigcup_{i=1}^{N} B_d(p_i, \epsilon), p_i \in \mathcal{P}\} \tag{1.54}$$

where $B_d(p_i, \epsilon) = \{p \in \mathcal{P} : d(p, p_i) \leq \epsilon\}$ is a ball in the metric d with centre p_i and radius ϵ. In other words, $N(\mathcal{P}, d, \epsilon)$ is the minimal number of $\epsilon-$balls required to cover $\mathcal{P}$.

We now present the basic contraction rate result for iid observations, Theorem 8.9 of [49].

Theorem 1.24 (Basic contraction, Theorem 8.9 of [49])**.** *Let $\mathcal{P}$ be a set of densities over $\mathcal{X}$ with respect to ν, and let Π_n be a (sequence of) prior distribution(s) over $\mathcal{P}$. Consider the Bayesian model with $p \sim \Pi$, $X_1, \ldots, X_n | p \overset{iid}{\sim} p$, and posterior $\Pi_n(p \in \cdot | X_{1:n})$, and let p^* be the frequentist true parameter. Suppose there exists a sequence of subsets $\mathcal{P}_n \subset \mathcal{P}$, and a constant $C > 0$, such that for sequences $\bar{\epsilon}_n \leq \epsilon_n$ with $n\bar{\epsilon}_n^2 \to \infty$, we have*

$$(i) \quad \Pi_n(B_2(p^*, \bar{\epsilon}_n)) \geq e^{-Cn\bar{\epsilon}_n^2}, \tag{1.55}$$

$$(ii) \quad \log N(\mathcal{P}_n, d, \frac{\epsilon_n}{2}) \leq n\epsilon_n^2, \tag{1.56}$$

$$(iii) \quad \Pi_n(\mathcal{P} \setminus \mathcal{P}_n) \leq e^{-(C+4)n\bar{\epsilon}_n^2}. \tag{1.57}$$

Then ϵ_n is a posterior rate of concentration at p^*, in the metric d.

A number of variations of Theorem 1.24 are presented in [49], for instance to avoid the use of conditions on the second-order variation neighbourhoods B_2.

The condition (ii) in Theorem 1.24 is used to construct tests of the null $p = p^*$ against complements of shrinking balls of radius ϵ_n; we can instead simply make the testing assumption as is done so in Theorem 8.12 of [49], and in some sense this assumption is 'minimal' as shown in Theorem 8.13 therein. These tests are typically constructed via covering arguments however, hence the original condition.

Finally, the theorem is straightforwardly adaptable to the non-iid context, with the neigh-

bourhoods B_0, B_k in the conditions of the theorem being replaced by analogous neighbourhoods for the joint densities $p_n(X_1, \ldots, X_n)$, as set out in Theorem 8.19 of [49]. That said, these conditions may take considerably more effort to validate in such contexts, with the HMM case being considered in [102]. We adapt the general result along the lines of the framework of [49] in our general concentration result presented in Chapter 2 for the HMM context, although the validation of those assumptions in terms of assumptions on the nonparametric emission densities relying heavily on prior work of [102].

1.5.2 Inversion Inequalities

In Theorem 1.24, and in subsequent refinements such as [102] for HMMs, we obtain the rate of convergence for a metric placed over the density of the observations. For instance, in the parametric mixture model setting, our marginal densities might take the form $p_{\pi,\theta} = \sum_{r=1}^{R} \pi_r f_r^{(\theta)}$, even if our parameter takes the form (π, θ). If we could apply Theorem 1.24, we would get a result of the form

$$\Pi_n(\{(\pi,\theta) : d(p_{\pi,\theta}, \underbrace{p_{\pi^*,\theta^*}}_{p^*}) \geq M_n \epsilon_n\}) \xrightarrow{p^*} 0. \tag{1.58}$$

Here d would be one of the aforementioned metrics on densities. In practice, we might be interested in understanding the posterior distribution over the parameters themselves, preferring instead a result of the form

$$\Pi_n(\{(\pi,\theta) : \tilde{d}((\pi,\theta),(\pi^*,\theta^*)) \geq M_n \epsilon_n\}) \xrightarrow{p^*} 0, \tag{1.59}$$

with $\tilde{d}$ now a metric on the parameters. To go from Equation (1.58) to Equation (1.59), we could seek a result of the form

$$d(p_{\pi,\theta}, p_{\pi^*,\theta^*}) \geq C^* \tilde{d}((\pi,\theta),(\pi^*,\theta^*)), \tag{1.60}$$

which should hold in some small (but fixed) neighbourhood of (π^*, θ^*). Results of the form (1.60) are called *inversion inequalities*. They are in some sense a strengthening of the notion of identifiability; not only should the parameter map $(\pi, \theta) \mapsto p_{\pi,\theta}$ be injective[3] and thus invertible, but this inverse should have some sort of local continuity (or Lipschitz-like property) with respect to d and $\tilde{d}$.

[3] As mentioned before, we ignore non-identifiability arising from label-swapping.

Such results are themselves non-trivial, and are of independent interest outside of the framework of concentration rates (as they would, for instance, apply equally well in the context of purely frequentist estimation). In [26], the authors prove such a result for the L^2 distance in the context of HMMs, and one of the contributions of Chapter 2 is such an inequality, also for HMMs, for the L^1 distance.

1.6 Bayesian Model Selection

Given a collection of N Bayesian models $\mathcal{P}_n^{(i)}$ with (disjoint) parameter spaces Θ_i and priors Π_i, we may wish to assess which model is most suitable for our data. Typically the models will differ in some parameter which affects the dimension of Θ_i, and so it is conceptually easier to think of N distinct models rather than a single model $\mathcal{P}_n = \bigcup_{i=1}^N \mathcal{P}_n^{(i)}$. In the context of mixtures and HMMs, the index i will refer to the number of hidden states, and the problem at hand is to decide how many states should be used in the model to represent our data.

When taking a Bayesian approach, we ultimately need to construct a prior over the full parameter space $\bigcup_{i=1}^N \Theta_i$ for the full model $\mathcal{P}_n = \bigcup_{i=1}^N \mathcal{P}_n^{(i)}$. We will assume for simplicity that $N = 2$, and we construct our prior in a hierarchical way to account for the differences between the models (e.g. for different dimensions). We set $\rho = \pi(\theta \in \Theta_1)$, so $1 - \rho = \pi(\theta \in \Theta_2)$. Let π_1, π_2 denote the respective priors over Θ_1, Θ_2, that is $\pi_i = \pi(\cdot | \theta \in \Theta_i)$. If the data X has likelihood f_θ^i (which is a density with respect to some measure ν) when $\theta \in \Theta_i$, then

$$\begin{aligned} \pi(\theta \in \Theta_i | X) &= \frac{\pi(\theta \in \Theta_i) m(X | \theta \in \Theta_i)}{m(X)} \\ &= \frac{\pi(\theta \in \Theta_i)}{m(X)} \int_{\theta \in \Theta_i} f_\theta^i(X) \nu(dx), \end{aligned} \tag{1.61}$$

where $m(X) = \rho m(X | \theta \in \Theta_1) + (1 - \rho) m(X | \theta \in \Theta_2)$. Taking the quotient of Equation (1.61) for the $i = 1$ and $i = 2$ case we get

$$\frac{\pi(\theta \in \Theta_1 | X)}{\pi(\theta \in \Theta_2 | X)} = \underbrace{\frac{\pi(\theta \in \Theta_1)}{\pi(\theta \in \Theta_2)}}_{\rho / 1 - \rho} \frac{m(X | \theta \in \Theta_1)}{m(X | \theta \in \Theta_2)}. \tag{1.62}$$

Looking at the posterior odds in Equation (1.62), the factor $\dfrac{\pi(\theta \in \Theta_1)}{\pi(\theta \in \Theta_2)}$ corresponds to the

prior odds, while the factor $\frac{m(X|\theta \in \Theta_1)}{m(X|\theta \in \Theta_2)}$ provides the update to the Bayesian's beliefs about the odds. This quantity is known as the *Bayes factor*, and is the object of study when seeking to understand Bayesian model selection, as we do so in Chapter 3.

1.7 Bayesian Analysis of Conflict Risk

In Chapter 4, we focus on the use of discrete time Hawkes processes as a means to model conflict. In Section 1.7.1, we provide an overview of these models. In Section 1.7.2, we introduce the particular topic of our application, namely conflict data.

1.7.1 Discrete-time Hawkes processes

First introduced in [53,54], Hawkes processes are self-exciting point processes, in which the intensity at a given time varies according to the events which precede it; for a review paper see [65]. As such, they have enjoyed use in a variety of contexts in which activity tends to evoke reaction, including seismology [83], finance [55], neurology [13,46], epidemiology [17,24], and social harm [81].

Univariate, discrete-time, Hawkes processes were proposed as means to model conflict in [16], and also for other applications, see for instance [17,80]. These processes are relevant when data is not received in a continuous way, but when the notions of self-excitation are nevertheless still relevant.

Mathematically, a univariate discrete time Hawkes process is a data-generating process for a discrete time series $Z_1, Z_2, \dots$, with $Z_i \in \{1, 2, \dots\}$, which satisfies

$$Z_t | Z_1, Z_2, \dots Z_{t-1} \sim \text{Po}(\lambda_t) \tag{1.63}$$

where

$$\lambda_t = \mu_t + \alpha \sum_{t'<t} Z_{t'} g(t - t'). \tag{1.64}$$

Here μ_t is a deterministic rate at which events occur sporadically, and α controls the average number of 'offspring' from an event (viewing Hawkes processes as a superposition of births caused by different events) and g is an excitation kernel with the property that $\sum_{i=1}^{\infty} g(i) = 1$, which ensures identifiability of α.

Hawkes processes, and their discretised versions as defined by (1.63) and (1.64), can be naturally extended to a multivariate setting where data is of k different types - these

may, for instance, be geographic regions as we consider in Chapter 4. This involves modelling jointly the observations $\{S_t^{(j)}\}_{t\leq T}^{j\leq k}$, corresponding to k distinct categories, by the relationship

$$S_t^{(j)} | \{S_{t'}^{(i)}\}_{t'<t}^{1\leq i\leq k} \sim \text{Po}(\lambda_t^{(j)}), \tag{1.65}$$

where the rates $\lambda_t^{(j)}$ are given as

$$\begin{aligned} \lambda_t^{(j)} &= \mu_j + \sum_{i=1}^{k} \lambda_{t,i}^{(j)}; \qquad\qquad 1 \leq j \leq k \\ \lambda_{t,i}^{(j)} &= \alpha_{ij} \sum_{t'<t} S_{t',i} g_{\beta_{ij}}(t-t'). \end{aligned} \tag{1.66}$$

We interpret the rate $\lambda_t^{(j)}$ as the intensity of events at time t of category j, with those events occurring sporadically at rate μ_j, and as a result of other events of category i at rate $\lambda_{t,i}^{(j)}$. In inferring such models, we seek also to uncover the latent structure encoded by the parameters $(\alpha_{ij})_{i,j\leq k}$, which can be thought of as the (weighted) edges of a graph of interactions between categories.

1.7.2 Statistical Modelling of Conflict and the ACLED Dataset

Statistical modelling of the occurrence of conflict events dates back to [90], who demonstrated the power-law nature of the scale of conflicts. More recently, [25] reexamined this notion in the context of terrorism. In [15], nonlinear spatial models were employed to understand the geographic contagion of civil unrest.

Understanding such events is of key importance for both the short-term assessment of risk, and the identification of longer-term trends. Such information is of interest to various actors, including both governmental organisations and nonprofits, and can inform policy and crisis response.

One such actor is the *Armed Conflict Location & Event Data Project* (ACLED), who curate a global collection of conflict data and produce corresponding analysis, with both aspects being open to the public. The data is provided at the daily level, with details about both the type of event (e.g. Battles, Riots) and the location (e.g. administrative district), among various other features.

The rich nature of this data, and the dependencies therein, motivates the use of statistical tools capable of modelling the evolution of conflict over space and time. Very recently, [16]

considered extensions of discrete-time Hawkes processes of the kind set out in Section 1.7.1, using a spatial kernel to incorporate spatiotemporal aspects of excitation, as a means to model this particular dataset. We discuss in the Section 1.8 the extensions of this work which we present in Chapter 4.

1.8 Summary of Contributions

In Chapter 2, we present the paper *Efficient Bayesian Estimation and use of cut posterior in semiparametric Hidden Markov models.*

In Section 2.3, we exhibit efficient estimators, in the semiparametric sense, of the transition matrix Q when the emissions are modelled nonparametrically as densities. To complement this result, we present also an analogous semiparametric Bernstein-von Mises result. The work of this section relates to the introductory material of Section 1.3, and is an extension of the work in multidimensional mixtures of [42]. Although we use the same ideas in our modelling approach, the proofs in the HMM context are considerably more technically involved, and do not follow straightforwardly from the mixtures settings, especially in order to prove efficiency.

In Section 2.4, we study the use of the cut posterior introduced in Section 1.4, in the sense of multiple prior models described in Section 1.4.2, to jointly infer the transition matrix Q and the densities $\mathbf{f}$. This is motivated by the requirement that we needed to choose a strongly undersmoothing prior Π_1 on $\mathbf{f}$, to obtain the good control over Q described in the previous paragraph. By using a second prior Π_2 to model the densities $\mathbf{f}$ given the value of Q, we achieve a 'best of both worlds' distribution over $(Q, \mathbf{f})$. The Q-marginal of this cut distribution behaves just as Π_1, on which we have theoretical guarantees, while we show separately that this distribution concentrates on neighbourhoods of the marginal density $g^{(3)}_{Q,\mathbf{f}}$ of three consecutive observations from the HMMs, in a result analogous to those discussed in Section 1.5. We then obtain a corresponding rate for the emission densities themselves by proving an inversion inequality for the L^1 metric, which is of independent interest for non-Bayesian approaches also. Finally, we control the smoothing probabilities under this cut posterior distribution, avoiding the need for splitting the sample.

In Chapter 3, we present a brief extension of the work of Chapter 2, titled *Bayesian Model Selection in Semiparametric Mixtures and Hidden Markov Models.*

In this work, we investigate a limitation of the work of Chapter 3 and also [45], namely the

requirement that the number of states R for the mixture or HMM is fixed and known in advance. In our investigation, we study a Bayesian model selection approach for choosing R, as introduced in Section 1.6. The value of R may be the sole object of inference in itself, or it may be used in combination with the work of Chapter 2 on the parameters $(Q, \mathbf{f})$ to ensure estimates are produced in the 'correct' class.

We first prove convergence of a penalised sequential Bayes factor approach for multinomial mixture models and HMMs, with these models providing a framework for semiparametric inference following the approach of [45]. Our conclusion is weaker than that of [23], but their assumptions do not hold in the multinomial setting and so a separate argument is instead presented. The need for such penalisation is not entirely clear, but it offsets the slower rate of posterior concentration which arises when the transition matrix is approximated, in overspecified models, by matrices which are almost singular. The result is most similar to that of [66], although the use of the marginal likelihood means that our approach has a Bayesian interpretation and is similar to the widely-used BIC.

We also give a negative result which indicates an issue with applying existing proof techniques to the setting we consider, even in the mixture setting where there is of course no issue arising from singular matrices. Finally, we demonstrate the procedure in practical examples.

In Chapter 4, titled *Bayesian Analysis of Conflict risk with Multivariate Point Processes* we consider a separate, though thematically related, line of work. Our methodology builds upon the approach of [16], although we take a multivariate approach to modelling spatial dependency which allows us to model more accurately the heterogeneity in our data, and provide interpretable measures of risk.

We begin by discussing the shortcomings of the spatial kernel approach of [16], as pertains to our case study in Bangladesh, and justify the need for more flexible spatial modelling alongside a heavier-tailed likelihood. We demonstrate the efficacy of this methodology on simulated data, which also demonstrates the possibility for, and benefit of, expert-tuning of the approach to eliminate spurious spatial links *a priori*. Finally, we demonstrate the model's application to the motivating case study of Bangladesh.

2. Efficient Bayesian Estimation and use of Cut Posterior in Semiparametric Hidden Markov Models

2.1 Introduction

Hidden Markov models (HMMs) are a broad and widely used class of statistical models, enjoying applications in a variety of settings where observed data is linked to some ordered process, for which an assumption of independently distributed data would be both inappropriate and uninformative. Specific applications include modelling of weather [3,60], circadian rhythms [59], animal behaviour [29,64], finance [75], information retrieval [79,99], biomolecular dynamics [56], genomics [107] and speech recognition [88].

In this paper, we consider inference in finite state space HMMs. Such models are characterised by an unobserved (latent) Markov chain $(X_t)_{t\geq 1}$ taking values in $[R] = \{1, 2, \ldots, R\}$ with $R < \infty$, evolving according to a transition matrix Q. Conditionally on $X_t = j$, $Y_t \sim F_j$ where F_j is the *emission distribution* with associated density f_j. These models generalise independent mixture models, which are obtained as a special case when the X_t are independent and identically distributed. Here we assume that R is known so that the parameters in such models are then Q and $\mathbf{F} = (F_1, \ldots, F_R)$.

Most of the work on HMMs considers parametric models, where the emissions are assumed to admit densities in a parametric class $\{f_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^d$ for some $d < \infty$, see for instance [18, 33, 39]. However such an assumption leads to inference which is strongly influenced by the choice of the parametric family $\{f_\theta : \theta \in \Theta\}$. This problem has been often discussed in the literature, see for instance [88], [36] or [107], especially but not solely in relation to clustering or segmentation. In the seminal paper [42], the authors show identifiability under weak assumptions on the emissions, provided the transition

matrix is of full rank, paving the way for estimation of semiparametric models. In [4], it is shown further that the number R of hidden states may be identified. In [8, 58] and [27], frequentist estimators of Q and $\mathbf{f} = (f_1, \dots, f_R)$ respectively have been proposed using spectral methods, showing in particular that Q can be estimated at the rate $1/\sqrt{n}$. Frequentist estimation of the emission densities has also been addressed using penalised least squares approaches [26] and spectral methods [27]. However, no results exist on the asymptotic distribution of frequentist estimators for Q, nor on efficient estimation for Q.

Although Bayesian nonparametric estimation methods have been considered in practice in Hidden Markov models, see for instance [107] or [36], little is known about their theoretical properties. While [102, 103] established posterior consistency under general conditions on the prior and refined the analysis to derive posterior concentration rates on the marginal density of ℓ successive observations, $g^{\ell}_{Q,\mathbf{F}}$, no results exist regarding the properties of Bayesian procedures when seeking to recover the parameters Q and $\mathbf{F}$, or other functionals of $Q, \mathbf{F}$ which are often of interests in HMMs. For instance, when interests lie in clustering or segmentation, the quantities of interest are the smoothing probabilities, being the conditional distribution of the latent states given the observations. In [27], the authors obtain rates of convergence for frequentist estimators of the smoothing probabilities, but their result requires either a sup-norm convergence for the estimator $\hat{\mathbf{f}}$ or splitting the data into two parts, with estimation of $\mathbf{f}$ based on one part of the data and estimation of the smoothing probabilities based on the other part of the observations. In this paper we intend to bridge this gap, concentrating on Bayesian semiparametric methods while also exhibiting non-Bayesian, semiparametric efficient estimators of Q.

We first construct a family of priors Π_1 on $(Q, \mathbf{F})$ which we show in Theorem 2.4 leads to an asymptotically normal posterior distribution for $\sqrt{n}(Q - \hat{Q})$, of variance V detailed therein. Here $\hat{Q}$ is a frequentist estimator, exhibited in Theorem 2.3, for which $\sqrt{n}(\hat{Q} - Q^*)$ converges in distribution to $N(0, V)$, with Q^* being the frequentist true parameter under which the data is assumed to be generated. Consequently, Bayesian point estimates associated to such posteriors, such as the posterior mean, enjoy parametric convergence rates to Q^* and, importantly, credible regions for Q are asymptotically confidence regions. In Theorem 2.6, we then refine this construction to obtain a scheme for which V is the optimal variance for semiparametric estimation of Q, i.e. it is the inverse efficient Fisher information. Semiparametric Bernstein-von Mises properties are highly non trivial results and correspondingly sparse in the literature, moreover in [21, 22, 38, 91] a number of coun-

terexamples are exhibited, where semiparametric Bernstein-von Mises results do not hold. Since this property is crucial to ensure that credible regions are also confidence regions and thus robust to the choice of the prior distribution, it is important to study and verify it.

Our approach to obtaining these results follows the ideas of [45], extending their work on mixtures to the more complex HMM setting. In particular, the construction of $\hat{Q}$ (together with the prior on $Q, \mathbf{F}$) is based on a parametric approximation to the nonparametric emission model, with the property that estimation in this model with appropriately 'coarsened' data leads to a well-specified parametric model for estimation of Q. Once we reduce to such a parametric model, we have asymptotic normality of the corresponding MLE as in [12] as well as Bernstein-von Mises for the posterior as in [28], although the intuition that our 'coarsened' data is less informative translates to an inefficient asymptotic variance V. We then show that we can construct efficient estimators of Q by adjusting the degree of coarsening with the sample size, similar to the approach of [45]. The proof techniques in our context are however significantly more complex, since the notions of Fisher information and score functions are much less tractable in hidden Markov models (see for instance [32]).

The prior distributions $\Pi_1(dQ, d\mathbf{F})$ considered above, which lead to the Bernstein-von Mises property of the posterior on Q, rely on a crude modelling of the emission densities and are therefore not well behaved as far as the estimation of $\mathbf{F}$ is concerned. To overcome this problem, we adapt in Section 2.4 the cut posterior approach (see [10, 20, 61, 74, 87]) to our semiparametric setting. Cut posteriors were originally proposed in the context of modular approaches to modelling, where a model is assembled from a number of constituent models, each with its own parameters θ_i and data Y_i. In the usual construction, the cut posterior has the effect of 'cutting feedback' of one of the (less reliable) data sources Y_i on the other parameters (associated to more reliable data).

Our approach, though also modular, departs from this setting in that we use a single data source but wish to choose different priors for different parts of the parameter. We consider a conditional prior $\Pi_2(d\mathbf{F}|Q)$, then combine $\Pi_1(dQ|Y_{1:n})$ with the conditional posterior $\Pi_2(d\mathbf{F}|Q, Y_{1:n})$ to produce a joint distribution over Q and $\mathbf{F}$. In this way, we construct a distribution over the parameters which is not a proper Bayesian posterior, but which simultaneously satisfies a Bernstein-von Mises theorem for the posterior marginal on Q, and is well-concentrated on the emission distributions and other functionals such as the

smoothing probabilities. Through this construction, we manage to combine the "best of both worlds" in the estimation of the parametric and nonparametric parts. We believe that this idea could be used more generally in other semiparametric models.

As previously mentioned, the existing posterior concentration result in the semiparametric HMMs covered only the marginal density $g^{\ell}_{Q,\mathbf{F}}$ of a fixed number ℓ of consecutive observations, see [102]. A key step in obtaining posterior contraction rates on the emission distributions $\mathbf{F}$ is an inversion inequality allowing us to deduce L^1 concentration of the posterior distribution of the emission distributions from concentration (at the same rate) of the marginal distribution of the observations. This is established in Theorem 2.10, from which we derive contraction rates for the cut-posterior on $\mathbf{f}$ in Theorem 2.12. This inversion inequality is of independent interest and can be used outside the framework of Bayesian inference. We finally show in Theorem 2.13 that these results lead to posterior concentration of the smoothing distributions, which are the conditional distributions of the hidden states given the observations, building on [27] but refining the analysis so that we require neither sup-norm contraction rate on the emissions $\mathbf{f}$, nor a splitting of the data set in two parts.

Organisation of the chapter The chapter is organised as follows. In Section 2.2, we introduce the model and the notions involved, together with a general strategy for inference on Q and $\mathbf{f}$ based on the cut posterior approach. In Section 2.3, we study the estimation of Q, proving asymptotic normality of the posterior and asymptotic efficiency. In Section 2.4, the cut posterior approach is studied, posterior contraction rates for $\mathbf{f}$ are derived together with posterior contraction rates for the smoothing probabilities. Theoretical results from both sections are then illustrated in Section 2.5. The proofs for the efficiency of the estimation of Q (Proposition 2.5), for the inversion inequality (Theorem 2.10), and for the convergence of the smoothing probabilities (Theorem 2.13), are presented in Section 2.6. The other proofs are deferred to Sections 2.8 and 2.10. Section 2.9 contains a general contraction rate theorem for cut posteriors, while Section 2.12 contains an exposition of relevant results in the literature, and Section 2.13 contains further details of simulations.

2.2 Inference in semiparametric finite state space HMMs

In this section we present our strategy to estimate Q and $\mathbf{F} = (F_1, \dots, F_R)$, based on a modular approach which consists of first constructing the marginal posterior distribution

of Q based on a first prior Π_1 on $(Q, \mathbf{F})$, and then combining it with the conditional posterior distribution of $\mathbf{F}$ given Q and $Y_{1:n}$ based on a different prior $\Pi_2(d\mathbf{F}|Q)$.

2.2.1 Model and notation

Hidden Markov models (HMMs) are latent variable models where one observes $Y_{1:n} = (Y_t)_{1\leq t\leq n}$ whose distribution is modelled via latent (non-observed) variables $X_{1:n} = (X_t)_{1\leq t\leq n} \in [R]^n$ which form a Markov chain. In this work we consider finite state space HMMs:

$$Y_t|X_t \overset{\text{ind}}{\sim} F_{X_t}, \quad \mathbb{P}_Q(X_t = r|X_{t-1} = s) = Q_{rs}, \quad r, s \leq R, \quad t \geq 1 \tag{2.1}$$

and the number of states R is assumed to be known throughout the paper.

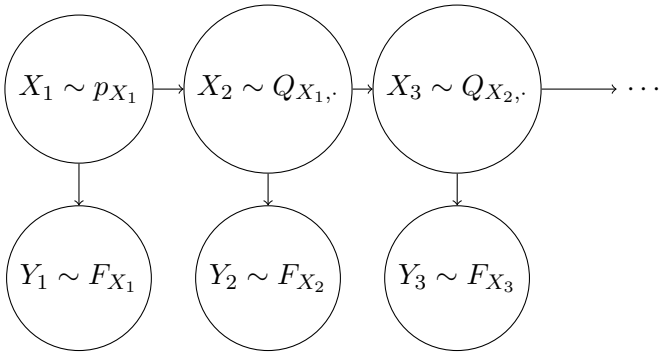


Figure 2.1: Visual representation of data generating process of HMM

The parameters are then the transition matrix $Q = (Q_{rs})_{r,s\leq R}$ of the Markov chain and the emission distributions $\mathbf{F} = \{F_r\}_{r\leq R}$, which represent the conditional distribution of the observations given the latent states. For a transition matrix Q we denote by p_Q its invariant distribution (when it exists).

Throughout the paper we denote by $\mathbf{F}^* = \{F_r^*\}_{r\leq R}$ and Q^* respectively the true emission distributions and the true transition matrix. The aim is to make inference on $\mathbf{F}^*, Q^*$, and some functionals of these parameters, using likelihood based methods and in particular Bayesian methods. We assume that the distributions F_j^*, $j = 1, \ldots, R$ are absolutely continuous with respect to some measure λ on $\mathcal{Y} \subset \mathbb{R}^d$, $d \geq 1$ and we denote by $f_1^*, \ldots, f_R^*$ their corresponding densities.

When the latent states $(X_t)_{t\geq 1}$ are independent and identically distributed on $[R]$, the parameters are not identifiable unless some strong assumptions are made on the F_j's, see [5]. However, in [42], it is proved that under weak assumptions on the data generating process, both Q and $\mathbf{F}$ are identifiable up to label swapping (or label switching). More

precisely, let $\mathbb{P}^{(\ell)}_{Q,\mathbf{F}}$ be the marginal distribution of ℓ consecutive observations from model 2.1 with parameters Q and $\mathbf{F}$, so that

$$\mathbb{P}^{(\ell)}_{Q,\mathbf{F}}(dy_{1:\ell}) = \sum_{i_1,\dots,i_\ell=1}^{R} p_Q(i_1) \prod_{t=1}^{\ell-1} \left(Q_{i_t,i_{t+1}} F_{i_t}(dy_t)\right) F_{i_\ell}(dy_\ell). \tag{2.2}$$

Under assumptions of ergodicity, p_Q exists and is unique. For example, this holds under the assumption $Q_{ij} > 0$ for all $i, j \in [R]$, see for instance equation (1) in [43]. Denote by $\mathcal{Q} = \Delta_R^R \subset [0,1]^{R\times R}$ the $R-$fold product of the $R-1$ dimensional simplex and let $\mathcal{P}$ be the set of probability measures on $\mathbb{R}^d$. Consider the following assumption:

Assumption 2.1. (i) The latent chain $(X_t)_{t\geq 1}$ has true transition matrix $Q^* = (Q^*_{ij})$ satisfying $\det Q^* \neq 0$.

(ii) The true emission distributions $\{F^*_r\}_{r=1}^R$ are linearly independent.

From [42], if **Assumption** 2.1 holds, then for any $Q \in \mathcal{Q}$ and any $F_1, \dots, F_R \in \mathcal{P}$, if $\mathbb{P}^{(3)}_{Q,\mathbf{F}} = \mathbb{P}^{(3)}_{Q^*,\mathbf{F}^*}$ then $Q = Q^*$ and $\mathbf{F} = \mathbf{F}^*$ up to label swapping. By "up to label swapping," we precisely mean that there exists a permutation τ of $[R]$ such that ${}^\tau Q = Q^*$ and ${}^\tau \mathbf{F} = \mathbf{F}^*$, where ${}^\tau Q = (Q_{\tau(r),\tau(s)}, r, s \in [R])$ and ${}^\tau \mathbf{F} = (F_{\tau(1)}, \dots, F_{\tau(R)})$. The requirement for such a permutation τ is unavoidable, since the labelling of the hidden states is fundamentally arbitrary. Correspondingly, the results which follow will always be given up to label swapping. In a slight abuse of notation, we will sometimes interchange $\mathbf{F}$ and $\mathbf{f} = (f_1, \dots, f_R)$, the latter being the densities of $\mathbf{F}$ with respect to some measure λ in a dominated model.

The likelihood associated to model (2.1), when $(F_r)_{r\in[R]}$ are dominated by a measure λ, is then given by

$$L_n(Q, \mathbf{F}) = g^{(n)}(Q, \mathbf{F})(y_1, \dots, y_n) = \sum_{i_1,\dots,i_n=1}^{R} p_Q(i_1) \prod_{t=1}^{n-1} Q_{i_t,i_{t+1}} f_{i_t}(y_t) f_{i_n}(y_n). \tag{2.3}$$

Extension to initial distributions different from the stationary one is straightforward, under the exponential forgetfulness of the Markov chain, which holds under our assumptions below, see Section 2.12.1.

If Π is a prior on $\mathcal{Q} \times \mathcal{F}^R$, with $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}_+ : \int f(y) d\lambda(y) = 1\}$ being the set of densities on $\mathbb{R}^d$ with respect to λ, then the Bayesian posterior $\Pi(\cdot|Y_{1:n})$ is defined as

follows: for any Borel subset A of $\mathcal{Q} \times \mathcal{F}^R$, we have

$$\Pi(A|Y_{1:n}) = \frac{\int_A L_n(Q, \mathbf{F})\Pi(dQ, d\mathbf{F})}{\int_{\mathcal{Q}\times\mathcal{F}^R} L_n(Q, \mathbf{F})\Pi(dQ, d\mathbf{F})}. \tag{2.4}$$

This posterior is well defined as soon as p_Q exists $\Pi-$almost surely, which holds for instance when $\Pi(\min_{i,j} Q_{ij} > 0) = 1$, which implies that the transition matrix is ergodic $\Pi-$almost surely by the earlier remarks. Throughout the paper we consider the parameterisation of Q given by $\tilde{Q} = (Q_{rs}, r \in [R], s \in [R-1])$ so that for each r we have $Q_{rR} = 1 - \sum_{s<R} Q_{rs}$. Hence, specification of the matrix Q amounts to specification of the $(R) \times (R-1)$ matrix $\tilde{Q} \in \tilde{\mathcal{Q}}$ for which $\sum_{s=1}^{R-1} Q_{rs} \leq 1$ for all r, and we will identify Q with $\tilde{Q}$ (and $\mathcal{Q}$ with $\tilde{\mathcal{Q}}$) when making statements about asymptotic distributions.

It will be helpful to consider that the stationary HMM $(X_t, Y_t)_{t\geq 1}$ arises as the restriction to $\{t \geq 1\}$ of a process $(X_t, Y_t)_{t\in\mathbb{Z}}$ indexed by $t \in \mathbb{Z}$, where the $(X_t)_{t\in\mathbb{Z}}$ and the $(Y_t)_{t\leq 0}$ are not observed. This is possible to define by considering the reversal of the latent chain.

We use $\mathbb{P}_*$ to denote the joint law of the variables $(X_t, Y_t)_{t\in\mathbb{Z}}$ under the stationary distribution associated with the parameters $Q^*, \mathbf{F}^*$. Estimators, given a sample of size n, are then understood to be random variables which are measurable with respect to the σ-algebra generated by the observations $(Y_t)_{1\leq t\leq n}$.

2.2.2 Cut posterior inference: A general strategy for joint inference on Q and $\mathbf{f}$

In Section 2.3.1 below we construct a family of prior distributions Π_1 on $\mathcal{Q} \times \mathcal{F}^R$ such that the associated marginal posterior distribution of Q satisfies a Bernstein-von Mises theorem centred at an asymptotically normal regular estimator $\hat{Q}$. Specifically, this means that with probability going to 1 under $\mathbb{P}_*$,

$$\exists \tau \in \mathcal{S}_R : \quad \|\Pi_1(\sqrt{n}({}^\tau Q - \hat{Q}) \in (\cdot)|Y_{1:n}) - N(0, V)\|_{TV} = o_{\mathbb{P}_*}(1) \tag{2.5}$$

and, for this permutation τ,

$$\sqrt{n}({}^\tau \hat{Q} - Q^*) \Rightarrow^* N(0, V), \tag{2.6}$$

where $\Rightarrow^*$ denotes convergence in distribution under $\mathbb{P}_*$ and where $\mathcal{S}_R$ is the set of permutations of $[R]$.

However, the choice of Π_1 which we require for this control over Q leads to a posterior distribution on $\mathbf{F}$ which is badly behaved, see Section 2.3. In order to jointly estimate well both Q and $\mathbf{F}$, we propose the following cut posterior approach. Consider a second conditional prior distribution $\Pi_2(d\mathbf{F}|Q)$ which leads to a conditional posterior distribution of $\mathbf{F}$ given $Q, Y_{1:n}$ of the form

$$\Pi_2(d\mathbf{F}|Q, Y_{1:n}) = \frac{L_n(Q, \mathbf{F})\Pi_2(d\mathbf{F}|Q)}{\int_{\mathcal{F}^R} L_n(Q, \mathbf{F})\Pi_2(d\mathbf{F}|Q)}.$$

The cut posterior is then the probability distribution on $\mathcal{Q} \times \mathcal{F}^R$ given by

$$\Pi_{cut}(d(Q, \mathbf{F})|Y_{1:n}) = \Pi_1(dQ|Y_{1:n})\Pi_2(d\mathbf{F}|Q, Y_{1:n}). \tag{2.7}$$

Note that, if $\Pi_2(d\mathbf{F}|Q) = \Pi_1(d\mathbf{F}|Q)$, then $\Pi_{cut}(d(Q, \mathbf{F})|Y_{1:n})$ is a proper posterior distribution and is equal to $\Pi_1(d(Q, \mathbf{F})|Y_{1:n})$. The motivation behind the use of the cut posterior $\Pi_{cut}(d(Q, \mathbf{F})|Y_{1:n})$ is to retain the good behaviour of $\Pi_1(dQ|Y_{1:n})$ in terms of estimating Q, while being flexible in the modelling of $\mathbf{F}$ to ensure that the posterior distribution over both $\mathbf{F}$ and Q (and functionals of these parameters) are well behaved.

Adapting the proof techniques from [48] to posterior concentration rates for cut posteriors, we derive in Section 2.4 contraction rates for cut posteriors in terms of the L^1 norm of $g^{(3)}_{Q,\mathbf{F}} - g^{(3)}_{Q^*,\mathbf{F}^*}$, where $g^{(3)}_{Q,\mathbf{F}}$ is the density of $\mathbb{P}^{(3)}_{Q,\mathbf{F}}$ with respect to the dominating measure λ. We show that, under suitable conditions and choice of $\epsilon_n = o(1)$, $\Pi_{cut}(\|g^{(3)}_{Q,\mathbf{F}} - g^{(3)}_{Q^*,\mathbf{F}^*}\|_1 \leq \epsilon_n|Y_{1:n}) = 1 + o_{\mathbb{P}_*}(1)$.

To derive cut posterior contraction rates in terms of the L^1 norm of $\mathbf{f} - \mathbf{f}^*$, we prove in Section 2.4.2 an inversion inequality in the form

$$\sum_{r=1}^{R} \|f_r - f_r^*\|_1 \lesssim \|Q - Q^*\| + \|g^{(3)}_{Q,\mathbf{F}} - g^{(3)}_{Q^*,\mathbf{F}^*}\|_{L^1},$$

which is also of independent interest.

We also derive cut-posterior contraction rates for the smoothing probabilities $(p_{Q,\mathbf{F}}(X_i = \cdot|Y_{1:n}))_{i=1,\ldots,n}$ in Section 2.4.3. In contrast with [27], concentration rates for $(p_{Q,\mathbf{F}}(X_i = \cdot|Y_{1:n}))_{i=1,\ldots,n}$ do not require a splitting of the data into 2 groups, nor do they require to to have a control of $f_r - f_r^*$ in sup-norm. We can avoid these difficulties thanks to the Bayesian approach as is explained in Section 2.4.3.

In our implementation of the cut posterior, we adopt a nested MCMC approach of the

kind detailed in [20] and [61], see Section 2.5 for details.

In the following section, we present Π_1 and show that the associated marginal posterior distribution $\Pi_1(dQ|Y_{1:n})$ is asymptotically normal in the sense of Equation (2.5).

2.3 Semi - parametric estimation of Q: Bernstein-von Mises property and efficient estimation

The prior Π_1 is based on a simple histogram prior on the $f_1, \dots, f_R$. For the sake of simplicity, we present the case of univariate data; the multivariate case can be treated similarly. Without loss of generality, we assume that the observations belong to $[0,1]$ and note that, if $\mathcal{Y} = \mathbb{R}$, we can transform the data to $[0,1]$ via some C^1 diffeomorphism ϕ prior to the analysis. The prior relies on a partition of the space $[0,1]$ into finitely many intervals $\{I_1, I_2, \dots\}$, and transforming the data is equivalent to constructing a prior based on the corresponding partition $\{\phi^{-1}(I_1), \phi^{-1}(I_2), \dots\}$ of $\mathbb{R}$. The construction of Π_1 is analogous to the construction considered in [45].

Let $M \in \mathbb{N}$ with $M \geq R$ and consider $\mathcal{I}_M = (I_m^{(M)}, m \leq \kappa_M)$ a partition of $[0,1]$ into κ_M bins, with $\kappa_{(\cdot)} : \mathbb{N} \to \mathbb{N}$ a strictly increasing sequence. Given $\mathcal{I}_M$, we consider the model of piece-wise constant densities as the set, $\mathcal{F}_M$, of densities with respect to Lebesgue measure, in the form:

$$f_\omega = \sum_{m=1}^{\kappa_M} \frac{\omega_m}{|I_m^{(M)}|} \mathbb{1}_{I_m^{(M)}}, \quad \min_m \omega_m \geq 0, \quad \sum_{m=1}^{\kappa_M} \omega_m = 1, \tag{2.8}$$

where $|I_m^{(M)}|$ denotes the length of interval $I_m^{(M)}$. For instance, one could consider a sequence of dyadic partitions with $\kappa_M = 2^M$. Such partitions are admissible for sufficiently large M in the sense we detail below, and are used for the empirical investigation of Section 2.5.

The parameters for this model are then Q and $\underline{\omega}_{(M)} = (\omega_{mr})_{m=1,\dots,\kappa_M}^{r=1,\dots,R}$, the latter of which varies in the set

$$\Omega_M = \{\underline{\omega}_{(M)} \in [0,1]^{\kappa_M \times R} \; ; \sum_{m=1}^{\kappa_M} \omega_{mr} = 1, \forall r \in [R]\}.$$

Through (2.8), we identify each $\underline{\omega}_{(M)} \in \Omega_M$ with a vector of emission densities $\mathbf{f}_{\underline{\omega}_{(M)}} \in \mathcal{F}_M^R$, and thus a prior distribution Π_M over the parameter space $\mathcal{Q} \times \Omega_M$ is identified with a prior distribution Π over $\mathcal{Q} \times \mathcal{F}_M^R$. The corresponding posterior distribution is denoted $\Pi_M(\cdot|Y_{1:n})$ and is defined through (2.4).

Throughout this section, we write $\underline{\omega}_r := (\omega_{mr}, m \leq \kappa_M)$ and we denote by $\mathcal{M}(\mathcal{I}_M)$ the hidden Markov model associated with densities of the form (2.8). Note that, for all $M > 0$, $\mathcal{M}(\mathcal{I}_M)$ is of dimension $(\kappa_M - 1)^R R(R-1) < \infty$.

A key argument used in [42], to identify Q^* from $g^{(3)}_{Q^*,\mathbf{f}^*}$, is to find a partition $\mathcal{I}_M$ for some $M > 0$ such that the matrix

$$\mathbf{F}^*(\mathcal{I}_M) := \begin{pmatrix} F_1^*(I_1^{(M)}), & \cdots, & F_1^*(I_{\kappa_M}^{(M)}) \\ F_2^*(I_1^{(M)}), & \cdots, & F_2^*(I_{\kappa_M}^{(M)}) \\ \vdots & \cdots. & \vdots \\ F_R^*(I_1^{(M)}), & \cdots, & F_R^*(I_{\kappa_M}^{(M)}) \end{pmatrix}$$

has full rank. We call such a partition an *admissible* partition.

Definition 2.1. A partition $\mathcal{I}_M$ is said to be *admissible* for $f_1^*, \ldots, f_R^*$ if the rank of $\mathbf{F}^*(\mathcal{I}_M)$ is equal to R.

Remark 2.2. Although we are using piece-wise constant functions to model the emission densities, we do not assume that the f_r^*, $r \in [R]$ are piece-wise constant and the simplified models M are not meant to lead to good approximation of the emissions densities. However, as far as the parameter Q is concerned, the likelihood induced by such a model is not misspecified but instead corresponds to a coarsening of the data. Specifically, it corresponds to the likelihood associated to observations of $Y_i^{(M)} = (\mathbb{1}_{I_1^{(M)}}(Y_i), \cdots, \mathbb{1}_{I_{\kappa_M}^{(M)}}(Y_i))$; for such observations this simplified model leads to a well-specified likelihood. Note also that, although we are modelling densities with respect to Lebesgue measure, we do not require F_r^* to have density with respect to Lebesgue measure since the quantities of importance are the probabilities $F_r^*(I_m^{M)})$, $r \in [R]$, $m \in [\kappa_M]$.

This particular coarsening was introduced in the context of mixtures by [45]. It is not at all obvious how other types of coarsening can be found, which are associated with well-specified parametric models for the coarsened data and the parameter of interest Q.

2.3.1 Asymptotic normality and Bernstein-von Mises

In this section, we study the asymptotic behaviour of the marginal posterior distribution of Q under model $\mathcal{M}(\mathcal{I}_M)$, together with the asymptotic normality of the maximum likelihood estimator $\hat{Q}_{n,M}$ in this model. As mentioned in Remark 2.2, the likelihood associated

to model $\mathcal{M}(\mathcal{I}_M)$ is given by

$$L_n(Q, \omega_{(M)}) = \sum_{i_1,\ldots,i_n=1}^{R} p_Q(i_1) \left(\prod_{t=1}^{n-1} Q_{i_t,i_{t+1}} f_{\underline{\omega}_{i_t}}(Y_t^{(M)}) \right) f_{\underline{\omega}_{i_n}}(Y_n^{(M)}), \tag{2.9}$$

with the abuse of notation $f_{\underline{\omega}_{i_t}}(Y_t^{(M)}) = \sum_{m=1}^{\kappa_M} \frac{\omega_{mi_t} Y_t^{(M)}(m)}{|I_m^{(M)}|}$ for $t = 1, \ldots, n$.

In other words, our likelihood becomes one of a hidden Markov model with finite state space and multinomial emission distributions. Under $\mathbb{P}_*$, the coarsened observations $Y_{1:n}^{(M)}$ arise from a hidden Markov model with multinomial emission distributions, governed by parameters $\underline{\omega}_{(M)}^*$ and transition matrix Q^*, with $\omega_{mr}^* = F_r^*(I_m^{(M)})$. We write $\theta = \theta^{(M)} := (Q, \underline{\omega}_{(M)})$ for the parameter of $\mathcal{M}(\mathcal{I}_M)$, and suppress the superscript on θ when the dependence on M is clear.

Asymptotic normality of the maximum likelihood estimator (MLE) of parametric finite state space hidden Markov models was considered for instance in [12], who showed that the MLE was asymptotically normal with covariance matrix given by the inverse of the Fisher information matrix, which is given by the limiting covariance matrix of the score statistics; see Lemma 1 of [12].

Let $J_M(\theta)$ be the Fisher information matrix associated to the likelihood (2.9):

$$J_M(\theta^*) = \lim_{n\to\infty} \frac{\mathbb{E}_* \left[-D^2 \log L_n(Q^*, \underline{\omega}_{(M)}^*) \right]}{n},$$

as defined in [12]. We also write $J_M[Q, Q], J_M[\omega, \omega]$, and $J_M[Q, \omega]$ to denote the submatrices corresponding to the second derivatives with respect to (Q, Q), (ω, ω) and (Q, ω) respectively.

The following theorem demonstrates that asymptotically normal, parametric-rate estimators of the transition matrix exist, although such estimators may not have optimal asymptotic variance.

Theorem 2.3. *Let $M > 0$ and $\mathcal{I}_M$ be an admissible partition for $\mathbf{F}^*$, and let $\hat{Q}_{n,M} = \hat{Q}_n^{(M)}(Y_{1:n})$ be the MLE in model $\mathcal{M}(\mathcal{I}_M)$, given observations $Y_{1:n}$. Grant **Assumption 2.1**, and assume also that the transition matrix Q^* is irreducible and aperiodic, and that $\omega_{mr}^* > 0$ for all $m \in [\kappa_M]$ and all $r \in [R]$. We then have (up to label-swapping)*

$$\sqrt{n}(\hat{Q}_{n,M} - Q^*) \Rightarrow^* N(0, \tilde{J}_M^{-1}),$$

where $\tilde{J}_M^{-1}$ *is positive definite and*

$$\tilde{J}_M^{-1} = (J_M^{-1})[Q,Q] = J_M[Q,Q] - J_M[Q,\omega]J_M[\omega,\omega]^{-1}J_M[\omega,Q].$$

The main difficulty in the proof of Theorem 2.3 is showing that the Fisher information matrix J_M is invertible.

Proof. The model $\mathcal{M}(\mathcal{I}_M)$ is a regular parametric HMM, hence using Theorem 1 of [12], we establish asymptotic normality of the MLE for the parameter $\theta = (Q, \underline{\omega}_{(M)})$ as soon as J_M is invertible. We prove invertibility of J_M in Section 2.10.1, then projecting onto the Q co-ordinates gives the result. □

To derive the Bernstein-von Mises Theorem associated to model $\mathcal{M}(\mathcal{I}_M)$, we need the following assumption on the prior distributions on Q and $\underline{\omega}_{(M)}$:

Assumption 2.2. (i) The prior Π_Q on Q has positive and continuous density on $\mathcal{Q}$.
(ii) The prior on $\underline{\omega}_{(M)}$ has positive and continuous density with respect to Lebesgue measure.

Theorem 2.4. *Let* $M > 0$, $\mathcal{I}_M$ *be an admissible partition for* $\mathbf{f}^*$, *and let* Π_M *be a prior satisfying* ***Assumption*** *2.2. Grant* ***Assumption*** *2.1, and assume also that that the transition matrix* Q^* *is irreducible and aperiodic, and that* $\omega^*_{mr} > 0$ *for all* $m \in [\kappa_M]$ *and all* $r \in [R]$. *We then have (up to label-swapping)*

$$\|\Pi_M(\sqrt{n}(Q - \hat{Q}_{n,M}) \in (\cdot)|Y_{1:n}) - N(0, \tilde{J}_M^{-1})\|_{TV} = o_{\mathbb{P}_*}(1).$$

As with the proof of Theorem 2.3, parametric results apply as soon as the Fisher information matrix is seen to be invertible.

Proof. Inspecting the proof of Theorem 1 of [12], we see (for $\theta = (Q, \underline{\omega}_{(M)})$, $J_M = J_M(\theta^*)$ the Fisher information and $l_n = \log g_n$ the log-likelihood) that

$$\sqrt{n}(\hat{\theta} - \theta^*) = n^{-\frac{1}{2}} J_M^{-1} \nabla_\theta l_n(\theta^*) + o_{\mathbb{P}_*}(1),$$

which up to $o_{\mathbb{P}_*}(1)$ is equal to the T_n of [28] at which their Bernstein-von Mises result (Theorem 2.1) is centred. Then this result implies the total variation convergence, in $\mathbb{P}_*$-probability, of the posterior to the given normal distribution, by considering the marginal posterior on Q. We remark further that the MLE $\hat{\theta}$ is regular, which follows from the

characterisations of Fisher information in Lemmas 1 and 2 of [12], the expansion of the MLE above, and an application of Le Cam's third lemma (Example 6.7 of [100]) along the lines of the proof of Lemma 8.14 in [100]. □

An interesting feature of Theorems 2.3 and 2.4 is that they essentially only require that $\mathcal{I}_M$ is an admissible partition of $\mathbf{F}^*$. For a given partition, this is an assumption on $\mathbf{F}^*$, and indeed the choice of the partition is important. Note, however, that under Assumption 2.1 and, for instance, if the (Lebesgue) densities f_r^* are positive and continuous, then for all sequences of embedded partitions $(\mathcal{I}_M)_M$ with radius going to 0 there exists an M such that $\mathcal{I}_M$ is admissible for $\mathbf{F}^*$. More discussion is provided in Section 2.5.

Combining Theorems 2.3 and 2.4, we see that credible regions for Q, based on the posterior associated to Π_M, are also asymptotic confidence regions. Their size may not be optimal however, even asymptotically. To ensure that such credible regions have optimal size while being asymptotic confidence regions, we would require that $\tilde{J}_M^{-1}$ is the best possible (asymptotic) variance, but this is not true in general; while $\tilde{J}_M^{-1}$ is the efficient covariance matrix for the estimation of Q in model $\mathcal{M}(\mathcal{I}_M)$, it is not necessarily the semiparametric efficient covariance matrix for the estimation of Q in model (2.3). The existence of an efficient estimator of Q in the semiparametric hidden Markov model, with likelihood (2.3), has not been established, although the fact that $\sqrt{n}$-convergent estimates of Q exist in the literature (see Section 2.10.1) indicate that semiparametric efficient estimation of Q should be possible.

In the following section we construct an efficient estimator for Q and a prior leading to an efficient Bernstein-von Mises theorem.

2.3.2 Efficient estimation

In [45], in the context of semiparametric mixture models with finite state space, the authors also consider the prior model (2.8) for the emission distributions. They derive for fixed M a Bernstein-von Mises theorem similar to Theorem 2.4 and they show that, if we let $M = M_n$ increase to infinity sufficiently slowly as $n \to \infty$, and if the corresponding partitions $\mathcal{I}_{M_n}$ are embedded with radius going to 0, then $\tilde{J}_{M_n}$ converges to the efficient semiparametric Fisher matrix for the mixture weights. This allows them to define an efficient estimator, whose value given n observations is the maximum likelihood estimator in the histogram model with M_n bins.

In this section we prove a similar result, however the proof is significantly more involved than in the case of mixture models. To study the theory on efficient semiparametric estimation of Q in the semiparametric HMMs, we follow the approach of [78]. We first prove the LAN expansion in local sub-models, which allows us to describe the tangent space, and then we prove that $\tilde{J}_M$ converges to $\tilde{J}$ as M goes to infinity where $\tilde{J}$ is the semiparametric efficient Fisher information matrix for Q. Throughout Section 2.3.2 we assume that the F_r^* have density f_r^* with respect to Lebesgue measure on $[0,1]$ and that the following holds:

Assumption 2.3. For all $i, j \in [R], \quad Q_{ij}^* > 0$.

2.3.2.1 Scores and tangent space in the semiparametric model

We begin by exhibiting the local asymptotic normality (LAN) expansion for our model, following the framework of [78]. As is usual in semiparametric efficiency arguments, this involves identifying the score functions and LAN norm along one-dimensional sub-models passing through the true parameter; see Chapter 25 of [100] for discussion in the iid setting. Since these sub-models are themselves parametric, this identification can be made by following the framework of [32], who considered asymptotic normality in the context of parametric HMMs. A more thorough treatment, alongside a recollection of the relevant definitions and results in [78], is provided in Section 2.12.2 - in what follows, we aim to give an overview.

The first step is to exhibit a LAN expansion for the semiparametric model we consider. The parameter space is $\Theta = \mathcal{Q} \times \mathcal{F}^R$ and we consider the tangent space, in the sense of Definition 1 of [78], at the parameter $(Q^*, \mathbf{f}^*)$ as

$$\mathcal{H} = \{(a, \mathbf{h}), a \in \mathbb{R}^{R \times (R-1)}, \quad \mathbf{h} = (h_r)_{r=1}^R, \quad h_r \in \mathcal{H}_r\},$$

where $\mathcal{H}_r = \mathcal{H}_r(f_r^*) = \{h_r \in L^2(f_r^*) : \int h_r f_r^* = 1, \|h_r\|_\infty < \infty\}$, and where we use the parameterisation $\tilde{Q}$ defined in Section 2.2.1. Write

$$\theta_n(a, \mathbf{h}) = \left(Q^* + \frac{a}{\sqrt{n}}, f_1^*(1 + \frac{h_1}{\sqrt{n}}), \ldots, f_R^*(1 + \frac{h_R}{\sqrt{n}})\right).$$

Then $\theta_n(a, \mathbf{h})$ is a perturbation of the parameter $\theta^* = (Q^*, \mathbf{f}^*)$ along the path characterised

by a given element $(a, \mathbf{h}) \in \mathcal{H}$. Consider the sub-model

$$\Theta_{a,\mathbf{h}} = \{\theta_t = (Q^* + ta, (f_r^*(1 + th_r) : r \in [R])), \ |t| \leq \epsilon\},$$

for some ϵ (depending on $a, \mathbf{h}$) sufficiently small that $\Theta_{a,\mathbf{h}} \subset \Theta$. Then, for given $a, \mathbf{h}$ and for sufficiently large n, the perturbed parameters $\theta_n(a, \mathbf{h})$ are elements of $\Theta_{a,\mathbf{h}}$. This means that, for each $a, \mathbf{h}$, we can make an asymptotic expansion of the log-likelihood ratio between $\theta_n(a, \mathbf{h})$ and θ^* by considering the likelihood in the sub-model $\Theta_{a,\mathbf{h}}$. To this end, we expand the gradient of the log-likelihood $\ell_n^{(a,\mathbf{h})}(t)$, in $\Theta_{a,\mathbf{h}}$ at $t_0 = 0$ as

$$\frac{1}{\sqrt{n}} \nabla_t \ell_n^{(a,\mathbf{h})}(t_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \Delta_{k,\infty}^{(a,\mathbf{h})} + o_{L^2(\mathbb{P}_*)}(1),$$

where

$$\begin{aligned} \Delta_{k,\infty}^{(a,\mathbf{h})} = \Delta_{k,\infty}^{(a,\mathbf{h})}(Y_{-\infty:k}) &= \mathbb{E}_*[\phi(X_{k-1}, X_k, Y_k)|Y_{-\infty:k}] \\ &+ \sum_{i=-\infty}^{k-1} \left(\mathbb{E}_*[\phi(X_{i-1}, X_i, Y_i)|Y_{-\infty:k}] - \mathbb{E}_*[\phi(X_{i-1}, X_i, Y_i)|Y_{-\infty:k-1}] \right), \end{aligned} \tag{2.10}$$

and where, writing $a_{rR} = -\sum_s a_{rs}$ and $Q_{r,R} = 1 - \sum_s Q_{rs}$,

$$\phi(X_{k-1}, X_k, Y_k) = \sum_{r=1}^{R} \sum_{s=1}^{R} \mathbb{1}\{X_{k-1} = r, X_k = s\} \frac{a_{r,s}}{Q_{r,s}^*} + \sum_{s=1}^{R} \mathbb{1}\{X_k = s\} h_s(Y_k). \tag{2.11}$$

These formulae arise as an application of the results of Section 6.1 of [32] to the (parametric) model $\Theta_{a,\mathbf{h}}$.

We note that the contribution of the first term of the right hand side of (2.11), which may be rewritten as

$$\sum_{r=1}^{R} \sum_{s=1}^{R-1} \left(\frac{\mathbb{1}\{X_{k-1} = r, X_k = s\}}{Q_{rs}^*} - \frac{\mathbb{1}\{X_{k-1} = r, X_k = R\}}{Q_{rR}^*} \right) a_{rs},$$

to the expression defined in (2.10) is precisely the score function for estimation in the model $\Theta_{a,\mathbf{0}}$, in which the emission densities are fixed and known. We see that this is equal to $a^T S_{Q^*}(Y_{-\infty:k})$, for S_{Q^*} the score at Q^* in the $R \times (R-1)$-dimensional parametric model with known emissions and unknown Q. Specifically, S_Q takes the form

$$S_{Q^*}(r,s)(Y_{-\infty:k}) = \frac{\mathbb{P}_*(X_{k-1}=r, X_k = s|Y_{-\infty:k})}{Q^*_{rs}} - \frac{\mathbb{P}_*(X_{k-1}=r, X_k = R|Y_{-\infty:k})}{Q^*_{rR}}$$
$$+ \sum_{i=-\infty}^{k-1} \left\{ \frac{\mathbb{P}_*(X_{i-1}=r, X_i = s|Y_{-\infty:k}) - \mathbb{P}_*(X_{i-1}=r, X_i = s|Y_{-\infty:k-1})}{Q^*_{rs}} \right.$$
$$\left. - \frac{\mathbb{P}_*(X_{i-1}=r, X_i = R|Y_{-\infty:k}) - \mathbb{P}_*(X_{i-1}=r, X_i = R|Y_{-\infty:k-1})}{Q^*_{rR}} \right\}, \tag{2.12}$$

where we index with $(r,s) \in [R] \times [R-1]$ for convenience sake, but consider S_{Q^*} as a vector of length $R(R-1)$.

We then rewrite (2.10), substituting also the expression in (2.11) as

$$\Delta^{(a,\mathbf{h})}_{k,\infty} = a^T S_{Q^*} + \sum_{r=1}^{R} \left\{ \mathbb{P}_*(X_k = r|Y_{-\infty:k}) h_r(Y_k) \right.$$
$$\left. + \sum_{i=-\infty}^{k-1} (\mathbb{P}_*(X_i = r|Y_{-\infty:k}) - \mathbb{P}_*(X_i = r|Y_{-\infty:k-1})) h_r(Y_i) \right\}. \tag{2.13}$$

To set notation for what follows, we will denote the r^{th} summand of the above display, for a given direction h_r, as

$$H_r(h_r)(Y_{-\infty:k}) := \mathbb{P}_*(X_k = r|Y_{-\infty:k}) h_r(Y_k)$$
$$+ \sum_{i=-\infty}^{k-1} (\mathbb{P}_*(X_i = r|Y_{-\infty:k}) - \mathbb{P}_*(X_i = r|Y_{-\infty:k-1})) h_r(Y_i). \tag{2.14}$$

The above discussion means that, through a Taylor expansion, we can write

$$\ell_n(\theta_n((a,\mathbf{h}))) - \ell_n(\theta_*) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\infty} \Delta^{(a,\mathbf{h})}_{k,\infty} - \frac{1}{2n} \left(-\frac{d^2}{dt^2}|_{t=t_1} \ell_n^{(a,\mathbf{h})}(t) \right)$$
$$= \Delta_{n,(a,\mathbf{h})} - \frac{1}{2}\|(a,\mathbf{h})\|^2_{\mathcal{H}} + o_{\mathbb{P}_*}(1),$$

with $t_1 \in (0, n^{-1/2})$ and by choosing $\|(a,\mathbf{h})\|^2_{\mathcal{H}}$ the Fisher information at $t=0$ in the model $\Theta_{a,\mathbf{h}}$, which is defined in [32] as the L^2 norm[1] of $\Delta^{(a,\mathbf{h})}_{0,\infty}$. We have $\Delta_{n,(a,\mathbf{h})} = \frac{1}{\sqrt{n}} \sum_{k=1}^{\infty} \Delta^{(a,\mathbf{h})}_{k,\infty}$, is linear in $(a,\mathbf{h})$ and satisfies $\Delta_{n,(a,\mathbf{h})} \to N(0, \|(a,\mathbf{h})\|^2_{\mathcal{H}})$, by the discussion directly preceding Theorem 2 in [32]. We also used above the local uniform conver-

[1] The L^2 norms of all $\Delta^{(a,\mathbf{h})}_{k,\infty}$ coincide by stationarity.

gence of the second derivative of the score to the Fisher information matrix at $t = 0$ in $\Theta_{a,\mathbf{h}}$, which is guaranteed by Theorem 3 of [32].

The preceding discussion shows that our model is LAN, which is the first step in understanding efficient estimators of a parameter of interest. In our case, the parameter of interest will be $v_n(P_{n,\theta_n(a,\mathbf{h})}) = Q$, which has 'derivative' $\dot{v}(a, \mathbf{h}) = a \in \mathbb{R}^p$ in the sense of [32], with $p = R(R-1)$.

Following what precedes, we apply the convolution theorem also described in [78], and originally proven in [101]. This theorem essentially states that the limiting law of a regular estimator is lower bounded by that of a Gaussian random variable, whose covariance matrix is the covariance of the *efficient influence function*, which is itself characterised by the tangent space and the parameter 'derivative' $\dot{v}$.

In Section 2.12.2, we detail the application of this theorem; the arguments are similar to those used in the iid setting, see for instance Section 25 of [100]. This essentially involves identifying influence functions $\dot{v}_b^T$ for estimation of one-dimensional functions $b^T Q$ at $b^T Q^*$, as we vary $b \in \mathbb{R}^p$. By considering the elements for which $a = 0$, we first find that $\dot{v}_b^T$ is in the orthogonal complement (with respect to the LAN norm on $\mathcal{H}$) of the linear span of the scores in the models $\Theta_{0,\mathbf{h}}$, which is the span of the $H_r(h_r)$ as we vary r and $h_r \in \mathcal{H}_r$. Write $\mathcal{A}$ for the projection onto this space, and write

$$\tilde{S}_Q = S_Q - \mathcal{A} S_Q \tag{2.15}$$

for the projection of the score function S_Q onto the orthogonal complement of this space, which we call the efficient score. By making further standard arguments, again detailed in Section 2.12.2, we find that the influence function $\dot{v}_b^T$ has variance $b^T \tilde{J}^{-1} b$, where $\tilde{J}$ is the covariance matrix of the efficient score $\tilde{S}_{Q^*}$, which is the efficient information matrix. By varying b, this characterises the optimal limiting covariance matrix as $\tilde{J}^{-1}$, the inverse efficient information matrix.

In the following section, we show that $\tilde{J}_M$ converges to $\tilde{J}$ when M goes to infinity and that, if M_n is a sequence going to infinity slowly enough, the sequence of estimators $(\hat{Q}_{n,M_n})_{n\geq 1}$ is asymptotically efficient and the posterior distributions $\Pi_{M_n}(dQ|Y_{1:n})$ satisfy the efficient Bernstein-von Mises theorem.

2.3.2.2 Approximation by the models $\mathcal{M}(\mathcal{I}_M)$

To prove that $\tilde{J}_M$ converges to $\tilde{J}$, we require some additional assumptions on the true generating process and the partitions:

Assumption 2.4. For all $r = 1, \dots, R$, the densities f_r^* are continuous, (Lebesgue) almost-everywhere positive, linearly independent and

$$\forall r, s \leq R \quad 0 < \inf_x \frac{f_r^*(x)}{f_s^*(x)} < \sup_x \frac{f_r^*(x)}{f_s^*(x)} < \infty.$$

We also consider a sequence of partitions with vanishing radius.

Assumption 2.5. The sequence $(\mathcal{I}_M)_{M \geq R}$ of partitions is embedded, so each $I \in \mathcal{I}_M$ is the union of some collection $I' \in \mathcal{I}_{M+1}$, and moreover

$$\max_{m \leq \kappa_M} \mathrm{Vol}(I_m^{(M)}) \to 0, \qquad \text{as } M \to \infty,$$

where $\mathrm{Vol}(A)$ denotes the Lebesgue measure of A.

Assumption 2.5 is directly under the control of the practitioner, and is verified for instance by considering nested dyadic partitions, as is done in Section 2.5. **Assumption** 2.4 is an assumption on the data generating process, although is very common in the context of semiparametric efficiency. Note also that, under **Assumptions** 2.4 and 2.5, there exists $M_0 > 0$ for which $\mathcal{I}_{M_0}$ is admissible for $\mathbf{f}^*$.

In Proposition 2.5, we show that the efficient Fisher information matrix is precisely the limit of $\tilde{J}_M$. To do this, we first define the efficient scores $\tilde{S}_{Q^*}^{(M)}$ for the histogram models, whose covariance matrix will be $\tilde{J}_M$, analogously to what was done in Section 2.3.2.1 in the context of the full semiparametric model.

Recall, from Remark 2.2, that the likelihood in the model with piece-wise constant emissions corresponds to a likelihood of a HMM with multinomial emission distributions and observations $Y_i^{(M)} = (\mathbb{1}_{I_1^{(M)}}(Y_i), \cdots, \mathbb{1}_{I_{\kappa_M}^{(M)}}(Y_i))$, and so we emphasize in our notation that the scores in $\mathcal{M}(\mathcal{I}_M)$ depend only on these summaries. A straightforward adaptation of the earlier presentation then leads us to define the score functions for Q in $\mathcal{M}(\mathcal{I}_M)$ by

$$S^{(M)}_{Q^*}(r,s)(Y^{(M)}_{-\infty:k}) = \frac{\mathbb{P}_*(X_{k-1}=r, X_k=s|Y^{(M)}_{-\infty:k})}{Q^*_{rs}} - \frac{\mathbb{P}_*(X_{k-1}=r, X_k=R|Y^{(M)}_{-\infty:k})}{Q^*_{rR}} + \sum_{i=-\infty}^{k-1} \left\{ \frac{\mathbb{P}_*(X_{i-1}=r, X_i=s|Y^{(M)}_{-\infty:k}) - \mathbb{P}_*(X_{i-1}=r, X_i=s|Y^{(M)}_{-\infty:k-1})}{Q^*_{rs}} - \frac{\mathbb{P}_*(X_{i-1}=r, X_i=R|Y^{(M)}_{-\infty:k}) - \mathbb{P}_*(X_{i-1}=r, X_i=R|Y^{(M)}_{-\infty:k-1})}{Q^*_{rR}} \right\}. \tag{2.16}$$

for $(r,s) \in [R] \times [R-1]$. For the model $\mathcal{I}_M$, the perturbations on the emissions vary in

$$\mathcal{H}_{r,M} := \{h = \sum_{m=1}^{\kappa_M} \alpha_m \mathbb{1}_{I_m}(y),\ \alpha_m \in \mathbb{R},\ \int h f^*_{r,M}(y)dy = 0\}.$$

and we then define, for $h_r \in \mathcal{H}_{r,M}$,

$$H_{r,M}(h_r)(Y^{(M)}_{-\infty:k}) = \mathbb{P}_*(X_j = r|Y^{(M)}_{-\infty:k})h_r(Y_k) + \sum_{i=-\infty}^{k-1} \left(\mathbb{P}_*(X_i = r|Y^{(M)}_{-\infty:k}) - \mathbb{P}_*(X_i = r|Y^{(M)}_{-\infty:k-1})\right) h_r(Y_i). \tag{2.17}$$

Following again the arguments of Section 2.3.2.1, we write $\mathcal{A}_M$ for the projection onto the space spanned by the $H_{r,M}(h_r)$, and finally define

$$\tilde{S}^{(M)}_{Q^*} = S^{(M)}_{Q^*} - \mathcal{A}_M S^{(M)}_{Q^*},$$

whose covariance matrix is $\tilde{J}_M$, the efficient Fisher information matrix for the model $\mathcal{M}(\mathcal{I}_M)$.

The convergence of $\tilde{S}^{(M)}_{Q^*}$ to $\tilde{S}_{Q^*}$ is then established by a martingale argument. To show convergence of the projection operators $\mathcal{A}_M$ to $\mathcal{A}$ is more involved; we use a deconvolution argument which shows that boundedness in the space of the nuisance scores $H_r(h_r)$ implies boundedness of the h_r in the index space $\mathcal{H}_r$ (and likewise for $H_{r,M}(h_r), \mathcal{H}_{r,M}$). These intermediate arguments, and the proof of the following results, are in Section 2.6.

Proposition 2.5. *Grant* ***Assumptions*** *2.1, 2.3-2.5. Then*

$$\tilde{S}^{(M)}_{Q^*} \to \tilde{S}_{Q^*}, \quad M \to \infty$$

where the convergence is in $L^2(\mathbb{P}_*)$. *Moreover, we have* $\tilde{J}_M \to \tilde{J}$ *as* $M \to \infty$ *where* $\tilde{J} \geq \tilde{J}_M$

is the efficient information for estimating Q in the full data model, and is invertible.

From Proposition 2.5, we deduce the following result.

Theorem 2.6. *Let $\hat{Q}_n = \hat{Q}_n^{(M_n)}$ where $M_n \to \infty$ sufficiently slowly. Then under **Assumptions** 2.1, 2.3-2.5, $\hat{Q}_n$ is a $\mathbb{P}_*$-regular estimator and satisfies (up to label-swapping)*

$$\sqrt{n}(\hat{Q}_n - Q^*) \Rightarrow^* N(0, \tilde{J}^{-1})$$

*where $\tilde{J}$ is the variance of the efficient score function, as defined in Proposition 2.5. In particular, $\hat{Q}_n$ is an efficient estimator of Q in the full semiparametric model. Moreover, for Π_n a sequence of priors placing its mass on models $\mathcal{M}(\mathcal{I}_{M_n})$ respectively, and satisfying **Assumption** 2.2, we have (up to label-swapping) that*

$$\left\| \Pi_n(\sqrt{n}(Q - \hat{Q}_n)|Y_{1:n}) - \mathcal{N}(0, \tilde{J}^{-1}) \right\|_{TV} = o_{\mathbb{P}_*}(1).$$

2.4 Cut posterior contraction

In what follows, we study contraction rates for the cut posterior $\Pi_{cut}(\cdot|Y_{1:n})$ defined in Section 2.2.

2.4.1 Concentration of marginal densities $g_{Q,\mathbf{f}}^{(\ell)}$

In this section, we present Proposition 2.7 which controls $\Pi_{cut}(\|g_{Q,\mathbf{f}}^{(\ell)} - g_{Q^*,\mathbf{f}^*}^{(\ell)}\| \leq \epsilon_n|Y_{1:n})$. This result follows from Theorem 2.27, which is an adaptation of the general approach of [48] and is of independent interest; see Section 2.9 for full details.

For the sake of simplicity, we consider a prior in the form $\Pi_2(d\mathbf{f}|Q) = \Pi_2(d\mathbf{f})$; extension to the case where the prior Π_2 depends on Q is straightforward from Section 2.9. Note that, in any case, the conditional posterior on $\mathbf{f}$ given Q depends on Q through the likelihood.

Hence, similarly to [102], we consider the following assumptions used to verify the Kullback-Leibler condition.

Assumption 2.6. Let $\epsilon_n, \tilde{\epsilon}_n > 0$ denote two sequences such that $\tilde{\epsilon}_n \leq \epsilon_n = o(1)$ and $n\tilde{\epsilon}_n^2 \to \infty$. Assume that the prior Π_2 on $\mathcal{F}^R$ satisfies the following conditions:

A. There exists $C_{\Pi_2} > 0$, depending on the choice of prior Π_2 and a sequence $S_n \subset \mathcal{F}^R$, such that $\Pi_2(S_n) \gtrsim \exp\left(-C_{\Pi_2} n\tilde{\epsilon}_n^2\right)$ and such that, for all $\mathbf{f} \in S_n$, there exists a set $S \subset \mathcal{Y}$ and functions $\tilde{f}_1, \ldots, \tilde{f}_R$ satisfying, for all $1 \leq i \leq R$ and for $C_{R,Q} = 4 + \log(\frac{2R}{\min_{ij} Q_{ij}^*}) +$

$10^4 \frac{R^2}{(\min_{ij} Q^*_{ij})^5}$,

$$\int_S \left[\frac{|f_i^*(y) - f_i(y)|^2}{f_i^*(y)} + \frac{\left|f_i^*(y) - \tilde{f}_i(y)\right|^2}{\tilde{f}_i(y)} \right] \lambda(dy) \leq \tilde{\epsilon}_n^2;$$

$$\int_S f_i^*(y) \max_{1 \leq j \leq R} \log\left(\frac{\tilde{f}_j(y)}{f_j(y)}\right) \lambda(dy) \leq \tilde{\epsilon}_n^2;$$

$$\int_{S^c} f_i^*(y) \max_{1 \leq j \leq R} \log\left(\frac{f_j^*(y)}{f_j(y)}\right) \lambda(dy) \leq C_{R,Q} \tilde{\epsilon}_n^2;$$

$$\int_{S^c} [\tilde{f}_i(y) + f_i^*(y)] \lambda(dy) \leq \tilde{\epsilon}_n^2.$$

B. For all constants $C > 0$, there exists a sequence $(\mathcal{F}_n)_{n \geq 1}$ of subsets of $\mathcal{F}^R$ and a constant $C' > 0$ such that

$$\Pi_2\left(\mathcal{F}_n^c\right) = o\left(\exp\left(-Cn\tilde{\epsilon}_n^2\right)\right), \quad \text{and} \quad N\left(\frac{\epsilon_n}{12}, \mathcal{F}_n, d\right) \lesssim \exp\left(C' n \epsilon_n^2\right),$$

where $N\left(\frac{\epsilon_n}{12}, \mathcal{F}_n, d\right)$ is the $\frac{\epsilon_n}{12}$-covering number of $\mathcal{F}_n$ with respect to $d(f, \tilde{f}) = \max_{i=1,\ldots,R} \|f_i - \tilde{f}_i\|_{L^1}$.

Proposition 2.7. *Let $(Y_t)_{t \geq 1}$ be observations from a finite state space HMM with transition matrix Q and emission densities $\mathbf{f} = (f_r)_{r=1,\ldots,R}$. Grant **Assumptions** 2.1-2.5, and consider the cut posterior obtained by choosing Π_1 as the prior Π_M of Theorem 2.4 associated to the admissible partition $\mathcal{I}_M$, and Π_2 such that **Assumption** 2.6 is verified for suitable $\epsilon_n, \tilde{\epsilon}_n$ satisfying $n\epsilon_n^2 \gtrsim \log n$. Then, for any $K_n \to \infty$,*

$$\Pi_{cut}(\{(Q, f) : \|g^{(3)}_{Q,\mathbf{f}} - g^{(3)}_{Q^*,\mathbf{f}^*}\|_{L_1} > K_n \epsilon_n\} | Y_n) = o_{\mathbb{P}_*}(1),$$

up to label-swapping.

Remark 2.8. Given Proposition 2.7, we may use the results of [102] to derive posterior contraction rates for a number of priors Π_2. We explore the case that Π_2 is a (product of) Dirichlet process mixture of normals in Section 2.4.4, as this is a popular choice for density estimation.

Remark 2.9. We could also choose $\Pi_1 = \Pi_{1,n}$ equal to Π_{M_n} for some $M_n \to \infty$ sufficiently slowly to give the refined control over the marginal cut posterior on Q as described in Theorem 2.6.

2.4.2 Concentration of emission distributions

Proposition 2.7 applies the general contraction result of Theorem 2.27 to obtain an estimation result for the marginal distribution of the observations. Given that we already have control over the transition matrix in this setting when using Π_1 as a histogram prior, it remains to establish concentration rates for the emission distributions. Theorem 2.10 allows us to translate a rate on a marginal distribution into a corresponding rate on the emission distribution.

Theorem 2.10. *Let the HMM satisfy the assumptions of Proposition 2.7, and let $g^{(3)}_{Q^*,f}$ be the marginal density for three consecutive observations under the parameters $(Q^*, \mathbf{f})$. Then there exists a constant $C = C(f^*, Q^*) > 0$ such that, for sufficiently small $\|g^{(3)}_{Q^*,f} - g^{(3)}_*\|_{L_1}$,*

$$\sum_{r=1}^{R} \|f_r - f_r^*\|_{L^1} \leq C \|g^{(3)}_{Q^*,f} - g^{(3)}_*\|_{L_1}.$$

Remark 2.11. Theorem 2.10 provides an inversion inequality from L^1 to L^1, which has interest in both Bayesian and frequentist estimation of emission densities. It is the first such result for the L^1 distance, with Theorem 6 of [26] establishing a similar inequality in the L^2 case. Given that the testing assumptions of Theorem 2.27 (and more generally, of results based on [48]) are much more straightforward to verify for the L^1 distance, our result has particular interest in the context of Bayesian (or pseudo-Bayesian) settings.

Using Proposition 2.7 and Theorem 2.10 together with Theorem 2.4, we easily deduce the following.

Theorem 2.12. *Let $Y_{1:n}$ be distributed according to the HMM with parameters Q^* and $\mathbf{f}^*$, and grant **Assumptions** 2.1-2.5. Let $\Pi_1 = \Pi_M$ and let Π_2 satisfy **Assumption** 2.6. Then, for ϵ_n as in Proposition 2.7 and any $K_n \to \infty$, we have (up to label-swapping)*

$$\Pi_{cut}(\{\|Q - Q^*\| > \frac{K_n}{\sqrt{n}}, \max_{r=1,\dots,R} \|f_r - f_r^*\|_{L^1} > K_n \epsilon_n\} | Y_{1:n}) = o_{\mathbb{P}_*}(1),$$

and

$$\|\Pi_{cut}(\sqrt{n}(Q - \hat{Q}_{n,M}) \in (\cdot) | Y_{1:n}) - N(0, \tilde{J}_M^{-1})\|_{TV} = o_{\mathbb{P}_*}(1).$$

2.4.3 Concentration of smoothing distributions

When clustering data, the *smoothing distribution* $\mathbb{P}_*(X_k = x | Y_{1:n})$ is often of interest. Our final main result concerns recovery of these probabilities using a (cut) Bayesian approach,

establishing contraction of the posterior distribution over these smoothing distributions in total variation. We do so by combining novel arguments with the inequality given in Proposition 2.2 of [27]. We recall the notation $\theta = (Q, \mathbf{f})$.

Theorem 2.13. *Grant the assumptions of Theorem 2.12, together with*

$$\max_{r\in[R]} \|\sqrt{f_r^*}\|_1 < \infty, \tag{2.18}$$

and assume that

$$\Pi_2(\max_i \|f_i\|_{L_2} > e^{\gamma n\epsilon_n^2}|Q) \leq C' e^{-2\gamma n\epsilon_n^2}; \qquad \max_i \|f_i^*\|_{L_2} \leq e^{\gamma n\epsilon_n^2}; \tag{2.19}$$

for some constants $\gamma, C' > 0$. Then, for ϵ_n as in Proposition 2.7 for which $n\epsilon_n^3 \to 0$ and for any $K_n \to \infty$, we have (up to label swapping)

$$\Pi_{cut}(\|\mathbb{P}_\theta(X_k = x|Y_{1:n}) - \mathbb{P}_*(X_k = x|Y_{1:n})\|_{TV} > K_n\epsilon_n|Y_{1:n}) = o_{\mathbb{P}_*}(1).$$

Remark 2.14. The requirement that $n\epsilon_n^3 \to 0$ is used in the proof, although we expect that this condition is not fundamental and it may be possible to weaken with appropriate proof techniques. It is not clear if assumption (2.19) is crucial, although it is anyway very weak and satisfied, for instance, by choosing a prior Π_2 for which $E_{\Pi_2}(\|f_r\|_{L^2}^2) < \infty$ for all r.

2.4.4 Example: Dirichlet process mixtures of Gaussians

In this section, we show that **Assumption** 2.6 is verified when Π_2 is a Dirichlet process mixture of Gaussians. We use the results of Section 4 of [102], in which **Assumption** 2.6 is verified for $\mathbf{f}^* = (f_r^*)_{r=1}^R$, with each f_r^* in the class

$$\begin{aligned}\mathcal{P}(\beta, L, \gamma) = \{f \in \mathcal{F} : &\log f \text{ is locally } \beta - \text{Hölder with derivatives } l_j = (\log f)^{(j)} \\ &\text{and } |l_{k_\beta}(y) - l_{k_\beta}(x)| \leq r!L(y)|y-x|^{\beta-k_\beta} \text{ when } |x-y| \leq \gamma\},\end{aligned}$$

and satisfying weak tail assumptions which we give in **Assumption** 2.8. Here $\beta > 0$, L is a polynomial function, $\gamma > 0$ and $k_\beta = \lceil\beta\rceil - 1$ for $\lceil\cdot\rceil$ the usual ceiling operator.

The following result, which is a corollary of Theorem 4.3 of [102], shows that, for prior

choices satisfying **Assumption** 2.9, the conditions of Theorem 2.12 are verified with

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}} \log(n)^t, \tag{2.20}$$

where $t > t_0 \geq \left(2 + \frac{2}{\gamma} + \frac{1}{\beta}\right)\left(\frac{1}{\beta} + 2\right)^{-1}$.

Corollary 2.15. *Let $Y_{1:n}$ be distributed according to the HMM with parameters Q^* and $\mathbf{f}^*$, and grant **Assumptions** 2.1-2.5. Let $\Pi_1 = \Pi_M$ and let the prior Π_2 on $\mathcal{F}^R$ take the form of an $R-$fold product of Dirichlet process mixtures of Gaussians, in which the base measure α and variance prior Π_σ satisfy **Assumption** 2.9. Suppose further that the true emission distributions $(f_i^*)_{i=1}^R$ satisfy **Assumption** 2.8 and that, for each i, $f_i^* \in \mathcal{P}(\beta, L, \gamma)$. Then Theorems 2.12 and 2.13 hold with ϵ_n as in (2.20), where $t > t_0 \geq \left(2 + \frac{2}{\gamma} + \frac{1}{\beta}\right)\left(\frac{1}{\beta} + 2\right)^{-1}$.*

Proof. The proof follows from [102] who showed that **Assumption** 2.6 is verified under the stated assumptions with $\lambda(dy)$ the Lebesgue measure, appropriately chosen B_n, $\tilde{\epsilon}_n = n^{-\frac{\beta}{2\beta+1}} \log(n)^{t_0}$ with t_0 as in the statement, and ϵ_n as in the statement. □

The rate $n^{-\frac{\beta}{2\beta+1}}$ is minimax optimal for the classes $\mathcal{P}(\beta, L, \gamma)$ under iid sampling assumptions, see [77]. Although not proved here, we strongly believe the rate ϵ_n to be minimax up to $\log n$ factors, since an iid sampling assumption corresponds to estimating $f_1, \ldots, f_R$ after observing $X_{1:n}, Y_{1:n}$ which is easier than estimating $f_1, \ldots, f_R$ from $Y_{1:n}$ only. We thank the anonymous reviewer for pointing this out to us.

Remark 2.16. It is also possible to verify the conditions of Proposition 2.7 in the case of a countable observation space, again by following the example set out in [102]. In this case, **Assumption** 2.6 is verified with $\lambda(dy)$ the counting measure on $\mathbb{N}$ by using a Dirichlet process prior on the emissions whose base measure satisfies a tail condition, and under a tail assumption on the true emissions. In this case, a parametric rate up to log factors is obtained.

2.5 Practical considerations and simulation study

In this section, we discuss the practical implication of the method and results described in the previous sections[2]. As described in Section 2.2.2, we consider a cut posterior approach where Π_1 is based on a histogram prior on the emission distributions and where Π_2 is

[2]Code is provided at https://github.com/dm1729/HMMs.

a Dirichlet process mixture of normals for the emission densities. We first describe the implementation of $\Pi_1(Q|Y_{1:n})$.

2.5.1 MCMC algorithm for Q

We first define the construction of the partition $\mathcal{I}_M$. In Section 2.3, it is defined as a partition on $[0,1]$ (or more generally on $[0,1]^d$), but this is easily generalized to a partition of $\mathbb{R}$ (or $\mathbb{R}^d$) via an appropriate C_1 diffeomorphism ϕ. In this section we restrict ourselves to dyadic partitions, i.e. $I_m^{(M)} = (\phi^{-1}((m-1)2^{-M}), \phi^{-1}(m2^{-M}))$, $m = 0, \cdots, 2^M - 1$, where $\phi : \mathbb{R} \longrightarrow [0,1]$ is a C_1 diffeomorphism.

Recall that, for Theorems 2.3 and 2.4 to be valid, we merely require that $\mathcal{I}^{(M)} = (I_m^{(M)}, m \leq 2^M - 1)$ is admissible and that, for all $r \leq R, m \leq 2^M - 1$, $F_r^*(I_m^{(M)}) > 0$. Interestingly Theorem 2.6 does not restrict the choice of ϕ, but restricts the choice of M. In practice, however, the choice of M and ϕ matters and more details are provided below.

Once $\mathcal{I}_M$ is chosen, we consider a prior on $Q, \underline{\omega}_M$ (suppressing the M henceforth) and, for the sake of computational simplicity, we consider the following family of Dirichlet priors:

$$\begin{aligned} \forall i \leq R, \quad Q_{i\cdot} &= (Q_{ij}, j \in [R]) \sim \mathcal{D}(\gamma_{i1}, \cdots, \gamma_{iR}) \\ \omega_{i\cdot} &= (\omega_{im}, m \in [2^M]) \sim \mathcal{D}(\beta_{i1}, \cdots, \beta_{i2^M}). \end{aligned}$$

We then use a Gibbs sampler on $(Q, \underline{\omega}, \mathbf{X})$ where given $\mathbf{X}, Y_{1:n}$,

$$\begin{aligned} Q_{i\cdot} &\overset{ind}{\sim} \mathcal{D}(\gamma_{i1} + n_{i1}, \cdots, \gamma_{iR} + n_{iR}), \quad n_{ij} = \sum_{t=2}^{n} \mathbb{1}_{X_{t-1}=i, X_t=j} \\ \omega_{i\cdot} &\overset{ind}{\sim} \mathcal{D}(\beta_{i1} + N_0^{(i)}, \cdots, \beta_{i2^M} + N_{2^M-1}^{(i)}), \quad N_m^{(i)} = \sum_{t=1}^{n} \mathbb{1}_{Y_t \in I_m^{(M)}, X_t=i}, \end{aligned}$$

and the conditional distribution of X given $Q, \underline{\omega}, Y$ is derived using the forward - backward algorithm (see [76] or [39]). To overcome the usual label-switching issue in mixtures and HMMs, we take the approach of Chapter 6 of [76] which deals with MCMC in the mixtures setting, in which the authors propose relabelling relative to the posterior mode as a post-processing step (with likelihood computed also with forward-backward).

In our simulation study, we have considered $R = 2$ hidden states with transition matrix

and emission distributions

$$Q^* = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}, \quad F_1^* \sim \mathcal{N}(-1, 1), F_2^* \sim \mathcal{N}(1, 1).$$

We have considered data of size $n = 1000, 2500, 5000, 10000$, obtained by restricting a single simulated data set of size 10000. To study the effect of M we run our MCMC algorithm targeting $\Pi_1(\cdot|Y_{1:n})$ with $\kappa_M = 2, 4, 8, 16, 64, 128$ bins.

We took $\gamma_{ij} = \beta_{ij} = 1$ so that independent uniform priors over the simplex were used for rows of the transition matrix and the histogram weights, and chose

$$\phi(y) = \begin{cases} (1 + e^{-y})^{-1} & |y| > 3 \\ \zeta + \eta y & |y| \leq 3 \end{cases}$$

with $\zeta, \eta \in \mathbb{R}$ chosen so that ϕ is continuous. The sigmoid $y \mapsto (1 + e^{-y})^{-1}$ is a popular transformation from $\mathbb{R}$ to $[0, 1]$. However, its steep gradient near $y = 0$ means that a uniform partition of $[0, 1]$ is mapped to a partition of $\mathbb{R}$ concentrated tightly around $y = 0$. This in turn leads to an over-coarsening of the data, with many data points lying in the two unbounded bins of the partition which extend to $\pm\infty$. The linear interpolation in the range $|y| \leq 3$ is thus intended to provide more discrimination between data points lying in this interval, which includes the vast majority (approx 97.5%) of the data.

We ran the MCMC for 150000 iterations for each binning, discarding 10000 iterations as burn-in and retaining one in every twenty of the remaining draws, for a total of 7000 posterior draws.

The fitted distributions for Q under the priors Π_1, with varying M, are shown in Figures 2.2, 2.3, 2.4 and 2.5, and demonstrate Theorems 2.4 and 2.6 as we detail below. To make an additional comparison with a typical Bayesian nonparametric approach, we also fitted a model with a prior Π', which uses Dirichlet priors on the transition matrix as in Π_1 and Dirichlet process mixtures of Gaussians to model the emission densities. We defer further details of Π' to Section 2.5.2 as it is more relevant as a comparison with Π_2, since both have a much higher computational cost in comparison to Π_1. In Figure 2.6, we compare distributions for the transition matrix under $\Pi'(\cdot|Y_{1:n})$ with the distribution under $\Pi_1(\cdot|Y_{1:n})$ for a selection of values of M, but we emphasize the large difference in compute time.

Even when $\kappa_M = 2$, the posterior distribution under the prior Π_1, as n increases, looks increasingly like a Gaussian, though its variance is quite large. For slightly larger values of κ_M, this Gaussian shape is preserved but with lower variance, demonstrating Proposition 2.24 which states that the Fisher information grows as we refine the partition. However, we can also see that taking κ_M too large leads to erroneous posterior inference, which may simply be biased (for instance when $n = 5000$ and we take $\kappa_M = 64$) or lose its Gaussian shape entirely (for instance when $n = 1000$ and $\kappa_M \in \{64, 128\}$). This demonstrates that the requirement of Theorem 2.6, that $M_n \to \infty$ sufficiently slowly, has practical consequences and is not merely a theoretical artefact.

While the latter issue (of a non-Gaussian shaped posterior) is easily diagnosed by eye, the former issue (of a biased, Gaussian-like posterior) is somewhat more worrisome, as the bias cannot be so easily identified when one does not have access to the true data generating process. In Figure 2.6, we see preliminary empirical evidence that this problem of bias may also be present when taking a fully Bayesian approach based on the prior Π', indicating that our approach may work better than a fully Bayes approach if M_n can be tuned well. We emphasize however that this simulation study is very limited in scope, and is only intended to demonstrate our results rather than to make conclusive comparisons.

For the histogram prior Π_1, we propose the following heuristic to tune κ_M after computing the posterior for a range of different numbers of bins. For a small number of bins (say $\kappa_M = 4$), take the posterior mean as a reference estimate (say $\hat{Q}_0$), which should have low bias and moderate variance. For higher κ_M values, compare $\hat{Q}_0$ κ_M-specific posterior mean and the $1 - \alpha$ credible sets C_α, for $\alpha = 0.05, 0.1$ say. We consider that κ_M is not too large if $\hat{Q}_0$ is well within the bounds of C_α. We have added some further markings to Figure 2.4 which illustrate this approach.

We emphasize that it is most important not to refine the partition too quickly, so even when adopting this heuristic one may wish to favour lower values of κ_M. When fitting Π_2 as discussed in Section 2.5.2, we used posterior draws based on one possible refinement, in which $\kappa_M = 4$ for $n = 1000$, $\kappa_M = 8$ for $n \in \{2500, 5000\}$ and $\kappa_M = 16$ for $n = 10000$ - see Figure 2.7.

2.5.2 MCMC algorithm for $\{f_r\}_{r=1}^R$

After simulating draws from the marginal posterior Π_1 of Q, we then use a Gibbs sampler to target the cut-posterior distribution $\Pi_2(f|Q, Y)$, for which the prior $\Pi_2(f|Q)$ is a Dirichlet

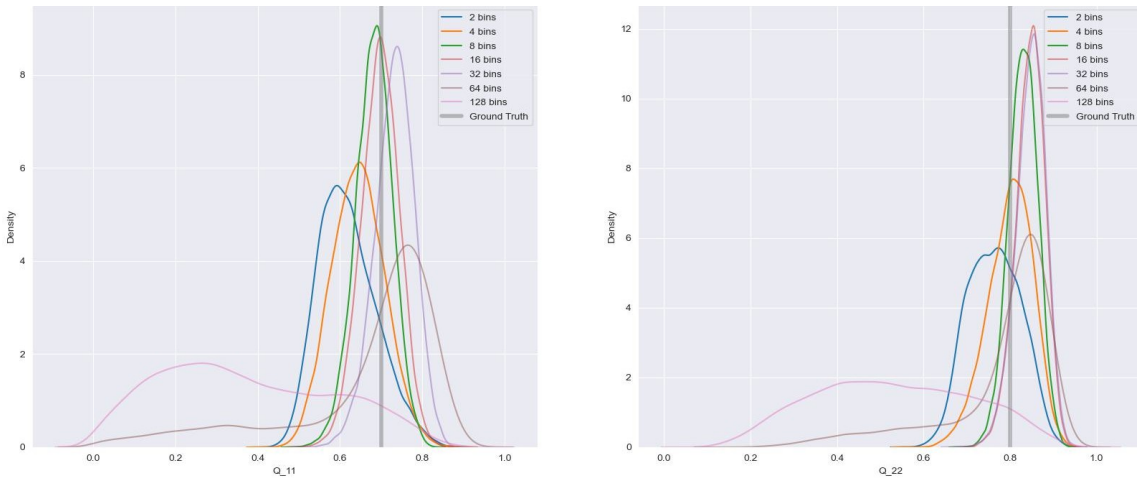


Figure 2.2: Posterior draws for Q under priors Π_1 with $\kappa_M \in \{2, 4, \ldots, 128\}$ bins when $n = 1000$.

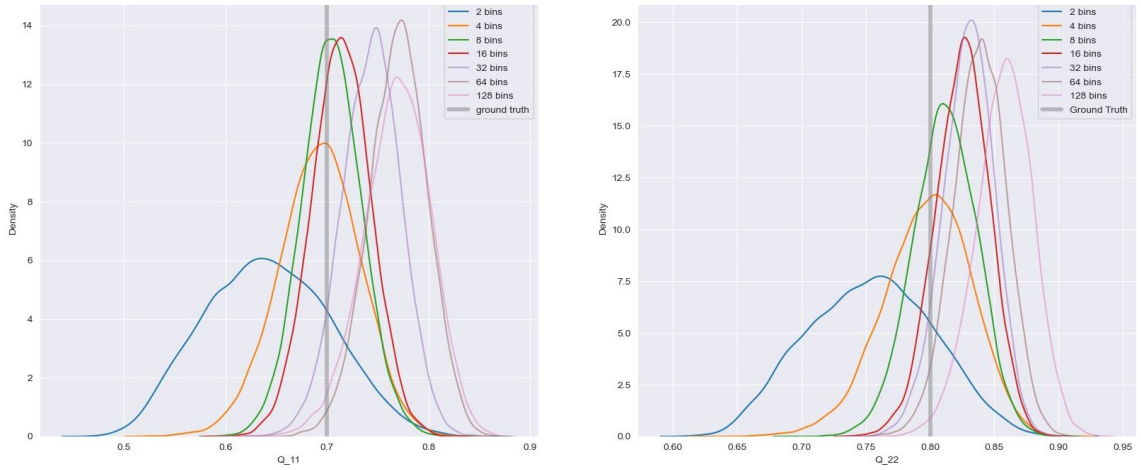


Figure 2.3: Posterior draws for Q under priors Π_1 with $\kappa_M \in \{2, 4, \ldots, 128\}$ bins when $n = 2500$.

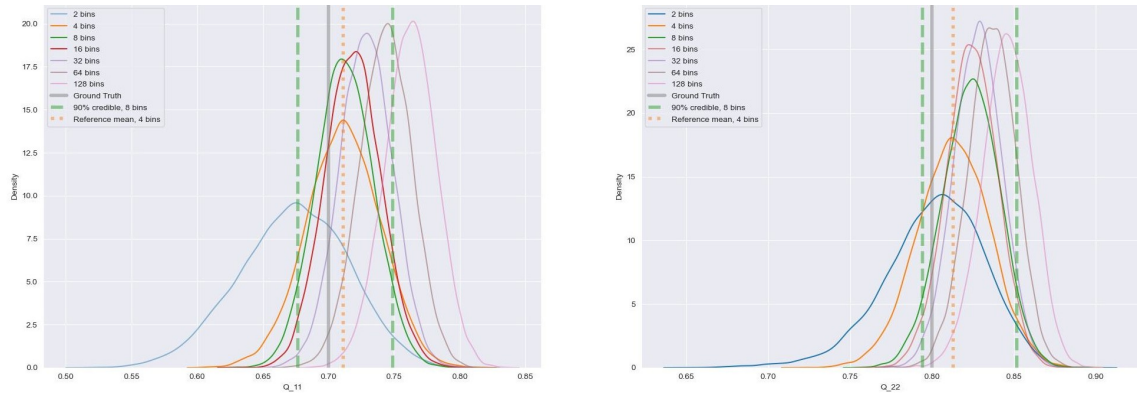


Figure 2.4: Posterior draws for Q under priors Π_1 with $\kappa_M \in \{2, 4, \ldots, 128\}$ bins when $n = 5000$. The dotted orange line is the reference mean, taken when $\kappa_M = 4$. The dashed green lines are the bounds of the 90% credible set when $\kappa_M = 8$.

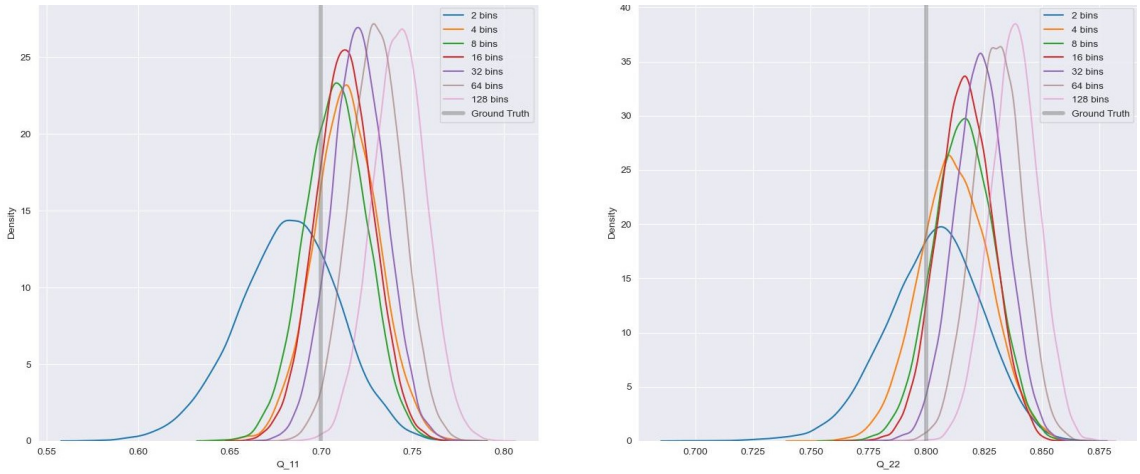


Figure 2.5: Posterior draws for Q under priors Π_1 when $n = 10000$, when using $\kappa_M \in \{2, 4, \ldots, 128\}$ bins.

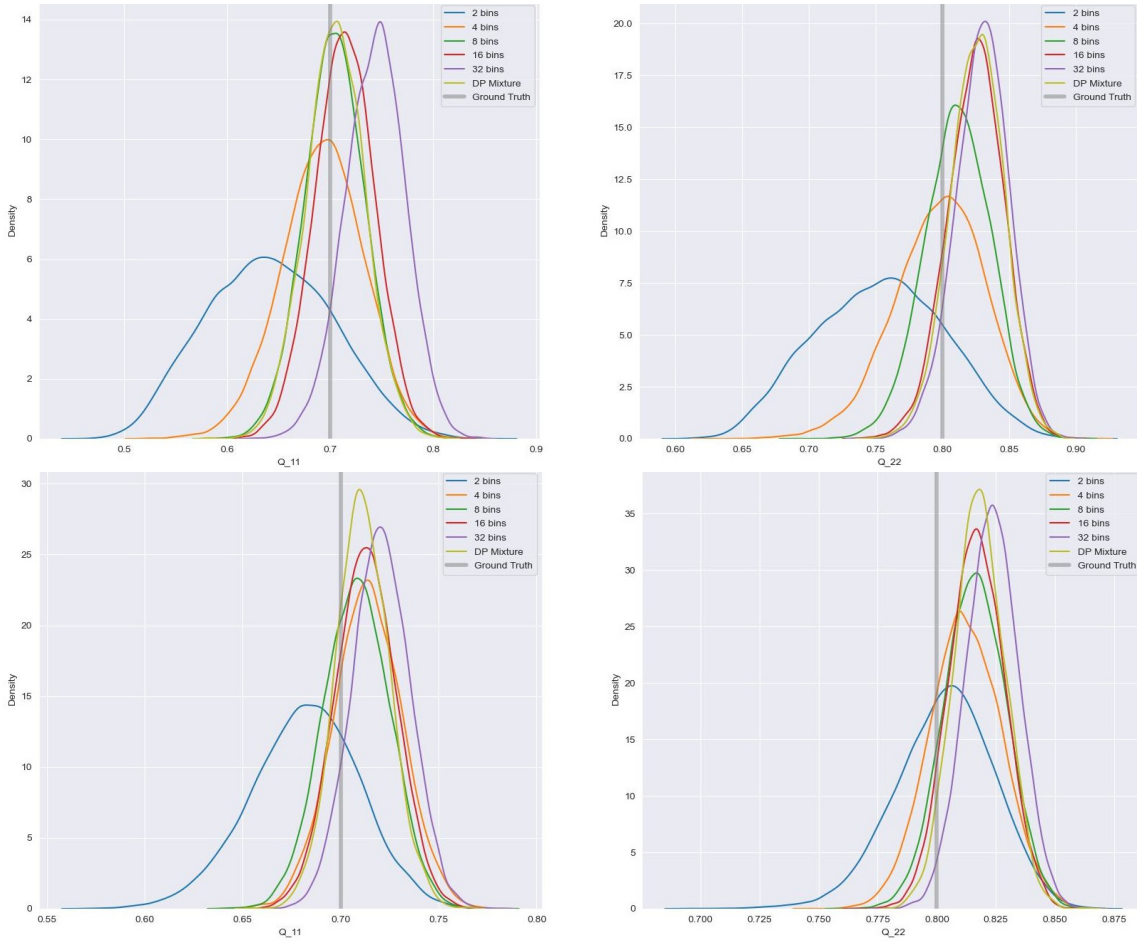


Figure 2.6: Posterior draws for Q under priors Π_1 (for $\kappa_M \in \{2, 4, 8, 16, 32\}$) and Π' (DP Mixture line) when $n = 2500$ (top) and $n = 10000$ (bottom).

mixture as follows. For each hidden state r (whose labels are fixed after relabelling in Section 2.5.1), we use independent dirichlet Process mixtures of normals to model f_r:

$$f_r(y) = \int \phi_{\sigma^{(r)}}(y-\mu)P^{(r)}(d\mu), \quad P^{(r)} \overset{iid}{\sim} DP(M_0, \mathcal{N}(\mu_c, \sigma_c^2)), \tag{2.21}$$

with $M_0 = 1$, $\mu_c = 0$ and $\sigma_c^2 = 1$. The kernels ϕ_σ are centred normal distributions with variance $v = \sigma^2$, and $v^{(r)}$ is equipped with an InvGamma$(\alpha_\sigma, \beta_\sigma)$ prior[3]. Note that the scale parameter σ is fixed across μ's for each r. The MCMC procedure will rely on the following stick breaking representation of (2.21) involving mixture allocation variables $s^{(r)} = (s_i^{(r)})_{i\leq n}$ and stick breaks $(V_j^{(r)})_{j\in\mathbb{N}}$ associated to a vector of weights $(W_j^{(r)})_{j\in\mathbb{N}}$:

$$\begin{aligned} Y_i|\mathbf{X}, \mu, s, \sigma, W &\overset{iid}{\sim} \phi_{\sigma^{(r)}}(\cdot - \mu_{s_i}^{(X_i)}); \quad s_i|V, \mathbf{X}, \mu, \sigma^2 \overset{indep}{\sim} W^{(X_i)}, \\ W_j^{(r)} &= V_j^{(r)} \prod_{i<j}(1-V_i^{(r)}); \quad \mathbf{X} \sim \text{MarkovChain}(Q) \\ V_j^{(r)} &\overset{iid}{\sim} \beta(1, M_0); \quad \mu_j^{(r)} \sim \mathcal{N}(\mu_c, \sigma_c^2); \quad v^{(r)} \sim \text{InvGamma}(\alpha_\sigma, \beta_\sigma). \end{aligned} \tag{2.22}$$

In order to approximately sample from the corresponding posterior, we replace the above prior with the Dirichlet-Multinomial process (see Theorem 4.19 of [49]) with truncation level $S_{\max} = \lfloor \sqrt{n} \rfloor$, in which W is instead sampled from a Dirichlet distribution with parameter

$$\alpha = \left(\frac{M_0}{S_{\max}}, \frac{M_0}{S_{\max}}, \cdots, \frac{M_0}{S_{\max}} \right). \tag{2.23}$$

This truncation level is suggested as a rule of thumb in the remark at the end of Section 5.2 of [49]. Further details on implementation can be found in Section 2.13.

Nested MCMC For the i^{th} draw from the cut posterior, we would ideally sample first Q_i from $\Pi_1(\cdot|Y)$, and then $f_i = (f_{ir})_r$ from $\Pi_2(\cdot|Q_i, Y)$. For the first step, we used Algorithm SA1 with burn in and thinning. However, we encounter difficulty when simulating from $\Pi_2(\cdot|Q_i, Y)$, as the Q_i changes when i changes, and so an MCMC approach effectively needs to mix i by i. In order to achieve this, we run a nested MCMC approach (see [87]) in which, for each i, an interior chain of length C is run, taking the final draw from the interior chain as our i^{th} global draw.

[3]We remark that Proposition 2.15 verifies the conditions of Proposition 2.7 under an inverse gamma prior on the standard deviation. We have instead adopted in our simulations the common practice of placing an inverse gamma prior on the variance for computational convenience.

Concerns about the computational cost of nested MCMC have been expressed, especially when there is strong dependence between the two modules (in our case, representing the transition and emissions) as in Section 4.2 of [72]. However, we found little improvement beyond $C = 10$ interior iterations, see Figure 2.8. We expect that this is down to the well localised posterior $\Pi_1(\cdot|Y)$ on Q which means that the Q_i, and hence the targets $\Pi_2(\cdot|Q_i, Y)$, don't vary too much i by i. Plots of the posterior mean when $C = 10$ are detailed in Figure 2.7. Since we used the thinned draws of Q from Π_1, we ran 7000 such interior chains.

When using such a small number of iterations for the interior chain, nested MCMC is not costly compared to other cut sampling schemes (see e.g. Table 1 of [72]). We further suggest that the computational cost compared to fully Bayesian approaches with fixed targets is not as high as it may seem, given that the use of such interior chains should, at least partially, subsume the need for thinning. Indeed running such interior chains with a fixed target for each i would be precisely the same as thinning. A more detailed development of this idea can be found in Section 4.4 of the supplement to [20].

Comparison to fully Bayesian approach As mentioned in Section 2.5.1, we also consider the fully Bayesian model with prior Π' as a means of comparison. The prior Π' independently places a Dirichlet prior on the rows of the transition matrix as discussed beforehand Π_1, as well as a Dirichlet process mixture prior over the emission densities as discussed earlier in this section for Π_2.

In the implementation of Π', we ran the MCMC for 70000 iterations, discarding the first 10000 as burn-in and thinning at a rate of one in ten observations. This was chosen so that we had an approximate matching with the computational cost of 7000 iterations of the cut posterior, each with 10 interior iterations. We plot the resulting emission densities in Figure 2.9. In comparison to Figure 2.7, the pointwise bands seem to capture the ground truth more accurately. We remark however that Theorem 2.12 only provides guarantees on L^1 concentration and so the plots will not entirely reflect the theory; L^1 credible sets are rather less easily visualised. We also emphasize our earlier comment from Section 2.5.1, that the simulation study is limited in scope, and is not intended to provide conclusive comparisons with other approaches.

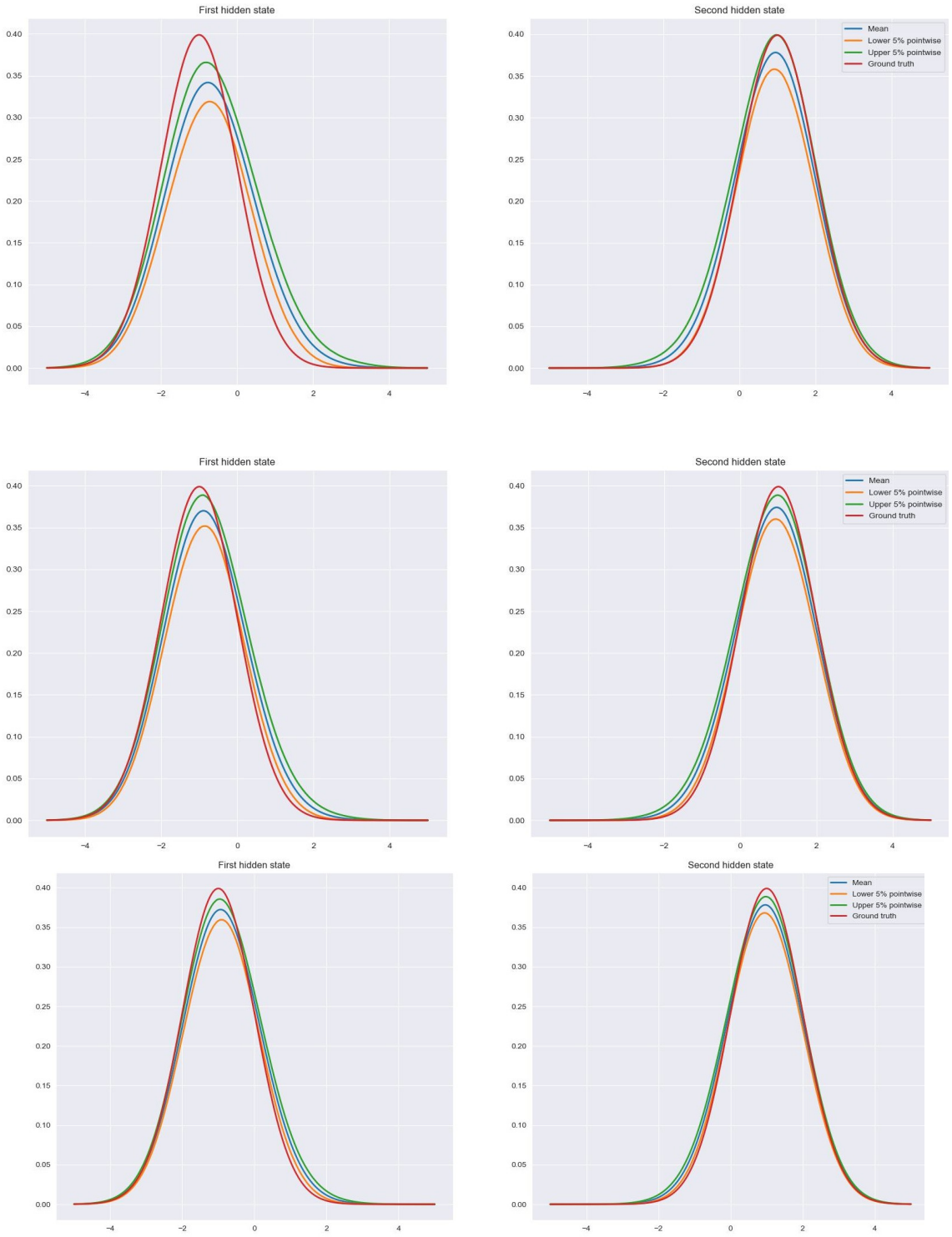


Figure 2.7: Cut posterior mean and pointwise 90% credible bands for emissions when refining the partition with $n = 2500$ (top), $n = 5000$ (middle) and $n = 10000$ (bottom), each with $C = 10$ interior iterations. See top plot of Figure 2.8 for $n = 1000$.

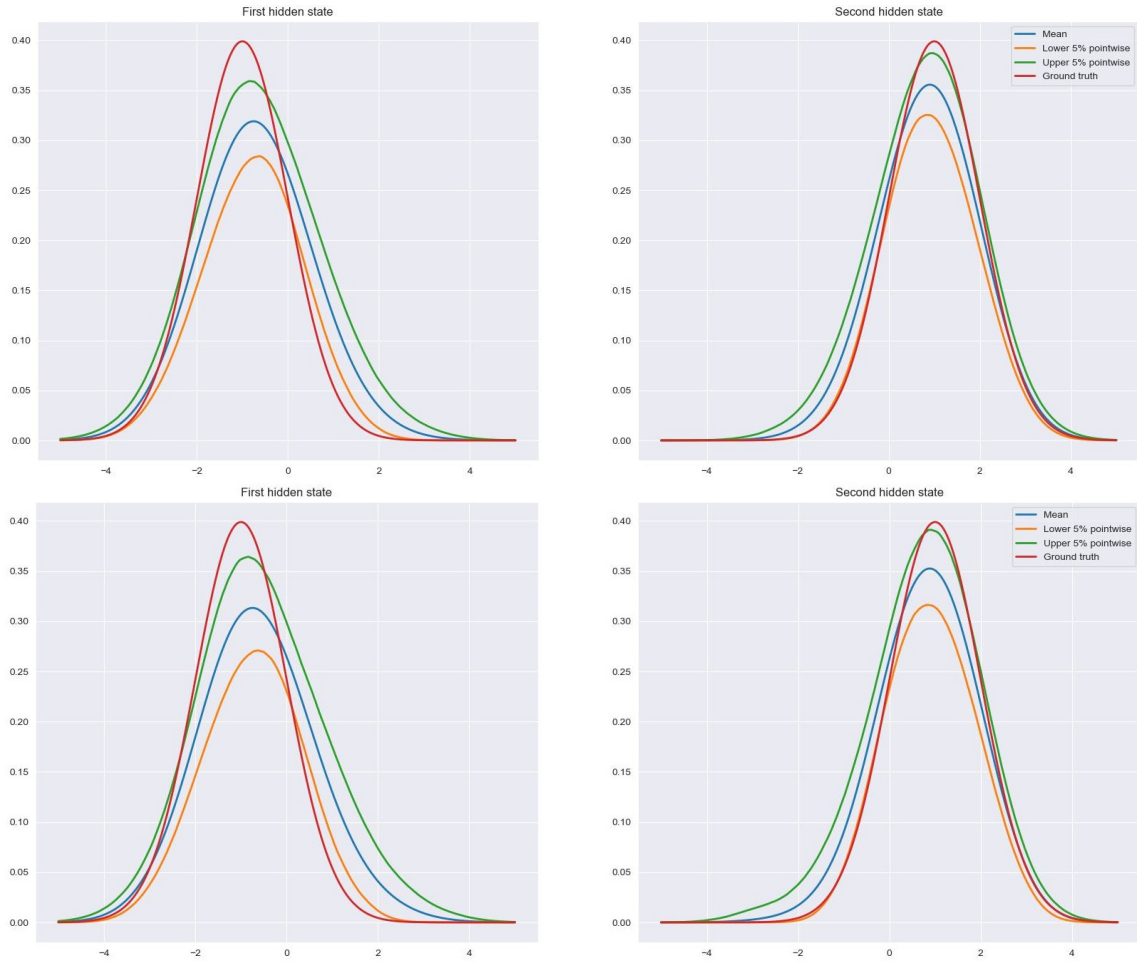


Figure 2.8: Cut posterior mean and pointwise 90% credible bands for emissions with $n = 1000$ and four bins. Top: $C = 10$ interior iterations; Bottom: $C = 100$ interior iterations.

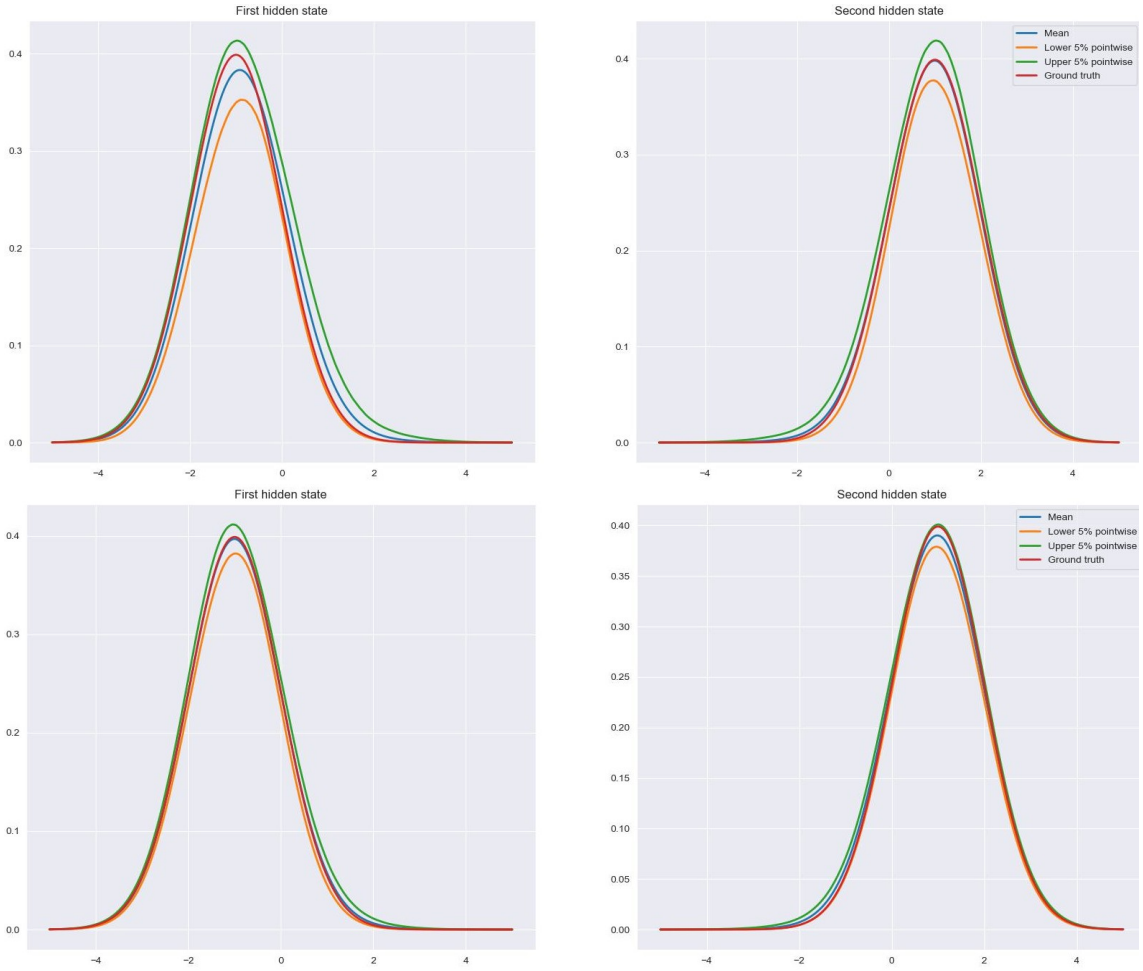


Figure 2.9: Posterior mean and pointwise 90% credible bands for emissions, with $n = 2500$ (top) and $n = 10000$ (bottom) under fully Bayesian posterior with prior Π'.

2.6 Proofs of main results

We first prove Proposition 2.5.

2.6.1 Proof of Proposition 2.5: Convergence of the scores

Throughout this section, **Assumptions** 2.1-2.5 are assumed to hold.

Recall that

$$\mathcal{H}_r = \{h \in L^2(f_r^* dx), \int h(x) f_r^* dx = 0\},$$

and

$$\mathcal{H}_{r,M} = \{h = \sum_{m=1}^{\kappa_M} \alpha_m \mathbb{1}_{I_m}(y),\ \alpha_m \in \mathbb{R},\ \int h f_{r,M}^*(y) dy = 0\}.$$

Define $\mathcal{P}$ to be the space spanned by the nuisance scores $\{H(h) : h = (h_r)_{r\leq R}, h_r \in \mathcal{H}_r\}$ in the full semiparametric model, and also $\mathcal{P}_M$ to be the space spanned by the nuisance scores $\{H_M(h_M) : h_M = (h_{r,M})_{r\leq R}, h_{r,M} \in \mathcal{H}_{r,M}\}$ in the histogram models $\mathcal{M}(\mathcal{I}_M)$. Here $H(h) = (H_r(h_r) : r \in [R])$ with H_r as in Equation (2.14), and also $H_M(h_M) = (H_{r,M}(h_r) : r \in [R])$ with $H_{r,M}$ as in Equation (2.17). Denote also $\mathcal{G}_M^\ell = \sigma(Y_t^{(M)} : t \leq \ell)$ and $\mathcal{G}^\ell = \sigma(Y_t : t \leq \ell)$, for $\ell \in \mathbb{Z}$.

To prove Proposition 2.5, we prove convergence of sequences of nuisance scores in $\mathcal{P}_M$ to nuisance scores in $\mathcal{P}$, and also convergence of the score functions $S_{Q^*}^{(M)}$ to S_{Q^*}. Recall that the scores S_{Q^*} in the model with known emissions are given by (2.12), with the scores $S_{Q^*}^{(M)}$ associated to data $Y^{(M)}$ being given by (2.16).

Proof of Proposition 2.5. Write $\mathcal{A}_M$ for the orthogonal projection onto $\mathcal{P}_M$ and $\mathcal{A}$ for the orthogonal projection onto $\mathcal{P}$. Define

$$\tilde{S}_{Q^*}^{(M)} := S_{Q^*}^{(M)} - \mathcal{A}_M S_{Q^*}^{(M)} \quad \text{, and} \quad \tilde{S}_{Q^*} := S_{Q^*} - \mathcal{A} S_{Q^*}.$$

To prove Proposition 2.5, it is enough to prove that $\tilde{S}_{Q^*}^{(M)}$ converges to $\tilde{S}_{Q^*}$. A major difficulty here is to prove the convergence of $\mathcal{A}_M S$ to $\mathcal{A}S$ in $L^2(\mathbb{P}_*)$, with $S \in L^2(\mathbb{P}_*)$, because, contrary to the mixture model case considered in [45], the sets $\mathcal{P}_M$ are not embedded.

In Lemma 2.19, we prove using a martingale argument that $S_{Q^*}^{(M)}$ converges to S_{Q^*} in

$L^2(\mathbb{P}_*)$ as M goes to infinity. Then

$$\begin{aligned}\|\tilde{S}^{(M)}_{Q^*} - \tilde{S}_{Q^*}\|_{L_2} &\leq \|S^{(M)}_{Q^*} - S_{Q^*}\|_{L_2} + \|\mathcal{A}_M(S_{Q^*} - S^{(M)}_{Q^*})\|_{L^2} + \|(\mathcal{A} - \mathcal{A}_M)S_{Q^*}\|_{L^2} \\ &= \|(\mathcal{A} - \mathcal{A}_M)S_{Q^*}\|_{L^2} + o(1).\end{aligned}$$

We now prove that $\|(\mathcal{A} - \mathcal{A}_M)S\|_{L^2} = o(1)$, for all $S \in L^2(\mathbb{P}_*)$. As mentioned earlier, the non trivial part of this proof comes from the fact that the sets $\mathcal{P}_M$ are not embedded. Hence we first prove that the set $\mathcal{P}_M$ *is close* to the set $\tilde{\mathcal{P}}_M$, where

$$\tilde{\mathcal{P}}_M = \{H(h_M),\, h_M = (h_{r,M})_{r\leq R}, h_{r,M} \in \mathcal{H}_{r,M}\}.$$

More precisely, we prove in Lemma 2.17 below that $\mathcal{A}_M S - \tilde{\mathcal{A}}_M S = o_{L^2}(1)$, where $\tilde{\mathcal{A}}_M$ is the projection onto $\tilde{\mathcal{P}}_M$. Since the partitions are nested, the $\tilde{\mathcal{P}}_M$ are nested; reasoning as in [45], we then obtain that $(\tilde{\mathcal{A}}_M S)_{M\in\mathbb{N}}$ is Cauchy, and so converges to some $\tilde{\mathcal{A}}S$. Then, arguing as in [45], we identify $\tilde{\mathcal{A}}S$ with $\mathcal{A}S$ in Lemma 2.21. The argument involves first showing that $\tilde{\mathcal{A}}$ is a projection onto a subspace of $\mathcal{P}$, and then showing that elements of $\mathcal{P}$ are well approximated by elements of $\tilde{\mathcal{P}}$, by approximating the corresponding $h_r \in L^2_0(f^*_r)$ by histograms $(h_{r,M})_{M\in\mathbb{N}}$. This terminates the proof of Proposition 2.5. □

Lemma 2.17. *For any element $S \in L^2(\mathbb{P}_*)$,*

$$|\mathcal{A}_M S - \tilde{\mathcal{A}}_M S|_2 = o(1).$$

Proof. As will appear in the proof, a key step is Lemma 2.18 which says that, for any sequence h_M bounded in L^2, if $H_M(h_M) = o_{L^2}(1)$, then $h_M = o_{L^2}(1)$. Lemma 2.20 also applies for L^2 bounded sequences, and implies that, for any bounded sequence h_M, $(H_M - H)(h_M) = o_{L^2}(1)$.

To prove that $\mathcal{A}_M S - \tilde{\mathcal{A}}_M S = o_{L^2}(1)$, we decompose

$$\mathcal{A}_M S - \tilde{\mathcal{A}}_M S = \mathcal{A}_M S - \tilde{\mathcal{A}}_M \mathcal{A}_M S + \tilde{\mathcal{A}}_M \mathcal{A}_M S - \tilde{\mathcal{A}}_M S.$$

We first prove that $\tilde{\mathcal{A}}_M \mathcal{A}_M S - \mathcal{A}_M S = o_{L^2}(1)$, and then that $\tilde{\mathcal{A}}_M(\mathcal{A}_M S - S) = o_{L^2}(1)$. Define h_M such that $\mathcal{A}_M S = H_M(h_M)$; we have that $\|H_M(h_M)\|_2 \leq \|S\|_2$. Also, by Lemma 2.20 and using the fact that orthogonal projections onto a space minimize the L^2

distance to that space (which includes $H(h_M)$),

$$\|\tilde{\mathcal{A}}_M\mathcal{A}_M S - \mathcal{A}_M S\|_2 \leq \|(H - H_M)(h_M)\|.$$

Suppose now that, along a subsequence, $\|h_M\|_2 \to \infty$. Then, along this subsequence, $\frac{H_M(h_M)}{\|h_M\|_2} = H_M(\frac{h_M}{\|h_M\|_2}) \to 0$ in L^2 and we can apply Lemma 2.18 to conclude that $\frac{h_M}{\|h_M\|_2} \to 0$ in L^2 along this subsequence, when in fact this subsequence has constant unit L^2 norm. We thus get a contradiction and conclude that $\sup_M \|h_M\|_2$ is bounded. Then, by Lemma 2.20, we have that $\mathcal{A}_M S = H_M(h_M) = H(h_M) + o_{L_2}(1)$. Thus $\|(H_M - H)(h_M)\|_2 \to 0$ and $\|\tilde{\mathcal{A}}_M\mathcal{A}_M S - \mathcal{A}_M S\|_2 = o(1)$.

We now control $\tilde{\mathcal{A}}_M(\mathcal{A}_M S - S)$. Considering $\tilde{\mathcal{A}}_M(S - \mathcal{A}_M S) = H(\tilde{h}_M)$ for some $\tilde{h}_M = (\tilde{h}_{r,M}, r \in [R])$ and $\tilde{h}_{r,M} \in L_0^2(f_r^*)$ in model M, we have

$$\begin{aligned}\|\tilde{\mathcal{A}}_M S - \tilde{\mathcal{A}}_M\mathcal{A}_M S\|_{L^2}^2 &= \langle S - \mathcal{A}_M S, \tilde{\mathcal{A}}_M(S - \mathcal{A}_M S)\rangle \\ &= \langle S - \mathcal{A}_M S, H(\tilde{h}_M)\rangle = \underbrace{\langle S - \mathcal{A}_M S, H_M(\tilde{h}_M)\rangle}_{=0 \text{ as } (S-\mathcal{A}_M S)\perp H_M(\tilde{h}_M)} + R_M,\end{aligned}$$

where

$$|R_M| \leq \|S\|_2\|(H - H_M)(\tilde{h}_M)\|_2 \to 0,$$

with the convergence following as before using Lemma 2.18 and the boundedness of $\|H(\tilde{h}_M)\|_{L_2}$ by $\|S\|_{L_2}$. Thus $\tilde{\mathcal{A}}_M(\mathcal{A}_M S - S) = o_{L_2}(1)$ and Lemma 2.17 is proved. □

Lemma 2.18. *Fix $r \in \{1, \ldots, R\}$. Let $h_M = (h_{r,M}, r \in [R])$ be a sequence of step functions on the partitions of model $\mathcal{M}(\mathcal{I}_M)$ with $\|h_M\| = \sum_r \|h_{r,M}\|_{L^2} \leq 1$ and $h_{r,M} \in \mathcal{H}_{r,M}$. Then*

$$H_M(h_M) \overset{L^2(\mathbb{P}_*)}{\to} 0 \implies h_M \overset{L^2(\mathbb{P}_*)}{\to} 0.$$

Proof. Assuming $H_M(h_M) \to 0$, Lemma 2.22 implies that, for all $k \geq 0$,

$$D_k(h_M) := \sum_{r=1}^{R}\sum_{k'=-k}^{0} h_{r,M}(y_{k'})\mathbb{P}_*(X_{k'} = r|Y_{-k:0}) = o_{L_2}(1).$$

For shortness sake we write $g^* = g^{(2)}_{Q^*,\mathbf{F}^*}$ and we will abuse notation and write $y_j := y_{-j}$ for $j \in \mathbb{N}$, so y_1 is y_{-1} etc. Expanding $D_1(h_M)$, we thus have that

$$\sum_r h_{r,M}(y_0)f_r^*(y_0)\sum_s \frac{p_s^* Q_{sr}^* f_s^*(y_1)}{g^*(y_0, y_1)} + \sum_r \frac{h_{r,M}(y_1)f_r^*(y_1)}{g^*(y_0, y_1)} p_r^* \sum_s f_s^*(y_0)Q_{rs}^* \overset{L^2(\mathbb{P}_*)}{\longrightarrow} 0.$$

Since, for any function H,

$$\mathbb{E}_* \left[\frac{H^2(Y_0, Y_1)}{g^*(Y_0, Y_1)^2} \right] = \int_{[0,1]^2} \frac{H^2(y_0, y_1)}{g^*(y_0, y_1)} dy_1 dy_0 \geq \frac{\|H\|_2^2}{\|g^*\|_\infty},$$

and, since $\|g^*\|_\infty \leq \max_r \|f_r^*\|_\infty < \infty$ by continuity, we obtain

$$\sum_r h_{r,M}(y_0) f_r^*(y_0) \sum_s p_s^* Q_{sr}^* f_s^*(y_1) + \sum_r h_{r,M}(y_1) f_r^*(y_1) p_r^* \sum_s f_s^*(y_0) Q_{rs}^* \overset{L^2(dy^2)}{\to} 0. \quad (2.24)$$

Note further that the family of functions of y_1 given by $y_1 \mapsto (\sum_s p_s^* Q_{sr}^* f_s^*(y_1))_{r=1}^k$ is linearly independent. Consider now a partition $\bar{I}_1, \bar{I}_2, \ldots, \bar{I}_R$ of $[0,1]$ such that

$$F_{rj} = \int_{\bar{I}_j} f_r^*(y_1) dy_1, \quad r, j \in \{1, \ldots R\},$$

defines a matrix $F = (F_{rj}^*)$ of rank R. Such a partition always exists by Lemma 2.45. Define also

$$A_{rj} = \int_{I_j} \sum_s p_s^* Q_{sr}^* f_s^*(y_1) dy_1, \quad r, j \in \{1, \ldots R\},$$

with $A = (A_{rj})$, define $D_p = \mathrm{diag}(p_1^*, \ldots, p_R^*)$ (which is invertible under **Assumption** 2.3) and recall $Q^* = (Q_{sr}^*)$. Then we can write $A^T = F^T D_p Q^*$ and note that A also has rank R. Now, integrating the y_1 coordinate out of Equation (2.24), we get, for all $j = 1, \ldots k$,

$$\sum_r \underbrace{h_{r,M}(y_0) f_r^*(y_0)}_{\tilde{h}_{r,M}} A_{rj} = -\sum_r \underbrace{\int_{\bar{I}_j} h_{r,M}(y_1) f_r^*(y_1) p_r^* dy_1}_{B_M(r,j)} \sum_s f_s^*(y_0) Q_{rs}^* + o_{L^2(dy_0)}(1).$$

Let $\tilde{h}_M = (\tilde{h}_{r,M})_{r=1}^R$, $h_M = (h_{r,M})_{r=1}^R$ and $B_M = (B_M(r,j))_{rj}$. Then the preceding display may be rewritten as

$$\tilde{h}_M(y_0) = \underbrace{-(A^T)^{-1} B_M^T Q^*}_{B_{0,M}} \mathbf{f}^*(y_0) + o_{L^2(dy_0)}(1). \quad (2.25)$$

Now, from its definition, $(B_M)_M$ is a bounded sequence in a finite dimensional space and so converges to some B along a subsequence. Working on this subsequence, we then get

the limit

$$\tilde{h}_M(y_0) \overset{L^2(dy_0)}{\to} \underbrace{-(A^T)^{-1}B^TQ^*}_{B_0}\mathbf{f}^*(y_0) := \tilde{h}(y_0).$$

Consider now $D_2(h_M)$, which again vanishes under our assumptions and by Lemma 2.22, and take the limit in M along the subsequence where we just established convergence. As before, abuse notation by replacing Y_{-k} by Y_k for $k \in \mathbb{N}$. Expanding and multiplying through by joint marginals, we get at the limit

$$\begin{aligned}
&\sum_r (B_0\mathbf{f}^*(y_0))_r \sum_{s_1,s_2} p^*_{s_2} Q^*_{s_2 s_1} Q^*_{s_1 r} f^*_{s_2}(y_2) f^*_{s_1}(y_1) \\
&+ \sum_r (B_0\mathbf{f}^*(y_1))_r \sum_{s_1,s_2} p^*_{s_2} Q^*_{s_2 r} Q^*_{r s_1} f^*_{s_1}(y_0) f^*_{s_2}(y_2) \\
&\quad + \sum_r (B_0\mathbf{f}^*(y_2))_r \sum_{s_1,s_2} p^*_r Q^*_{r s_1} Q^*_{s_1 s_2} f^*_{s_1}(y_1) f^*_{s_2}(y_0) = 0. \qquad (2.26)
\end{aligned}$$

Rewriting (2.26) into matrix products, we obtain

$$\begin{aligned}
&\sum_{s_1} [\mathbf{f}^*(y_2)^T D_p Q^*]_{s_1} f^*_{s_1}(y_1) (Q^* B_0 \mathbf{f}^*(y_0))_{s_1} \\
&+ \sum_{s_1} [B_0\mathbf{f}^*(y_1)]_{s_1} (Q^*\mathbf{f}^*(y_0))_{s_1} (\mathbf{f}^*(y_2)^T D_p Q^*)_{s_1} \\
&\quad + \sum_{s_1} [\mathbf{f}^*(y_2)^T B_0^T D_p Q^*]_{s_1} f^*_{s_1}(y_1) (Q^*\mathbf{f}^*(y_0))_{s_1} = 0.
\end{aligned}$$

By linear independence of the the set of functions $\{f^*_s(y_1)\}_{s=1,\dots,R}$, we have that, for all s_1, the coefficients vanish so that, for all s_1, we get

$$\begin{aligned}
&[\mathbf{f}(y_2)^T D_p Q^*]_{s_1} (Q^* B_0 \mathbf{f}(y_0))_{s_1} + \sum_{r'} B_{0 r' s_1} (Q^*\mathbf{f}^*(y_0))_{r'} (\mathbf{f}^*(y_2)^T D_p Q^*)_{r'} \\
&\quad + [\mathbf{f}^*(y_2)^T B_0^T D_p Q^*]_{s_1} (Q^*\mathbf{f}^*(y_0))_{s_1} = 0.
\end{aligned}$$

We can now repeat the argument, instead viewing the expression as a linear combination of the linearly independent set of functions $\{f^*_s(y_2)\}_{s=1,\dots,R}$. Expanding what precedes for

each s_1, we get

$$\sum_r f_r^*(y_2)p_r^*Q_{rs_1}^*(Q^*B_0\mathbf{f}^*(y_0))_{s_1} + \sum_{r,r'} B_{0r's_1}(Q^*\mathbf{f}^*(y_0))_{r'}(f_r^*(y_2))p_r^*Q_{rr'}^* + \sum_r f_r^*(y_2)(B_0^TD_pQ^*)_{rs_1}(Q^*\mathbf{f}^*(y_0))_{s_1} = 0,$$

so that, for each r, s_1 we get that

$$p_r^*Q_{rs_1}^*(Q^*B_0\mathbf{f}^*(y_0))_{s_1} + \sum_{r'} B_{0r's_1}(Q^*\mathbf{f}^*(y_0))_{r'}p_r^*Q_{rr'}^* + (B_0^TD_pQ^*)_{rs_1}(Q^*\mathbf{f}^*(y_0))_{s_1} = 0.$$

Expanding in terms of the linearly independent functions $\{(Q^*\mathbf{f}^*)_r(y_0)\}_{r=1,\ldots,R}$, we get

$$\sum_{r'} p_rQ_{rs_1}^*(Q^*B_0Q^{*-1})_{s_1r'}(Q^*\mathbf{f}^*)_{r'}(y_0) + \sum_{r'} B_{0r's_1}(Q^*\mathbf{f}^*(y_0))_{r'}p_r^*Q_{rr'}^* + \sum_{r'} \mathbb{1}_{s_1=r'}(B_0^TD_pQ^*)_{rr'}(Q^*\mathbf{f}^*(y_0))_{r'} = 0,$$

which then gives, for each r, r', s_1, that

$$p_r^*Q_{rs_1}^*(Q^*B_0Q^{*-1})_{s_1r'} + B_{0r's_1}p_r^*Q_{rr'}^* + \mathbb{1}_{s_1=r'}(B_0^TD_pQ^*)_{rr'} = 0.$$

In particular, setting $s_1 = r'$ we get that, for all r, r',

$$p_r^*Q_{rr'}^*(Q^*B_0Q^{*-1})_{r'r'} + B_{0r'r'}p_r^*Q_{rr'}^* + (B_0^TD_pQ^*)_{rr'} = 0.$$

By **Assumption** 2.3, $Q_{ij}^* > 0\ \forall i, j$ and so we can divide through to get

$$\frac{(B_0^TD_pQ^*)_{rr'}}{Q_{rr'}^*} = p_r^*a(r'),$$

for some a depending on r' alone. Write D_a for the diagonal matrix whose diagonals are the $a(r')$. Then

$$(B_0^TD_pQ^*)_{rr'} = p_r^*Q_{rr'}^*a(r') = (D_pQ^*D_a)_{rr'}.$$

Equivalently, we can write $B_0^T = (D_pQ^*)D_a(D_pQ^*)^{-1}$. Recalling the earlier definition of B_0, we had

$$B_0 = -(A^T)^{-1}B^TQ^*,$$

with each column of B expressible as the integral over $\bar{I}_j$, with respect to y_1, of $D_p\tilde{h}(y_1) =$

$D_p B_0 \mathbf{f}^*(y_1)$, and thus expressible as $B = D_p B_0 F$. We can then substitute into the previous display to see that

$$B_0 = -(A^T)^{-1} F^T B_0^T D_p Q^*.$$

The expression of A as $A^T = F^T D_p Q^*$ allows us to write

$$B_0 = -Q^{*-1} D_p^{-1} (F^T)^{-1} F^T B_0^T D_p Q^* = -Q^{*-1} D_p^{-1} B_0^T D_p Q^* = -D_a.$$

This means that $B_0^T = -D_a = (D_p Q^*) D_a (D_p Q^*)^{-1}$. So $Q^{-1} D_a Q^* = -D_a$ which gives

$$a(r) Q^*_{rs} = -Q^*_{rs} a(s).$$

Taking $r = s$ and dividing through by Q^*_{rr} gives $a_r = 0$ for all r, hence $B_0 = 0$ and $\tilde{h} = 0$. This means $\tilde{h}_m(y_0) \to 0$ in $L^2(dy_0)$ and so $h_{r,M} \to 0$ in $L^2(f_r(y_0)dy_0)$, and so in $L^2(\mathbb{P}_*)$ under **Assumption** 2.4, which is what we wanted to show. □

Lemma 2.19. *For $S^{(M)}_{Q^*}, S_{Q^*}$, as in (2.16) and (2.12), and for any choice of j, we have that*

$$S^{(M)}_{Q^*}(Y^{(M)}_{-\infty:j}) \to S_{Q^*}(Y_{-\infty:j})$$

in $L^2(\mathbb{P}_*)$.

Proof. Without loss of generality we take $j = 0$. We first show that we can, uniformly in M, truncate the series in (2.16) to an arbitrary degree of accuracy. We prove that, uniformly in M,

$$\sum_{j \le -J} \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^0_M) - \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^{-1}_M) = o_{L^2(\mathbb{P}_*)}(1), \tag{2.27}$$

as J goes to infinity. Using Lemma 2.49, we have that there exists $\rho < 1$ and $C > 0$ such that, for all $j \le -1$and for all M,

$$\left\| \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^0_M) - \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^{-1}_M) \right\|_{L^2(\mathbb{P}_*)} \le C\rho^{-j}.$$

Analogous results hold with $\mathcal{G}^0, \mathcal{G}^1$ in place of $\mathcal{G}^0_M, \mathcal{G}^1_M$. Therefore, for all $\epsilon > 0$, there

exists $J(\epsilon) > 0$ for which

$$\sum_{j \leq -J(\epsilon)} \left\| \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}_M^0) - \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}_M^{-1}) \right\|_{L^2(\mathbb{P}_*)} + \left\| \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^0) - \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^{-1}) \right\|_{L^2(\mathbb{P}_*)} < \epsilon.$$

It now suffices to prove for all j that

$$\mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}_M^l) \overset{L^2(\mathbb{P}_*)}{\to} \mathbb{P}_*(X_{j-1} = a, X_j = b | \mathcal{G}^l). \tag{2.28}$$

Firstly, we show that, for each $\ell = -1, 0$, the $\mathcal{G}_M^\ell$ form an increasing sequence of σ-algebras $\mathcal{G}_M^\ell \subset \mathcal{G}_{M'}^\ell$ for $M < M'$, and that the limiting σ-algebra $\mathcal{G}_\infty^\ell := \sigma\left(\bigcup_{M \in \mathbb{N}} \mathcal{G}_M^\ell\right)$ is equal to $\mathcal{G}^l$. We suppress the dependence on ℓ, since the proofs for each case are the same. For the sake of presentation, we code the coarsened observations $Y_i^{(M)}$ as their projection onto the left endpoint of whichever interval of the partition $\mathcal{I}_M$ contains Y_i: $Y_i^{(M)} = e_j$ if $Y_i \in I_j^{(M)} = (e_j, e_{j+1})$.

To see that the $\mathcal{G}_M$ form a filtration, we wish to show that every set in $\mathcal{G}_M$ is also in $\mathcal{G}_{M+1}$. Every set in these $\sigma-$algebras is generated by the preimages of the left endpoints of intervals under Y_i^M for some i. Suppose the partition $\mathcal{I}_{M+1}$ is obtained from $\mathcal{I}_M$ by splitting one bin. The events $\{Y_i^M = e_j\}$, for $j = 1, \ldots, M$ and e_j the endpoint of bin j, generate $\mathcal{G}_M$. For the split bin, if we split bin M in $\mathcal{I}_M$ into bins M and $M+1$ in $\mathcal{I}_{M+1}$, then we can express $\{Y_i^M = e_j\} = \{Y_i^{M+1} = e_j\}$ for $j < M$ and $\{Y_i^M = e_M\} = \{Y_i^{M+1} = e_M\} \cup \{Y_i^{M+1} = e_{M+1}\}$. Hence we can express these events as events in $\mathcal{G}_{M+1}$, and by induction the argument remains true for any nested sequence of intervals.

To see that each $\mathcal{G}_M \subset \mathcal{G}$ it suffices to note that $\{Y_i^M = e_j\} = \{Y_i \in [e_j, e_{j+1})\} \in \mathcal{G}$. This shows that $\mathcal{G}_\infty \subset \mathcal{G}$. To show that $\mathcal{G} \subset \mathcal{G}_\infty$, note that the Borel $\sigma-$algebra on $[0, 1]$ is generated by the sets $(t, 1]$. So it suffices to check that the sets $\{Y_i > t\}$ are in $\mathcal{G}_\infty$, as these sets generate $\mathcal{G}$. But, for all $t \in (0, 1)$, there exists a sequence $e_{j(t,M)} \leq t$ converging to t and such that $e_{j(t,M)+1} > t$ also converges to t. Thus $\{Y_i^M > t\} = \{Y_i \geq e_{j(t,M)+1}\}$ increases to $\{Y_i > t\}$ and $\{Y_i > t\} = \lim_M \{Y_i^M > t\} \subset \mathcal{G}_\infty$. Therefore, the conditional probabilities appearing in the histogram score expression (2.16) form a bounded martingale in L^2 with $\mathcal{G}_\infty^\ell = \mathcal{G}^\ell$, $\ell = -1, 0$. The L^2 martingale convergence theorem them implies that (2.28) holds for any j, ℓ, and Lemma 2.19 is proved. □

The proofs of the following results, which are used in the proofs of Lemmas 2.17 and 2.18, are deferred to Section 2.10.

Lemma 2.20. *If* $h_M = (h_{r,M})_{r\le R}$ *with* $\sup_M \sum_r \|h_{r,M}\|_{L^2(f_r^*)} \le L < \infty$, *then*

$$H_M(h_M) = H(h_M) + o_{L_2(\mathbb{P}_*)}(1).$$

Lemma 2.21. *Let* $\tilde{\mathcal{A}}_M$ $\tilde{\mathcal{P}}_M$, $\mathcal{A}$, $\mathcal{P}$ *be as in the statement of Lemma 2.17. Then, if* $\tilde{\mathcal{A}} := \lim_{M\to\infty} \tilde{\mathcal{A}}_M$, *we have* $\tilde{\mathcal{A}} = \mathcal{A}$.

Lemma 2.22. *Let* h_M *be a sequence of step functions on the partitions of model* $\mathcal{M}(\mathcal{I}_M)$ *with* $\|h_M\| = \sum_r \|h_{r,M}\|_{L^2} \le 1$, *such that* $h_{r,M} \in L_0^2(f_r^*)$. *For any* $k \ge 0$,

$$H_M(h_M) \overset{L^2(\mathbb{P}_*)}{\to} 0 \implies \sum_{r=1}^{R} \sum_{k'=-k}^{0} h_M(y_{k'}) \mathbb{P}_*(X_{k'} = r | Y_{-k:0}) \overset{L^2(\mathbb{P}_*)}{\to} 0.$$

2.6.2 Proof of Theorem 2.10

The proof of Theorem 2.10 works by contradiction. If there exists no $C > 0$ such that, for all $\mathbf{f}$ with $\|g^{(3)}_{Q,\mathbf{f}} - g^{(3)}_{Q,\mathbf{f}^*}\|_{L^1}$ sufficiently small,

$$\frac{\|g^{(3)}_{Q,\mathbf{f}} - g^{(3)}_{Q,\mathbf{f}^*}\|_{L^1}}{\sum_{r=1}^R \|f_r - f_r^*\|_{L^1}} \ge C,$$

then there exists a sequence of emission densities $\mathbf{f}^{(n)}$ such that $\|g^{(3)}_{Q,\mathbf{f}} - g^{(3)}_{Q,\mathbf{f}^*}\|_{L^1} = o(1)$ and

$$\liminf_n \frac{\|g^{(3)}_{Q,\mathbf{f}^n} - g^{(3)}_{Q,\mathbf{f}^*}\|_{L^1}}{\sum_{r=1}^R \|f_r^n - f_r^*\|_{L^1}} = 0.$$

For $j = 2, 3$, $r \in [R]$, and $\mathbf{f}^{(n)}$ as above, set

$$\Delta_{n,j} = \frac{\|g^{(j)}_{Q,\mathbf{f}^{(n)}} - g^{(j)}_{Q,\mathbf{f}^*}\|_{L^1}}{\sum_{r=1}^R \|f_r^n - f_r^*\|_{L^1}}; \quad h_r^n = \frac{f_r^n - f_r^*}{\sum_{r=1}^R \|f_r^n - f_r^*\|_{L^1}},$$

and write $\mathbf{h}^n = (h_r^n : r \in [R])$.

Since $\liminf_n \Delta_{n,3} = 0$, there exists a subsequence $\phi_1(n)$ such that $\Delta_{\phi_1(n),3} \to 0$, which implies also that $\Delta_{\phi_1(n),2} \to 0$ since $\Delta_{n,2} \le \Delta_{n,3}$. We have by assumption that $\|g^{(j)}_{Q,\mathbf{f}^{(\phi_1(n))}} - g^{(j)}_{Q,\mathbf{f}^*}\|_{L^1} \to 0$, and by applying Lemma 2.23 we can find a further subsequence $\phi_2(\phi_1(n))$ such that $\mathbf{f}^{(\phi_2(\phi_1(n)))} \to \mathbf{f}^*$ in L^1. For notational convenience we write $\Delta_{n,j}$ for the terms $\Delta_{\phi_2(\phi_1(n)),j}$ along this subsequence, and in general any index with n is now interpreted as

being along this subsequence. Since $\sum_{r=1}^{R}\|f_r^n - f_r^*\|_{L^1} \to 0$, $\Delta_{n,2} \to 0$ implies that

$$\sum_{r,s}(PQ^*)_{rs}(f_r^*(y_1)h_s^n(y_2) + h_r^n(y_1)f_s^*(y_2)) = o_{L^1}(1),$$

where $P = \mathrm{diag}(p_1^*, \ldots, p_R^*)$. Partition $I_1, \ldots, I_R$ such that $\mathbf{F}^*$, as defined in Lemma 2.23, has rank R. Define also $\mathbf{H}^n = (H_{ir}^n)_{ir} = \int_{I_i} h_r^n(y)dy$. We have, for any integrable function G and any $i \in [R]$, that

$$\int_{\mathbb{R}}\int_{\mathbb{R}}|G(y_1,y_2)|dy_1dy_2 \geq \int_{\mathbb{R}}\int_{I_i}|G(y_1,y_2)|dy_1dy_2 \geq \int_{\mathbb{R}}\left|\int_{I_i}G(y_1,y_2)dy_1\right|dy_2,$$

which implies in particular that

$$\sum_{r,s}(PQ^*)_{rs}(F_{ir}^*h_s^n(y_2) + H_{ir}^n f_s^*(y_2)) = o_{L^1}(1).$$

This gives us

$$(\mathbf{F}^*)^T(PQ^*)\mathbf{h}^n = -\mathbf{H}_n^T(PQ)\mathbf{f}^* + o_{L^1}(1).$$

The sequence $\mathbf{H}_n$ is bounded, and so converges to some $\mathbf{H}_0$ along a subsequence $\phi_3(n)$ which we pass to. Then, since $\mathbf{F}^*$ is of rank R, we may write

$$\mathbf{h}^n = \underbrace{-((\mathbf{F}^*)^T(PQ))^{-1}(\mathbf{H}^*)^T(PQ)}_{=B_1}\mathbf{f}^* + o_{L^1}(1),$$

and so $h_r^n \to (B_1\mathbf{f}^*)_r$ in L^1 for each $r = 1, \ldots, R$. Next, since we have that $\Delta_{n,3} \to 0$ in L^1, we get

$$\sum_{r,s,t}(PQ)_{rs}Q_{st}\left(f_r^*(y_1)f_s^*(y_2)h_t^n(y_3) + f_r^*(y_1)h_s^n(y_2)f_t^*(y_3) + h_r^n(y_1)f_s^*(y_2)f_t^*(y_3)\right) = o_{L_1(1)}.$$

Replacing h_r^n by its limit $(B_1\mathbf{f}^*)_r$, we get that, for Lebesgue-almost-all y_1, y_2, y_3

$$\sum_{r,s,t}(PQ)_{rs}Q_{st}\left(f_r^*(y_1)f_s^*(y_2)(B_1\mathbf{f}^*)_t(y_3) + f_r^*(y_1)(B_1\mathbf{f}^*)_s(y_2)f_t^*(y_3) + (B_0\mathbf{f}^*)_r(y_1)f_s^*(y_2)f_t^*(y_3)\right) = 0$$

Under **Assumption** 2.4, we have continuity of f^* and so in fact the above display holds for all y_1, y_2, y_3. This expression is of the form of Equation (2.26) in the proof of Lemma 2.18, and the same proof techniques show that $B_1 = 0$ and hence $\|\mathbf{h}^n\| = o_{L^1}(1)$. This yields the desired contradiction as $\sum_{r=1}^{R}\|h_r^n\|_{L^1} = 1$ for all $n \in \mathbb{N}$. We conclude that no subsequence can exist for which $\Delta_{n,3} \to 0$, and hence that $\liminf_{n\to\infty}\Delta_{n,3} > 0$, which

terminates the proof of Theorem 2.10.

The proof of Theorem 2.10 uses the following result, which we prove in Section 2.10.5.

Lemma 2.23. *Consider a sequence of densities* $\mathbf{f}^{(n)}$ *such that* $g^{(3)}_{Q^*,\mathbf{f}^{(n)}} \overset{L^1(dy^3)}{\to} g^{(3)}_{Q^*,\mathbf{f}^*}$. *Then* $\mathbf{f}^{(n)} \overset{L^1(dy)}{\to} \mathbf{f}^*$ *along a subsequence.*

2.6.3 Proof of Theorem 2.13

To prove Theorem 2.13, we make use of Lemma 2.47 which allows us to approximate the MLE by an estimator which does not depend on the observation for which we control the respective smoothing probability. Intuitively, this result says that a single observation does not influence our MLE too much.

Proof of Theorem 2.13. The starting point of the proof is Proposition 2.2 of [27]. Let $\theta = (Q, \mathbf{f})$ and $\|Q - Q^*\| \leq M_n/\sqrt{n}$ so that $Q_{i,j} \geq \bar{q}/2 > 0$ with $\bar{q} = \min_{i,j} Q^*_{i,j} < \frac{1}{2}$, under **Assumption** 2.3. Set $\rho = 1 - \bar{q}/(1-\bar{q})$, then

$$\Delta_k := \|\mathbb{P}_\theta(X_k = \cdot|Y_{1:n}) - \mathbb{P}_*(X_k = \cdot|Y_{1:n})\|_{TV} \leq \frac{C^*}{\bar{q}}\Big[\rho^k\|\mu_Q - \mu^*\| + \|Q - Q^*\| + \sum_{\ell=1}^{n}\frac{\bar{q}\rho^{|\ell-k|}}{c_*(Y_\ell)}\max_s|f_s(y_\ell) - f^*_s(Y_\ell)|\Big],$$

where $c^*(y) = \min_{r\in[R]}\sum_{s\in[R]} Q^*_{rs}f^*_s(y) \geq \bar{q}\sum_s f^*_s(y)$, and C^* is a constant depending only on $Q^*, \mathbf{f}^*$. Throughout the proof, we reuse C^* to denote a generic constant depending only on $Q^*, \mathbf{f}^*$. First, for any $K_n \to \infty$, if

$$B_n = \{\theta,\ \|Q^* - \hat{Q}\| \leq \frac{K_n}{\sqrt{n}};\ \max_s\|f^*_s - \hat{f}_s\|_{L^1} \leq K_n\epsilon_n\} \cap \{\max_r\|f_r\|_2 \leq e^{\gamma n\epsilon_n^2}\},$$

for γ defined in (2.19), then Theorem 2.12 together with condition (2.19) gives $\Pi_{cut}(B_n^c|Y_{1:n}) = o_{\mathbb{P}_*}(1)$. Recall also that $n\epsilon_n^2 \to \infty$, so that, for $z_n > 0$,

$$\Pi_{cut}(\Delta_k > z_n\epsilon_n|Y_{1:n}) \leq \Pi_{cut}(\{\Delta_k > z_n\epsilon_n\} \cap B_n|Y_{1:n}) + o_{\mathbb{P}_*}(1).$$

Since $\|\Pi_1(dQ|Y_{1:n}) - \phi_n(Q|Y_{1:n})dQ\|_{TV} \to 0$, where $\phi_n(Q|Y_{1:n})$ is a Gaussian density centred at the MLE of variance $\tilde{J}^{-1}/n$, we have $\|\Pi_{cut} - \tilde{\Pi}_{cut}\|_{TV} = o_{\mathbb{P}_*}(1)$, where $\tilde{\Pi}_{cut}$ is the measure $\phi_n(Q|Y_{1:n})dQ\Pi_2(d\mathbf{f}|Q, Y_{1:n})$. Moreover, on B_n,

$$\Delta_k \leq \frac{C^*K_n}{\sqrt{n}} + \sum_{\ell=1}^{n}\frac{\bar{q}\rho^{|\ell-k|}}{c_*(Y_\ell)}\max_s|f_s(Y_\ell) - f^*_s(Y_\ell)|. \tag{2.29}$$

To prove Theorem 2.13, we thus need to control the sum in (2.29) under the posterior distribution. In [1], the authors use either a control in sup-norm of $\hat{f}_s - f_s^*$, or split their data into a *training set* used to estimate $\mathbf{f}^*$, and a *test set* used to estimate the smoothing probabilities. We show here that, with the Bayesian approach, we do not need to split the data in two parts nor do we need a sup-norm control. We believe that our technique of proof might still be useful for other approaches.

We split the sum in (2.29) into $|\ell - k| \leq T_n$ and $|\ell - k| > T_n$, where $T_n \epsilon_n$ converges arbitrarily slowly to 0. We first control, for all $r \in [R]$ and any $z_n > 0$ going to infinity,

$$\begin{aligned}\tilde{\Pi}_{cut}\left[B_n \cap \left\{\sum_{|\ell-k|\leq T_n} \frac{\rho^{|\ell-k|}|f_r(Y_\ell) - f_r^*(Y_\ell)|}{c_*(Y_\ell)} \geq z_n\epsilon_n \sum_{|\ell-k|\leq T_n} \rho^{|\ell-k|/2}\right\}\middle| Y_{1:n}\right] \\ \leq \sum_{|\ell-k|\leq T_n} \tilde{\Pi}_{cut}\left[B_n \cap \left\{\frac{|f_r(Y_\ell) - f_r^*(Y_\ell)|}{c_*(Y_\ell)} \geq z_n\epsilon_n\rho^{-|\ell-k|/2}\right\}\middle| Y_{1:n}\right]. \qquad (2.30)\end{aligned}$$

To bound (2.30), we need to study the quantities $\pi_{n,\ell}(z) = \Pi_{cut}\left[B_n \cap A_n(z;\ell)\middle| Y_{1:n}\right]$, where $z > 0$ and

$$A_n(z;\ell) = \left\{\frac{|f_r(Y_\ell) - f_r^*(Y_\ell)|}{\sum_s f_s^*(Y_\ell)} > z\epsilon_n\right\}.$$

Write $L_n(Y_{1:n}, \theta) = \sum_s f_s(Y_\ell) L_{-\ell}(\theta, s)$, where $L_{-\ell}(\theta, s)$ depends on $Y_{1:n}^{-\ell}$ and is given by

$$L_{-\ell}(\theta, s) = L(y_{1:\ell-1}, \theta)\mathbb{P}_\theta(X_\ell = s|Y_{1:\ell-1})L(y_{\ell+1:n}|X_\ell = s, \theta).$$

Note that

$$\begin{aligned}L(y_{\ell+1:n}|X_\ell = s, \theta) &= \sum_{s'} Q_{ss'} L(y_{\ell+1:n}|X_{\ell+1} = s', \theta) \\ &\in \left[\frac{\bar{q}}{2}\sum_{s'} L(y_{\ell+1:n}|X_{\ell+1} = s', \theta), \sum_{s'} L(y_{\ell+1:n}|X_{\ell+1} = s', \theta)\right],\end{aligned}$$

and also that $\mathbb{P}_\theta(X_\ell = s|Y_{1:\ell-1}) \in [\frac{\bar{q}}{2}, 1]$, which together implies, denoting $\bar{L}_{-\ell}(\theta) = \max_s L_{-\ell}(\theta, s)$, that

$$L_{-\ell}(\theta, s) \in \left[\frac{\bar{q}^2}{4}\bar{L}_{-\ell}(\theta), \bar{L}_{-\ell}(\theta)\right].$$

This implies that the cut posterior mass of $A_n(z,\ell) \cap B_n$ is bounded as

$$\begin{aligned}&\Pi_{cut}(A_n(z;\ell) \cap B_n|Y_{1:n}) \\ &\quad \leq \frac{4}{\bar{q}^2} \frac{\int_{B_n} \mathbb{1}_{A_n(z;\ell)}(f_r)\bar{L}_{-\ell}(\theta)(\sum_s f_s(Y_\ell))\Pi_2(d\mathbf{f}|Q)\phi_n(Q|Y_{1:n})dQ}{\int_{B_n} \bar{L}_{-\ell}(\theta)(\sum_s f_s(Y_\ell))\Pi_2(d\mathbf{f}|Q)\phi_n(Q|Y_{1:n})dQ}. \qquad (2.31)\end{aligned}$$

Applying Lemma 2.47, we also have, uniformly in $\ell \in [k-T_n, k+T_n] \cap \mathbb{N}$, $\|\phi_n(Q|Y_{1:n})dQ - \phi_n(Q|Y_{1:n}^{-\ell})dQ\|_{TV} \to 0$, where $\phi_n(Q|Y_{1:n}^{-\ell})$ is a Gaussian centred at the estimator defined in Lemma 2.47 and also of variance $\tilde{J}^{-1}/n$, which depends only on $Y_{1:n}^{-\ell}$. Hence, with

$$\tilde{\Pi}_{cut}^{(\ell)}(A|Y_{1:n}) = \frac{\int_A L_n(Y_{1:n},\theta)\Pi_2(d\mathbf{f})\phi(Q|Y_{1:n}^{(-\ell)})dQ}{D_n(\ell)},$$

where $D_n(\ell) = \int_{B_n} \bar{L}_{-\ell}(\theta)(\sum_s f_s(Y_\ell))\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ$, we have

$$\max_{|\ell-k|\leq T_n} \|\tilde{\Pi}_{cut}^{(\ell)}(\cdot|Y_{1:n}) - \tilde{\Pi}_{cut}(\cdot|Y_{1:n})\|_{TV} = o_{\mathbb{P}_*}(1).$$

Hence, it suffices to control each summand of (2.30) with $\tilde{\Pi}_{cut}$ replaced by $\tilde{\Pi}_{cut}^{(\ell)}$. We first focus on lower bounding the denominator $D_n(\ell)$. Writing $\Omega_n = \{\sum_s f_s(Y_\ell) \geq \sum f_s^*(Y_\ell)/2\}$, it is immediate that

$$D_n(\ell) \geq \int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{1}_{\Omega_n}\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ \left(\sum_s f_s^*(Y_\ell)/2\right).$$

We will show that we can replace the above bound with a bound in probability, in which the integrand does not feature the indicator $\mathbb{1}_{\Omega_n}$. This is similar in spirit to concentration results based around the framework of [48] (including our own Theorem 2.27), where a key ingredient of the proof is to lower bound the denominator by a suitable quantity. We have that

$$\begin{aligned}
&\mathbb{P}_*\left(\int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{1}_{\Omega_n}\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ) < \frac{1}{2}\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ\right) \\
&= \mathbb{P}_*\left(\int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{1}_{\Omega_n^c}\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ \geq \frac{1}{2}\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ\right) \\
&\leq 2\mathbb{E}_*\left[\mathbb{E}_*\left[\frac{\int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{1}_{\Omega_n^c}\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}\middle| Y_{1:n}^{-\ell}\right]\right] \\
&\leq 2\mathbb{E}_*\left[\frac{\int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{P}(\Omega_n^c|Y_{1:n}^{-\ell})\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}\right].
\end{aligned}$$

Now, we can bound the probability in the integrand by

$$\begin{aligned}\mathbb{P}_*(\Omega_n^c|Y_{1:n}^{-\ell}) &= \mathbb{P}_*\left(\sum_s f_s(Y_\ell) < \frac{1}{2}\sum_s f_s^*(Y_\ell)|Y_{1:n}^{-\ell}\right)\\ &= \mathbb{P}_*\left(\sum_s (f_s^*(Y_\ell) - f_s(Y_\ell)) \geq \sum_s f_s^*(Y_\ell)/2\middle|Y_{1:n}^{-\ell}\right)\\ &\leq 2\int \frac{\sum_s |f_s^*(y_\ell) - f_s(y_\ell)|}{\sum_s f_s^*(y_\ell)} f^*(y_\ell|Y_{1:n}^{-\ell})dy_\ell.\end{aligned}$$

We can bound the conditional density by $f^*(y_\ell|Y_{1:n}^{-\ell}) \leq \sum_s f_s^*(y_\ell)$, so that

$$\mathbb{P}_*(\Omega_n^c|Y_{1:n}^{-\ell}) \leq 2\sum_s \|f_s^* - f_s\|_{L_1} \leq 2RK_n\epsilon_n,$$

which in turns implies that

$$\begin{aligned}\mathbb{P}_*\Bigg(\int_{B_n} \bar{L}_{-\ell}(\theta)\mathbb{1}_{\Omega_n}\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ < \\ \frac{1}{2}\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ\Bigg) \leq 4RK_n\epsilon_n.\end{aligned}$$

Thus, on this event of probability at least $1 - 4RK_n\epsilon_n$, we can bound the denominator as

$$D_n(\ell) \geq \frac{1}{4}\sum_s f_s^*(y_\ell) \underbrace{\int_{B_n} \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}_{=:M(Y_{1:n}^{-\ell})}, \tag{2.32}$$

and so, on this event,

$$\begin{aligned}\tilde{\Pi}_{cut}^{(\ell)}(A_n(z;\ell)\cap B_n|Y_{1:n}) &\leq \frac{16}{\bar{q}^2}\frac{\int_{B_n}\mathbb{1}_{A_n(z;\ell)}(\sum_s f_s(Y_\ell))\bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{\sum_s f_s^*(Y_\ell)M(Y_{1:n}^{-\ell})}\\ &\leq \frac{16}{\bar{q}^2}\Bigg[\underbrace{\frac{\int_{B_n}\mathbb{1}_{A_n(z;\ell)}(\sum_s |f_s(Y_\ell) - f_s^*(Y_\ell)|)\bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{\sum_s f_s^*(Y_\ell)M(Y_{1:n}^{-\ell})}}_{J_1}\\ &\quad + \underbrace{\frac{\int_{B_n}\mathbb{1}_{A_n(z;\ell)}\bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{M(Y_{1:n}^{-\ell})}}_{J_2}\Bigg].\end{aligned}$$

Now, using $\mathbb{E}_*[J_1] = \mathbb{E}_*[\mathbb{E}_*[J_1|Y_{1:n}^{-\ell}]]$, together with the fact that $f(y_\ell|Y_{1:n}^{(-\ell)})dy_\ell \leq \sum_s f_s^*(y_\ell)dy_\ell$, as well as the bound on $\|f_s - f_s^*\|_{L^1}$ on B_n, we obtain

$$\mathbb{E}_*[J_1] \leq \mathbb{E}_*\left[\frac{\int_{B_n}\sum_s\|f_s - f_s^*\|_1\bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{M(Y_{1:n}^{-\ell})}\right] \leq RK_n\epsilon_n.$$

To bound J_2, note that $\mathbb{1}_{A_n(z;\ell)} \leq |f_r(Y_\ell) - f_r^*(Y_\ell)|/(z\epsilon_n \sum_r f_r^*(Y_\ell))$, so that

$$J_2 \leq \frac{\int_{B_n} |f_r(Y_\ell) - f_r^*(Y_\ell)| \bar{L}_{-\ell}(\theta)\Pi_2(df)\phi_n(Q|Y_{1:n}^{-\ell})dQ}{z\epsilon_n M(Y_{1:n}^{-\ell}) \sum_s f_s^*(Y_\ell)},$$

and bounding as before with J_1, we obtain $\mathbb{E}_*[J_2] \leq \frac{RK_n}{z}$. This implies that

$$\begin{aligned}
&\mathbb{P}_* \left[\sum_{|\ell-k|\leq T_n} \tilde{\Pi}_{cut}^{(\ell)}(A_n(\rho^{-|\ell-k|/2} z_n; \ell) \cap B_n | Y_{1:n}) > \epsilon \right] \\
&\leq \sum_{|\ell-k|\leq T_n} \mathbb{P}_* \left[D_n(\ell) \leq \frac{c_*(Y_\ell) M(Y_{1:n}^{-\ell})}{4} \right] + T_n \frac{16RK_n\epsilon_n}{\epsilon \bar{q}^2} + \sum_{|\ell-k|\leq T_n} \frac{\rho^{|\ell-k|/2} K_n}{\epsilon z_n} \\
&\lesssim \frac{T_n K_n \epsilon_n}{\epsilon} + \frac{K_n}{z_n \epsilon}.
\end{aligned}$$

Recall that $T_n \leq \epsilon^2 [K_n\epsilon_n]^{-1}$ and $z_n \geq K_n/\epsilon^2$, so that the above bound becomes

$$\mathbb{P}_* \left[\sum_{|\ell-k|\leq T_n} \tilde{\Pi}_{cut}^{(\ell)}(A_n(\rho^{-|\ell-k|/2} z_n; \ell) \cap B_n | Y_{1:n}) > \epsilon \right] \leq c_1 \epsilon, \tag{2.33}$$

for some constant c_1 independent on n and ϵ.

Returning to the original goal of bounding (2.29), and having now bounded the sum over $|\ell - k| \leq T_n$, we study the sum over $|\ell - k| > T_n$. For this, we will bound the sum under $\tilde{\Pi}_{cut}$ defined previously. First note that

$$\begin{aligned}
\mathbb{P}_* \left[\exists \ell,\, c_*(Y y_\ell) \leq e^{-L_n n\epsilon_n^2} \right] &\leq n \int_{c_*(y) \leq e^{-L_n n\epsilon_n^2}} \sum_{s=1}^{R} p_s^* f_s^*(y) dy \\
&\leq n \int_{c_*(y) \leq e^{-L_n n\epsilon_n^2}} \sqrt{c^*(y)} e^{-L_n n\epsilon_n^2/2} dy \\
&\leq C^* n e^{-L_n n\epsilon_n^2/2},
\end{aligned}$$

where we use that $c_*(y) \geq \sum_s p_s^* f_s^*(y)$, and that $\sqrt{c_*}$ is L^1 bounded by some C^* from condition (2.18). As soon as $L_n n\epsilon_n^2 \geq L_0 \log n$ with L_0 large enough, which is possible since $n\epsilon_n^2 \gtrsim \log n$ by assumption, we have $\mathbb{P}_* \left[\exists \ell,\, c_*(Y y_\ell) \leq e^{-L_n n\epsilon_n^2} \right] = o(1)$ and we can bound $c_*(Y y_\ell) \geq e^{-L_n n\epsilon_n^2}$ for all ℓ. Next, we define

$$\bar{A}_n(\ell) = \left\{ \theta;\, \sum_{|\ell-k|>T_n} \rho^{|\ell-k|} |f_r(Y y_\ell) - f_r^*(Y y_\ell)| \leq z_n \epsilon_n e^{-L_n n\epsilon_n^2} \right\}.$$

We write

$$\tilde{\Pi}_{cut}(B_n \cap \bar{A}_n^c(\ell)|Y_{1:n}) = \frac{\int_{B_n \cap \bar{A}_n^c(\ell)} e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})\phi_n(Q|Y_{1:n})dQ}{\int_{B_n} e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})\phi_n(Q|Y_{1:n})dQ} =: \frac{N_n(\bar{A}_n(\ell)^c)}{D_n},$$

where we write the denominator as D_n and write $N_n(\bar{A}_n(\ell)^c)$ for the numerator. Using Lemma 2.30 which shows that $\mathbb{P}_*[D_n < e^{-L_n n\epsilon_n^2}] \lesssim L_n^{-1}$, together with $\phi_n(Q|Y_{1:n}) \lesssim n^{R(R-1)/2}$, we obtain, with probability greater than $1 - O(1/L_n)$, and with L_n going arbitrarily slowly to infinity,

$$\begin{aligned}
\tilde{\Pi}_{cut}(B_n \cap \bar{A}_n(\ell)^c|Y_{1:n}) &\leq e^{L_n n\epsilon_n^2}\int_{B_n} \mathbb{1}_{\bar{A}_n(\ell)^c} e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})\phi_n(Q|Y_{1:n})dQ \\
&\leq \frac{C_* e^{2L_n n\epsilon_n^2} n^{R(R-1)/2}}{z_n\epsilon_n} \sum_{|\ell-k|>T_n} \rho^{|\ell-k|}\int_{B_n} |f_r(Yy_\ell) - f_r^*(Yy_\ell)|e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})dQ \\
&\leq C_* e^{3L_n n\epsilon_n^2} \sum_{|\ell-k|>T_n} \rho^{|\ell-k|}\int_{B_n} |f_r(Yy_\ell) - f_r^*(Yy_\ell)|e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})dQ,
\end{aligned}$$

where, in the final line, we again use that $\log n \lesssim n\epsilon_n^2$ to consolidate the exponential. This implies that $\tilde{\Pi}_{cut}(B_n \cap \bar{A}_n(\ell)^c|Y_{1:n})$ is controlled under $\mathbb{P}_*$ as

$$\begin{aligned}
&\mathbb{P}_*\left[\tilde{\Pi}_{cut}(B_n \cap \bar{A}_n(\ell)^c|Y_{1:n}) \geq 2/L_n\right] \leq o(1) \\
&\quad + \mathbb{P}_*\left[\sum_{|\ell-k|>T_n} \rho^{|\ell-k|}\int_{B_n} |f_r(Yy_\ell) - f_r^*(Yy_\ell)|e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})dQ > \frac{e^{-3L_n n\epsilon_n^2}}{L_n}\right] \\
&\quad \leq \mathbb{P}_*\left[\sum_{|\ell-k|>T_n} \int_{B_n} |f_r(Yy_\ell) - f_r^*(Yy_\ell)|e^{\ell_n(\theta)-\ell_n(\theta^*)}\Pi_2(d\mathbf{f})dQ > \frac{\rho^{-T_n}e^{-3L_n n\epsilon_n^2}}{L_n}\right] + o(1) \\
&\quad \leq \frac{C_* L_n \rho^{T_n} e^{3L_n n\epsilon_n^2}}{L_n} \sum_{|\ell-k|>T_n} \int_{B_n} \mathbb{E}_{Q,\mathbf{f}}[|f_r(Yy_\ell) - f_r^*(Yy_\ell)|]\Pi_2(d\mathbf{f})dQ + o(1) \\
&\quad \leq C_* L_n \rho^{T_n} e^{3L_n n\epsilon_n^2} \sum_{s=1}^{R} \int_{B_n}\int f_s(y)(f_r(y) + f_r^*(y))dy\Pi_2(d\mathbf{f}) + o(1) \\
&\quad \leq C_* L_n \rho^{T_n} e^{3L_n n\epsilon_n^2} \sum_{s} \int_{B_n} \|f_s\|_2(\|f_r\|_2 + \|f_r^*\|_2)\Pi_2(d\mathbf{f}) + o(1) \\
&\quad \leq C_* \rho^{T_n} e^{4L_n n\epsilon_n^2} + o(1) = o(1),
\end{aligned} \tag{2.34}$$

with the final equality holding as soon as $T_n \log(1/\rho) \geq 5nL_n\epsilon_n^2$. We recall that, in order to establish Equation (2.33) earlier in the proof, we required that $T_n\epsilon_n \to 0$ arbitrarily slowly. We thus use the assumption $n\epsilon_n^3 \to 0$ to choose T_n so that both conditions may hold simultaneously.

Finally, combining (2.34) with (2.33) and (2.29), and since K_n and L_n can be chosen to go

arbitrarily slowly to infinity and $n\epsilon_n^2 \gtrsim \log n$, we obtain that, for any z_n going to infinity,

$$\Pi_{cut}(\Delta_k > z_n\epsilon_n|Y_{1:n}) = o_{\mathbb{P}_*}(1),$$

which terminates the proof. □

2.7 Conclusion and discussion

In this paper we use the cut posterior approach of [61] for inference in semiparametric models, which we apply to the nonparametric Hidden Markov models. A difficulty with the Bayesian approach in high or infinite dimension is that it is difficult (if not impossible) to construct priors on these complex models which allow for *good* simultaneous inference on a collection of parameters of interest, where *good* means having good frequentist properties (and thus leading to some robustness with respect to the choice of the prior). As mentioned in Section 4.1, a number of examples have been exhibited in the literature where *reasonable* priors lead to poorly behaved posterior distributions for some functionals of the parameters. We believe that the cut posterior approach is an interesting direction to pursue in order to address this general problem, and we demonstrate in the special case of semiparametric Hidden Markov models that it leads to interesting properties of the posterior distribution and is computationally tractable.

Our approach is based on a very simple prior Π_1 on $Q, \mathbf{f}$ used for the estimation of Q based on finite histograms with a small number of bins. This enjoys a Bernstein-von Mises property, so that credible regions for Q are also asymptotic confidence regions. Moreover, by choosing M large (but not too large), we obtain efficient estimators. Proving efficiency for semiparametric HMMs is non trivial and our proof has independent interest. Another original and important contribution of our paper is an inversion inequality (stability estimate) comparing the L^1 distance between $g^{(3)}_{Q,\mathbf{f}_1}$ and $g^{(3)}_{Q,\mathbf{f}_2}$ and the L^1 distance between $\mathbf{f}_1$ and $\mathbf{f}_2$. Finally, another interesting contribution is our control of the error of the estimates of the smoothing probabilities using a Bayesian approach, which is based on a control under the posterior distribution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|f_r(Y_i) - f_r^*(Y_i)|}{\sum_r f_r^*(Y_i)},$$

despite the double use of the data (in the posterior on f_r and in the empirical distance above). It is a rather surprising result, which does not hold if f_r is replaced with $\hat{f}_r$

constructed using $y_1, \ldots, y_n$, unless a sup-norm bound on $\hat{f}_r - f_r^*$ is obtained.

Section 2.5 demonstrates clearly the estimation procedures and highlights the importance of choosing a small number of bins in practical situations. Here, there remains open the question of how to choose this number in a principled way, and an interesting extension of the work would follow along the lines of the third section of [45], in which the authors produce an oracle inequality to justify a cross-validation scheme for choosing the number of bins.

Finally, the paper deals with the case where R is known. This assumption is rather common both in theory (see for instance the works of [44], [27] or [7]) and in practice (for instance in genomic applications as in [107]). The identifiability results of [4] and [42] do show that R can be identified as well, leaving open the possibility to jointly estimate R, together with Q and $\mathbf{f}$. We leave this direction of research for future work.

2.8 Further Proofs of main results

In the following section, we detail the proofs of Theorem 2.6 and Proposition 2.7.

2.8.1 Proof of Theorem 2.6

The proof of Theorem 2.6 follows from Proposition 2.5 and Lemma 2.24.

Lemma 2.24. *Let $(\mathcal{I}_M)_M$ be a sequence of embedded partitions, let M be such that $\mathcal{I}_M$ is admissible for $\mathbf{f}^*$, and grant Assumptions 2.1, 2.3 and 2.4. Then the matrix*

$$\tilde{J}_{M+1} - \tilde{J}_M$$

is positive semi-definite.

Proof. Consider the MLE for $\theta = (Q(\theta), \underline{\omega}_M(\theta))$ for parameters in $\mathcal{Q} \times \Omega_M$ in the model with partition $\mathcal{I}_M$.

By the remarks at the end of the proof of Theorem 2.3, the MLE is regular. This means that, for all parameter sequences θ_n of the form $\theta_0 + n^{-\frac{1}{2}}h$ with $h \in [-H, H]$ for some fixed H, $\sqrt{n}(\hat{\theta}^{(n)} - \theta_n)$ converges in distribution under $\mathbb{P}_{\theta_n}$ to a variable, say Z, of fixed distribution, say μ_Z. Explicitly, this means that for Borel sets A we have

$$\mathbb{P}_{\theta_n}\left(\sqrt{n}(\hat{\theta}^{(n)} - \theta_n) \in A\right) \to \mu_Z(A) \text{ as } n \to \infty.$$

Here θ_n is a sequence of parameters in $\mathcal{Q} \times \Omega_M$. However, any likelihood-based estimator depends on the observations only through the probabilities of each bin assignment, by considering the multinomial model with the count data. Thus, we can replace $\mathbb{P}_{\theta_n}$ with any law from the full semiparametric model for which the transition matrix Q_n is the same and the functions f_n have the corresponding $\underline{\omega}_M(\theta_n)$ as the bin assignment probabilities, and the above convergence remains true.

Since $\mathcal{I}_{M+1}$ is made by splitting bins in $\mathcal{I}_M$, a collection of weights $\omega^{(M+1)} = (\omega_m^{(M+1)} : m \leq \kappa_{M+1})$ for the partition $\mathcal{I}_{M+1}$ is identified with weights $\omega^{(M)}$ for the partition $\mathcal{I}_M$. We set $\omega_m^{(M+1)} = \sum_{m' \in \mathrm{Ch}_M(m)} \omega_{m'}^{(M)}$, where $\mathrm{Ch}_M(m)$ satisfies $I_m^{(M)} = \bigcup_{m' \in \mathrm{Ch}_M(m)} I_{m'}^{(M+1)}$. This implies that the MLE for the model $\mathcal{M}(\mathcal{I}_M)$ may be obtained by maximising the likelihood in model $\mathcal{M}(\mathcal{I}_{M+1})$ subject to the constraint that the quantities $\dfrac{\omega_{m'}^{(M+1)}}{|I_{m'}^{(M+1)}|}$ coincide for all $m' \in \mathrm{Ch}(m)$. Regularity of the constrained MLE within $\mathcal{I}_{M+1}$ is then inherited from its regularity as a global MLE in $\mathcal{I}_{\mathcal{M}}$, by the preceding discussion.

Moreover, its asymptotic variance, which is given by the inverse Fisher information $\tilde{J}_M^{-1}$ for the model with partition $\mathcal{I}_M$, is lower bounded by the inverse Fisher information for the model with partition $\mathcal{I}_{M+1}$ by Lemma 2.39, hence $\tilde{J}_M^{-1} - \tilde{J}_{M+1}^{-1}$ is positive semi definite. □

Proof of Theorem 2.6. We have, by inspecting the part of the proof of Lemma 2.24 concerning regularity, that under any sequence of the form $\mathbb{P}_*^n = \mathbb{P}_{\theta^*(a,\mathbf{h})}$, where $\theta^*(a, \mathbf{h}) = (Q^* + \frac{a}{\sqrt{n}}, f_r^*(1 + \frac{1}{\sqrt{n}} h_r) : r \in [R])$ with $a \in \mathbb{R}^{R(R-1)}$ and $h_r \in L_0^2(f_r^*)$,

$$R_{M,n} := d_{BL}\left(\mathbb{P}_*^n\left(\sqrt{n}(\hat{Q}_n^{(M)} - (Q^* + \frac{a}{\sqrt{n}})) \in \cdot\right), \mu_M\right) \to 0,$$

where μ_M is Gaussian of covariance equal to $\tilde{J}_M^{-1}$ by Theorem 2.3. Here d_{BL} is the bounded Lipschitz metric which metrizes weak convergence. By taking $M_n \to \infty$ sufficiently slowly, we have $R_{M_n,n} \to 0$ also. Since, for any $t \in \mathbb{R}^{R(R-1)}$, $(t^T \tilde{J}_{M_n} t)^{-1}$ is decreasing by Lemma 2.24 and bounded below by $(t^T \tilde{J} t)^{-1}$ the efficient variance, it converges and the limit is equal to $(t^T \tilde{J} t)^{-1}$ by Proposition 2.5. This implies weak convergence of the measures μ_{M_n} to μ, the centred Gaussian measure whose variance is $\tilde{J}^{-1}$, and applying the triangle inequality in the metric d_{BL} proves the first claim.

For the Bernstein-von Mises result, we have from Theorem 2.4 that

$$\left\| \Pi_M(\sqrt{n}(Q - \hat{Q}_n) | Y_{1:n}) - \mathcal{N}(0, \tilde{J}_M^{-1}) \right\|_{TV} = \tilde{R}_{M,n},$$

where $\tilde{R}_{M,n} \to 0$ in $\mathbb{P}_*-$probability as $n \to \infty$. Refining the sequence from before, so that $M_n \to \infty$ sufficiently slowly that we additionally have $\tilde{R}_{M_n,n} \to 0$, and noting that $\tilde{J}_{M_n} \to \tilde{J}$, we get

$$\begin{aligned}
&\left\|\Pi_{M_n}(\sqrt{n}(Q-\hat{Q}_n)|Y_{1:n}) - \mathcal{N}(0,\tilde{J}_\infty^{-1})\right\|_{TV} \\
&\leq \left\|\Pi_{M_n}(\sqrt{n}(Q-\hat{Q}_n)|Y_{1:n}) - \mathcal{N}(0,\tilde{J}_{M_n}^{-1})\right\|_{TV} + \left\|\mathcal{N}(0,\tilde{J}^{-1}) - \mathcal{N}(0,\tilde{J}_{M_n}^{-1})\right\|_{TV} \\
&= o_{\mathbb{P}_*}(1).
\end{aligned}$$

□

2.8.2 Proof of Proposition 2.7

The proof of Proposition 2.7 is a direct consequence of the following result, which shows that Theorem 2.27 applies to semiparametric HMMs of the type described in Section 2.2.2. For the construction of Π_1, a C_1 diffeomorphism ϕ is implicitly used (if necessary) in order to consider histograms on $[0,1]$

Proposition 2.25. *Let $(Y_t)_{t\geq 1}$ be observations from a finite state space HMM with transition matrix Q and emission densities $\mathbf{f} = (f_r)_{r=1,\dots,R}$. Consider the cut posterior based on Π_1 associated to the partition $\mathcal{I}_M$ and Π_2.*

*(i) Under **Assumptions** 2.1-2.5, **Assumption** 2.7 is satisfied with $\mathcal{T}_n = \{Q; \|Q-Q^*\| \leq z_n/\sqrt{n}\}$, with $z_n \to \infty$ sufficiently slowly and ϕ_n the restriction to $\mathcal{T}_n$ of the Gaussian density centered at $\hat{Q}_{n,M}$ with variance $\tilde{J}_M^{-1}/n$.*

*(ii) Choosing Π_2 such that **Assumption** 2.6 is verified for suitable $\epsilon_n, \tilde{\epsilon}_n$ with $n\epsilon_n^2 \gtrsim \log n$, the assumptions of Theorem 2.27 are verified for the same ϵ_n and any $K_n \to \infty$.*

Proof. *(i)* The idea of the proof is as follows: We know from Theorem 2.4 that $\Pi_1(\cdot|Y_{1:n})$ is close in total variation distance to the normal distribution centred at the estimator $\hat{Q}$. Then the restriction of this distribution to a ball of radius z_n, centred at Q^*, where $M_n \to \infty$ slowly, will be close in total variation to the original normal distribution with high probability as $n \to \infty$.

Let $p_{1,n}$ be the normal density of variance $\tilde{J}_M^{-1}/n$ centred at $\hat{Q}$, where $\hat{Q}$ is the estimator from Theorem 2.4. Let $p_{2,n}$ be the restriction of this normal density to $\{Q : \|Q-Q^*\| \leq z_n/\sqrt{n}\}$ where $z_n \to \infty$. Fix $\epsilon > 0$ and choose n large enough that $\mathbb{P}_*(\|\hat{Q}-Q^*\| > cx_n'/\sqrt{n}) < \epsilon$ where $1 > c > 0$ is a constant. Let $p_{3,n}$ be the density obtained by restriction

of $p_{1,n}$ to $\{Q : \|\hat{Q} - Q^*\| \leq (1-c)z_n/\sqrt{n}\}$. Then, for sufficiently large n, we have with probability exceeding $1-\epsilon$ that the support of $p_{3,n}$ is a subset of the support of p_2, in which case

$$\|p_{1,n} - p_{2,n}\|_{L^1} \leq \|p_{1,n} - p_{3,n}\|_{L^1}.$$

But the variance of $p_{1,n}$ is of the order $O(n^{-1})$ and $z_n \to \infty$ so, as $n \to \infty$, the support of $p_{3,n}$ approaches the support of $p_{1,n}$ and $\|p_{1,n} - p_{3,n}\|_{L^1} \to 0$. But then $\|p_{1,n} - p_{2,n}\|_{L^1} \to 0$. Thus, with probability exceeding $1-\epsilon$, we have that $\|p_{1,n} - p_{2,n}\|_{L^1} \to 0$ as $n \to \infty$. Since $\epsilon > 0$ is arbitrary, we conclude $\|p_{1,n} - p_{2,n}\|_{L^1} = o_{\mathbb{P}_*}(1)$.

Taking $\phi_n = p_{2,n}$ as above, for any $z_n \to \infty$, we verify Assumption 2.7 for any ϵ_n for which $n\epsilon_n^2 \gtrsim \log n$.

(ii) This is a consequence of Theorem 3.1 in [102]. The choice of $\mathcal{T}_n$ also provides verification of the hypothesis of Lemma 3.2 of [102] and establishes the required control on KL divergence described in (2.35). To satisfy the testing assumption (2.36), it suffices to note that, for large enough n, $\mathcal{T}_n$ is a subset of $\{Q : Q_{i,j} \geq \frac{Q^*_{ij}}{2} \ \forall i,j \leq R\}$ under **Assumption 2.3**. □

2.9 General theorem for cut posterior contraction

In this section, we present a general theory for contraction of cut posteriors, which is developed in the style of the usual theory for Bayesian posteriors of [48]. The main result of this section is Theorem 2.27, from which Proposition 2.7 follows.

Consider a general semiparametric model in which there is a finite dimensional parameter ϑ and an infinite-dimensional parameter η. Suppose we wish to estimate the pair (ϑ, η) governing the law $\mathbb{P}^n_{\vartheta,\eta}$ of a random sample $Y_{1:n} \in \mathcal{Y}^n$. Assume the data is generated by some true distribution $\mathbb{P}^n_* = \mathbb{P}^n_{\vartheta^*,\eta^*}$ and consider the following two models:

Model 1: Consider a model $\mathcal{T} \times \mathcal{F}_1$ on pairs (ϑ, η) such that $\vartheta^* \in \mathcal{T}$ but for which we do not require that $\eta^* \in \mathcal{F}_1$. Consider a joint prior Π_1 over this space yielding a marginal posterior on ϑ given by

$$\Pi_1(d\vartheta|Y_{1:n}) = \int_{\mathcal{F}_1} \Pi_1(d\vartheta, d\eta|Y_{1:n}).$$

Model 2: Consider a model on η, conditional on ϑ, with a prior $\Pi_2(\cdot|\vartheta)$. We denote the parameter set for η by $\mathcal{F}_2$ and we assume that $\eta^* \in \mathcal{F}_2$. We obtain in this model, a

conditional posterior distribution $\Pi_2(d\eta|Y_{1:n}, \vartheta)$.

Definition 2.26. The *cut posterior* on (ϑ, η) is given by

$$\Pi_{\text{cut}}(d\vartheta, d\eta|Y_{1:n}) = \Pi_1(d\vartheta|Y_{1:n})\Pi_2(d\eta|\vartheta, Y_{1:n}).$$

Write $\mathbb{P}^n_{\vartheta,\eta}$ for the law of the observations $Y_{1:n}$. Define for some $\ell > 0$ and some loss function $d(\cdot,\cdot)$ on $\mathbb{P}^\ell_{\vartheta,\eta}$,

$$B_\ell(\epsilon) = \{(\vartheta,\eta) : d(\mathbb{P}^\ell_{\vartheta,\eta}, \mathbb{P}^\ell_{\vartheta^*,\eta^*}) < \epsilon\}, \quad M, \epsilon > 0,$$

with $B^c_\ell(\epsilon)$ its complement. Define also the Kullback-Leibler neighbourhoods of $\mathbb{P}^n_* = \mathbb{P}^n_{\vartheta^*,\eta^*}$ as

$$V_n(\mathbb{P}^n_*, \epsilon) = \{(\vartheta,\eta) : \mathcal{K}(\mathbb{P}^n_{\vartheta^*,\eta^*}|\mathbb{P}^n_{\vartheta,\eta}) < n\epsilon^2\},$$

where $\mathcal{K}(\mathbb{P}^n_{\vartheta^*,\eta^*}|\mathbb{P}^n_{\vartheta,\eta}) = \mathbb{E}^n_* \log \frac{d\mathbb{P}^n_{\vartheta^*,\eta^*}}{d\mathbb{P}^n_{\vartheta,\eta}}$ is the Kullback-Leibler divergence.

We now present a general theorem to characterize cut-posterior contraction result, in the spirit of the now classical result of [48]. Our main additional assumption for the cut setup is that we have sufficiently good control over Π_1, similar to the kind established in Section 2.2 in the HMM setting, which we detail now.

Denote by $\pi_1(\cdot|Y_{1:n})$ the marginal posterior density of ϑ with respect to some measure μ_1 on $\mathcal{T}$, associated to the prior Π_1 on ϑ, η.

Assumption 2.7. For all sequences $z_n \to \infty$ there exist $\mathcal{T}_n \subset \mathcal{T}$ with $\vartheta^* \in \mathcal{T}_n$, $\epsilon_n = o(1)$, and a non-negative, random function ϕ_n on $\mathcal{T}_n$ with $\text{supp}(\phi_n) \subset \mathcal{T}_n$, such that

$$\|\pi_1(\cdot|Y_{1:n}) - \phi_n(\cdot|Y_{1:n})\|_{L^1(\mu)} = o_{\mathbb{P}^n_*}(1),$$

and

$$\mu(\mathcal{T}_n)\mathbb{E}^n_* \left(\int_{\mathcal{T}_n} \phi_n^2(\vartheta)\mu(d\vartheta) \right) = O(z_n),$$

and $\exists K > 0$ such that

$$\mathbb{P}^n_*[\sup_{\vartheta\in\mathcal{T}_n} \phi_n(\vartheta) > e^{Kn\epsilon_n^2}] = o(1)$$

Assumption 2.7 is mild, and holds for instance if ϕ_n is a Gaussian distribution with mean $\hat{\vartheta}$ and variance $\mathbf{i}_*/n$ for some semi definite matrix $\mathbf{i}_*$, where $\hat{\vartheta} - \vartheta^* = O_{\mathbb{P}^n_*}(1/\sqrt{n})$. In this case, $\mathcal{T}_n = \{\|\vartheta - \vartheta^*\| \leq K_n/n\}$ with $K_n \to \infty$ arbitrarily slowly, since we have

$$\mathbb{E}_*^n\left(\int_{\mathcal{T}_n}\phi_n^2(\vartheta)\mu(d\vartheta)\right)\lesssim n^{d/2},\quad \mathrm{vol}(\mathcal{T}_n)\lesssim n^{-d/2}K_n^{d/2},$$
$$\sup_{\vartheta\in\mathcal{T}_n}\phi_n(\vartheta)\lesssim n^{d/2}=o(e^{K_n n\epsilon_n^2}),$$

as soon as $n\epsilon_n^2\gtrsim\log n$.

Theorem 2.27. *Let **Assumption** 2.7 hold with $n\epsilon_n^2\to\infty$, and assume that there exists $C>0$ such that, for any $\vartheta\in\mathcal{T}_n$, sets $S_n(\vartheta)\subset\mathcal{F}_2$ satisfying*

$$\bigcup_{\vartheta\in\mathcal{T}_n}\{\vartheta\}\times S_n(\vartheta)\subset V_n(\mathbb{P}_*^n,\epsilon_n),\quad \inf_{\vartheta\in\mathcal{T}_n}\Pi_2(S_n(\vartheta)|\vartheta)\geq e^{-Cn\epsilon_n^2}. \tag{2.35}$$

Assume also that there exist $L_n\to\infty$, $\mathcal{F}_{2,n}(\vartheta)\subset\mathcal{F}_2$ and $\psi_n:\mathcal{Y}^n\longrightarrow\{0,1\}$ such that

$$\sup_{\vartheta\in\mathcal{T}_n}\Pi_2(\mathcal{F}_{2,n}(\vartheta)^c|\vartheta)\leq e^{-L_n n\epsilon_n^2},\quad \textit{set}\quad \mathcal{U}_n=\bigcup_{\vartheta\in\mathcal{T}_n}\{\vartheta\}\times\mathcal{F}_{2,n}(\vartheta),$$

$$\mathbb{E}_*^n\psi_n=o(1),\qquad \sup_{\substack{(\vartheta,\eta)\in\mathcal{U}_n\\ d(\mathbb{P}^n_{\vartheta,\eta},\mathbb{P}^n_{\vartheta^*,\eta^*})>K_n\epsilon_n}}\mathbb{E}^n_{\vartheta,\eta}(1-\psi_n)\leq C'e^{-2L_n n\epsilon_n^2}, \tag{2.36}$$

for some $K_n,K>0$. Then, as $n\to\infty$,

$$\Pi_{cut}(B_\ell^c(K_n\epsilon_n)|Y_{1:n})=o_{\mathbb{P}_*^n}(1).$$

Remark 2.28. Typically $K_n=O(\sqrt{L_n})$, when the tests (2.36) are constructed as a supremum of local tests. Given some $L_n\to\infty$ for which the conditions are satisfied, we can choose $L_n\to\infty$ arbitrarily slowly, and consequently we can choose $K_n\to\infty$ arbitrarily slowly.

The proof of Theorem 2.27 which we present below is a rather simple adaptation of [48]. As in [49], we have simplified the common Kullback-Leibler neighbourhood assumption involving variances of the log-likelihood ratio using the technique of Lemma 6.26 therein.

Remark 2.29. By placing an additional assumption on neighbourhoods for higher-order Kullback-Leibler variations, the assumption on the existence of the sequence $L_n\to\infty$ can be replaced with an assumption that $L_n=L$ for a constant $L>1+C$ - the proof is similar but uses a different (though standard) technique for proving the evidence lower bound. In this case, we can choose K_n a constant.

Proof of Theorem 2.27. The proof is an adaptation of the proof on posterior contraction

rates, as in [48]. Under **Assumption** 2.7, it suffices to prove the claim when we replace the cut posterior Π_{cut} with the distribution

$$\tilde{\Pi}(d\vartheta, d\eta) = \phi_n(\vartheta)\mu(d\vartheta)\Pi_2(d\eta|\vartheta, Y_{1:n}).$$

Write also $\bar{B}_\ell = B_\ell^c(K_n\epsilon_n) \cap (\mathcal{T}_n \times \mathcal{F}^R)$, the subset of $B_\ell^c(K_n\epsilon_n)$ on which $\tilde{\Pi}$ is supported, and $\delta_n = \mu(\mathcal{T}_n)$.

Writing $J_n(\vartheta)$ for the random variable $\Pi_2(\bar{B}_\ell|\vartheta, Y_{1:n})$, we have $J_n(\vartheta) = \dfrac{N_n(\vartheta)}{D_n(\vartheta)}$ where

$$N_n(\vartheta) = \int_{\bar{B}_\ell} e^{l_n(\vartheta,\eta)-l_n(\vartheta^*,\eta^*)}\Pi_2(d(\eta|\vartheta); \quad D_n(\vartheta) = \int_{\mathcal{F}_2} e^{l_n(\vartheta,\eta)-l_n(\vartheta^*,\eta^*)}\Pi_2(d(\eta|\vartheta).$$

Write also $\tilde{\Pi}(\bar{B}_\ell|Y_{1:n}) = \tilde{\Pi}(\bar{B}_\ell|Y_{1:n})\psi_n + \tilde{\Pi}(\bar{B}_\ell|Y_{1:n})(1-\psi_n)\mathbb{1}_{\Omega_n^c} + \tilde{\Pi}(\bar{B}_\ell|Y_{1:n})(1-\psi_n)\mathbb{1}_{\Omega_n}$, where $\Omega_n = \{\sup_{\vartheta\in\mathcal{T}_n}\phi_n(\vartheta) \leq e^{Kn\epsilon_n^2}\}$. By the testing assumption, the first term vanishes in probability, while the second term vanishes under Assumption 2.7. For the remaining term, we have, for $\tilde{L}_n \to \infty$ to be chosen later, that

$$\begin{aligned}
&\mathbb{1}_{\Omega_n}(1-\psi_n)\int_{\mathcal{T}_n} J_n(\vartheta)\phi_n(\vartheta)\mu(d\vartheta) \\
&\qquad \leq \int_{\mathcal{T}_n} \mathbb{1}_{\{D_n(\vartheta)\leq e^{-(C+\tilde{L}_n)n\epsilon_n^2}\}}\phi_n(\vartheta)\mu(d\vartheta) \qquad (I_1) \\
&\qquad +(1-\psi_n)\int_{\mathcal{T}_n} \mathbb{1}_{\Omega_n} N_n(\vartheta)e^{(C+\tilde{L}_n)n\epsilon_n^2}\phi_n(\vartheta)\mu(d\vartheta), \qquad (I_2)
\end{aligned}$$

using the fact that $J_n(\vartheta) \leq 1$. We can bound $\mathbb{E}_*[I_1]$ as

$$\mathbb{E}_*[I_1] \leq \left(\int_{\mathcal{T}_n} \mathbb{P}_*\{D_n(\vartheta) \leq e^{-(C+\tilde{L}_n)n\epsilon_n^2}\}\mu(d\vartheta)\right)^{\frac{1}{2}} \left[E_*\left(\int_{\mathcal{T}_n}\phi_n^2(\vartheta)\mu(d\vartheta)\right)\right]^{\frac{1}{2}},$$

by an application of Cauchy-Schwartz. Now, noting that by Lemma 2.30 and under the assumption of sufficient prior mass on the $S_n(\vartheta)$, we have for any $\tilde{L}_n \to \infty$ that

$$\mathbb{P}_*\{D_n(\vartheta) \leq e^{-(\tilde{L}_n+C)n\epsilon_n^2}\} = O(\tilde{L}_n^{-1}),$$

and so

$$\mathbb{E}_*[I_1] \lesssim \sqrt{\delta_n}\tilde{L}_n^{-1/2}\left(\int_{\mathcal{T}_n}\phi_n^2(\vartheta)\mu(d\vartheta)\right)^{\frac{1}{2}} = o(1)$$

under **Assumption** 2.7, by taking $z_n = o(\tilde{L}_n)$. We bound $\mathbb{E}_*[I_2]$ by

$$\mathbb{E}_*[I_2] \leq e^{(C+\tilde{L}_n)n\epsilon_n^2}\mathbb{E}_*\left(\int_{\mathcal{T}_n}\int_{\bar{B}_\ell}[e^{l_n(\vartheta,\eta)-l_n(\vartheta^*,\eta^*)}(1-\psi_n)\phi_n]\Pi_2(d(\eta|\vartheta)\mu(d\vartheta)\right)$$
$$\leq e^{(C+\tilde{L}_n)n\epsilon_n^2}\left(\int_{\mathcal{T}_n}\int_{\bar{B}_\ell}\mathbb{E}_{\vartheta,\eta}[(1-\psi_n)\phi_n\mathbb{1}_{\Omega_n}]\Pi_2(d(\eta|\vartheta)\mu(d\vartheta)\right).$$

Using Fubini's theorem again, alongside the assumption on type II errors and on the sieves, and the deterministic bound on $\phi_n(\vartheta)$ over Ω_n from **Assumption** 2.7, we bound what precedes by

$$\leq e^{(K+C+\tilde{L}_n)n\epsilon_n^2}\left(\int_{\mathcal{T}_n}\int_{\bar{B}_\ell}\mathbb{E}_{\vartheta,\eta}[(1-\psi_n)(\mathbb{1}_{\eta\in\mathcal{F}_{2,n}(\vartheta)}+\mathbb{1}_{\eta\in\mathcal{F}_{2,n}(\vartheta)^c})]\Pi_2(d(\eta|\vartheta)\mu(d\vartheta)\right)$$
$$\leq e^{(K+C+\tilde{L}_n)n\epsilon_n^2}\left\{\left(\int_{\mathcal{T}_n}\int_{\bar{B}_\ell}C'e^{-L_nn\epsilon_n^2}\Pi_2(d(\eta|\vartheta)\mu(d\vartheta)\right)+\int_{\mathcal{T}_n}e^{-L_nn\epsilon_n^2}\mu(d\vartheta)\right\}$$
$$\leq e^{(K+C+\tilde{L}_n)n\epsilon_n^2}\left(C'e^{-L_nn\epsilon_n^2}+e^{-L_nn\epsilon_n^2}\right).$$

Choosing $\tilde{L}_n = o(L_n)$, we get the required convergence. □

Lemma 2.30 provides the lower bound on the denominator used in the proof of Theorem 2.27. The proof is standard but we include it for completeness, it follows almost exactly the proof of Lemma 6.26 in [49].

Lemma 2.30. *Let* $A_n(\vartheta) = \{D_n(\vartheta) \geq \Pi_2(S_n(\vartheta)|\vartheta)e^{-n\tilde{L}_n\epsilon_n^2}\}$ *where* $\tilde{L}_n \to \infty$ *and* $S_n(\vartheta)$ *is such that there exists* $\mathcal{T}_n$ *with*

$$\bigcup_{\vartheta\in\mathcal{T}_n}\{\vartheta\}\times S_n(\vartheta) \subset V_0(P_0,\epsilon_n).$$

Then

$$\sup_{\vartheta\in\mathcal{T}_n}\mathbb{P}_*(A_n(\vartheta)^c) = O(\tilde{L}_n^{-1}) = o(1)$$

as $n\to\infty$.

Proof. We show that, for each ϑ, with $\mathbb{P}_*$-probability tending to one,

$$\int_{\mathcal{F}_2}e^{l_n(\vartheta,\eta)}-e^{l_n(\vartheta^*,\eta^*)}\Pi_2(d(\eta|\vartheta)\geq\Pi_2(S_n(\vartheta)|\vartheta)e^{-nM_n\epsilon_n^2},$$

for any $M_n\to\infty$. It suffices to show the above equation holds when we restrict the integral to $S_n(\vartheta)\subset\mathcal{F}_2$, for which we have $\{\vartheta\}\times S_n(\vartheta)\subset V_n(\mathbb{P}_*,\epsilon_n)$. By dividing both sides by $\Pi_2(S_n(\vartheta)|\vartheta)$, we see that it suffices to show that, for any probability measure ν

supported on $S_n = S_n(\vartheta)$, that

$$\int_{S_n} e^{l_n(\vartheta,\eta)} - e^{l_n(\vartheta^*,\eta^*)} d\nu(\eta) \geq e^{-nM_n\epsilon_n^2}.$$

By applying the logarithm to both sides and using Jensen's inequality, it suffices to show that, with high $\mathbb{P}_*$ probability,

$$Z := \int_{S_n} l_n(\vartheta,\eta) - l_n(\vartheta^*,\eta^*) d\nu(\eta) \geq -nM_n\epsilon_n^2.$$

Arguing as in [49], we obtain

$$\mathbb{P}_*(Z < -nM_n\epsilon_n^2) = O(M_n^{-1}) = o(1)$$

as $n \to \infty$ for any $M_n \to \infty$. Since the bound does not depend on ϑ and uses only that $\{\vartheta\} \times S_n(\vartheta) \subset V_0(P_0, \epsilon_n)$, we obtain the desired bound on the supremum.

□

2.10 Proofs of technical results

Section 2.10 is devoted to a number of technical proofs which are required the main results, but are reasonably standard in their approach.

In Section 2.10.1, we prove that the Fisher information matrix is invertible for general, discrete state-space, discrete observation HMMs. This is necessary to apply the results of [12] and [28] in the proofs of Theorems 2.3 and 2.4.

In Section 2.10.2, we gather some properties of the Fisher information matrix, showing a Cramér-Rao bound for estimation in HMMs, and showing a local uniform convergence result for the expected information from n observations.

In Section 2.10.3, we collect technical lemmas used for the deconvolution argument which forms the key part of the proof of Theorem 2.6.

In Section 2.10.4, we state a result which implies the existence of an admissible partition.

In Section 2.10.5, we collect technical lemmas which feature in the proofs of Theorems 2.10 and 2.13

2.10.1 Non-singularity of Fisher Information

In what follows, we establish invertibility of the Fisher Information matrix for general multinomial Hidden Markov models.

Proposition 2.31. *Consider a multinomial HMM with latent states $X_t \in \{1, \dots, R\}$ and discrete observations Y_t taking values in the set of basis vectors of $\mathbb{R}^\kappa$, which we denote $\{e_1, \dots, e_M\}$. Denote by $Q \in \mathbb{R}^{R\times R}$ the transition matrix of (X_t), and let $\Omega \in \mathbb{R}^{\kappa\times R}$ be the matrix whose R columns $\omega_r = (\omega_{mr})_{m=1}^M$ are the emission probabilities for the r^{th} state, with $r \in [R]$.*

Denote $\theta = (Q, \Omega) \in \mathbb{R}^p$ the HMM parameter, and write $J(\theta)$ for the Fisher Information matrix with entries given by

$$J(\theta)_{ij} = -\lim_{n\to\infty} \underbrace{\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(Y_1, \dots, Y_n)\right]}_{-J_n(\theta)_{ij}}.$$

Then, if Q, Ω have rank R, J is non-singular.

The idea of the proof is to exhibit estimators with L^2 risk of order $n^{-1/2}$. We then show that the local asymptotic minimax result of [**?**] implies that the existence of such estimators guarantees a non-singular Fisher information. We use the spectral estimators proposed in [8], the control of which is established in [**?**].

Proof of Proposition 2.31. By an application of the van Trees inequality, analogously to the derivation preceding Theorem 16Equation (12) in Theorem 4 of [**?**], we obtain for the HMM that (for p the parameter dimension)

$$\int_{B_p(0,1)} \mathbb{E}^{(n)}_{\theta^*+ch/\sqrt{n}}\left[\left(U^T(\hat\theta^{(n)} - (\theta^* + \frac{ch}{\sqrt{n}}))\right)^2\right] q(h)dh$$
$$\geq U^T\left(\frac{n}{c^2}J_q + n\int_{B_p(0,1)}\frac{1}{n}J_n(\theta^* + \frac{ch}{\sqrt{n}})q(h)dh\right)^{-1}U,$$

which holds for any vector U. Here J_n is the joint Fisher information for n observations as in Proposition 2.40, and q is a density on $\mathbb{R}^p$ such that $J_q := \int_{\mathbb{R}^P}\nabla q\nabla q^T\frac{\mathbb{1}_{\{q>0\}}}{q}dx$ is

non-singular. Rescaling, we get that, for any vector U,

$$\int_{B_p(0,1)} \mathbb{E}^{(n)}_{\theta^*+ch/\sqrt{n}} \left[\left(\sqrt{n} U^T (\hat{\theta}^{(n)} - (\theta^* + \frac{ch}{\sqrt{n}})) \right)^2 \right] q(h)dh$$
$$\geq U^T \left(\frac{1}{c^2} J_q + \int_{B_p(0,1)} \frac{1}{n} J_n(\theta^* + \frac{ch}{\sqrt{n}}) q(h)dh \right)^{-1} U.$$

Taking a limit in n and applying Lemma 2.40 stated below gives that the limit inferior of the left hand side is at least

$$\lim_n U^T \left(\frac{1}{c^2} J_q + \int_{B_p(0,1)} \frac{1}{n} J_n(\theta^* + \frac{ch}{\sqrt{n}}) q(h)dh \right)^{-1} U = U^T \left(\frac{1}{c^2} J_q + J(\theta^*) \right)^{-1} U.$$

Call the matrix on the right hand side which we invert J_c. It is indeed invertible for sufficiently large c, since J_q is invertible and the set of invertible matrices is open. Denote its matrix square root by $J_c^{\frac{1}{2}}$. Now suppose $\exists V^*$ such that $(V^*)^T J(\theta^*) V^* = 0$. Then, by writing $V = J_c^{-\frac{1}{2}} U$, we get

$$\sup_{\|U\|=1} U^T J_c^{-1} U = \sup_{\|U\| \neq 0} \frac{U^T J_c^{-1} U}{U^T U}$$
$$= \sup_{\|V\| \neq 0} \frac{V^T V}{V^T J_c V} \geq \frac{V^{*T} V^*}{V^{*T} J_c V^*} = Ac^2,$$

with A a fixed constant not depending on c. Taking the limit as $c \to \infty$ gives that (upper bounding also the averaging over the law qdh by the supremum over h)

$$\liminf_{c\to\infty} \sup_{\|U\|=1} \liminf_{n\to\infty} \sup_{\|h\|<1} \mathbb{E}^{(n)}_{\theta^*+ch/\sqrt{n}} \left[\left(\sqrt{n} U^T (\hat{\theta}^{(n)} - (\theta^* + \frac{ch}{\sqrt{n}})) \right)^2 \right] = +\infty,$$

or equivalently,

$$\liminf_{c\to\infty} \sup_{\|U\|=1} \liminf_{n\to\infty} \sup_{\|\theta-\theta^*\|<\frac{c}{\sqrt{n}}} \mathbb{E}^{(n)}_{\theta} \left[\left(\sqrt{n} U^T (\hat{\theta}^{(n)} - \theta) \right)^2 \right] = +\infty,$$

which contradicts the local uniform bound of Proposition 2.32. □

The following proposition establishes the existence of an estimator with suitable risk, as required for the arguments of Proposition 2.10.1.

Proposition 2.32. *Let $\theta = (Q, \Omega)$ be the parameter for the HMM described in Proposition 2.31 and let θ^* be such that the identifiability conditions of **Assumption** 2.1 hold. Then*

there exists an estimator $\hat{\theta}$ *such that, for any* $\epsilon > 0$ *sufficiently small, we have, as* $n \to \infty$,

$$\sup_{\|\theta-\theta^*\|<\epsilon} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \leq \frac{C}{n}(1 + o(1)),$$

up to label-swapping.

Proof. Let $\hat{\theta}$ be the spectral estimator constructed in Section 2.10.1.1. We have that

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = \int_0^\infty t\mathbb{P}_\theta(|\hat{\theta} - \theta\| > t)dt.$$

Setting $t = \frac{Cx}{\sqrt{n}}$ we get

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 = \int_0^\infty \frac{C^2 x}{n} \mathbb{P}_\theta(\|\hat{\theta} - \theta\| > \frac{Cx}{\sqrt{n}})dx.$$

The bound of Lemma 2.38 is valid for $x \geq 1$, and so

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \leq \frac{C^2}{n} + \frac{C^2}{n} \int_0^\infty x e^{-x^2} dx \leq \frac{3C^2}{2n}.$$

Since the bound is valid on a small neighbourhood around θ, it holds in supremum over that neighbourhood. □

2.10.1.1 Construction of spectral estimators

Spectral estimation of HMMs has been addressed by a number of previous works [1,2,7,8, 27,58]. In [1], the authors exhibit estimators for emissions in parametric HMMs with a $\sqrt{n}$ rate in probability, but we require convergence in expectation. In [27], a convergence in expectation is shown but the rate contains a logarithmic factor which is not sufficient for our use in the proof of Proposition 2.10.1. In [2], the authors exhibit estimators for which the concentration is sub-Gaussian, which would permit integration to an in-expectation bound, but their work only concerns two-state HMMs.

We will construct estimators which are very similar to those constructed in these works. To get the convergence properties we require, we use as a starting point the sub-Gaussian concentration of the empirical estimator of $E_{ijk} = \mathbb{P}(Y_1 = i, Y_2 = j, Y_3 = k)$. To extract the HMM parameters, we follow the approach set out in [8], who prove guarantees for their *Tensor power method* which outputs, from an estimate $\hat{T}$ of a given symmetric, orthogonal tensor T, estimates of the associated eigenvalues and eigenvectors whose error is of the same order as the input error $\|\hat{T} - T\|$.

Since T is required to be symmetric and orthogonal, we cannot directly apply their algorithm to extract estimates with appropriate concentration properties from estimates $\hat{E}$ of the tensor E. Instead, we construct an estimate $\hat{T}$ of a symmetrised, orthogonalised version T of E, the eigenvalues and eigenvectors of which can then be extracted through the Tensor power method and related back to the parameters of the HMM. This approach is similar to that taken in [6, 57]; [57] consider the setting of Spherical Gaussians, which requires a different symmetrisation step but whose orthogonalisation procedure we adopt here. Our symmetrisation step is instead based on Theorem 3.6 of [8].

The construction First define the following matrices and tensors:

$$E_\theta^{(12)} = \mathbb{E}_\theta[Y_1 \otimes Y_2], \quad E_\theta^{(13)} = \mathbb{E}_\theta[Y_1 \otimes Y_3], \quad E_\theta^{(123)} = \mathbb{E}_\theta[Y_1 \otimes Y_2 \otimes Y_3],$$

$$\hat{E}^{(12)} = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_i \otimes Y_{i+1}, \quad \hat{E}^{(13)} = \frac{1}{n-2}\sum_{i=1}^{n-2} Y_i \otimes Y_{i+2}, \quad \hat{E}^{(123)} = \frac{1}{n-2}\sum_{i=1}^{n-2} Y_i \otimes Y_{i+1} \otimes Y_{i+2}.$$

Next, we will symmetrise about Y_3 as follows: define

$$\tilde{Y}_1 = \underbrace{E_\theta^{(21)}(E_\theta^{(12)})^{-1}}_{A_\theta} Y_1, \quad \tilde{Y}_2 = \underbrace{E_\theta^{(31)}(E_\theta^{(21)})^{-1}}_{B_\theta} Y_2,$$

where $E_\theta^{(ij)}$ is the transpose of $E_\theta^{(ji)}$. Define $\tilde{E}_\theta^{(ij)}$ and $\tilde{E}_\theta^{(123)}$ in the same way as $E_\theta^{(ij)}, E_\theta^{(123)}$, but with Y_1, Y_2 replaced by $\tilde{Y}_1, \tilde{Y}_2$ respectively; these matrices are then symmetric. Define estimators $\check{E}^{(12)}$, $\check{E}^{(13)}$ and $\check{E}^{(123)}$ of the symmetric quantities as follows: Split the sample in two and, from the first sample, produce estimates $\hat{E}^{(\cdot)}$. This yields estimates $\hat{A}$, $\hat{B}$ of A_θ, B_θ. Then define

$$\check{E}^{(12)} = \frac{1}{n - \lceil \frac{n}{2} \rceil} \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes \hat{B}Y_{i+1}, \quad \check{E}^{(13)} = \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes Y_{i+2},$$

and

$$\check{E}^{(123)} = \frac{1}{n - \lceil \frac{n}{2} \rceil - 1} \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes \hat{B}Y_{i+1} \otimes Y_{i+2}.$$

We remark that Theorem 3.6 of [8] gives us the decomposition of the symmetric tensor $\tilde{E}_\theta^{(123)}$ as

$$\tilde{E}_\theta^{(123)} = \sum_{i=1}^{R} p_i \mu_i \otimes \mu_i \otimes \mu_i,$$

where $p_i = \mathbb{P}_\theta(X_2 = i)$ is taken to be the stationary distribution and $\mu_i = \mathbb{P}_\theta(Y_3 =$

$\cdot|X_2 = i) = (\Omega Q^T)_i$. In a slight abuse of notation, we suppress the dependence on θ in our notation for p_i, μ_i.

This justifies that indeed $\tilde{E}_\theta^{(123)}$ is a symmetric tensor. However, the μ_i are generally not orthonormal, which is required for the application of Theorem 5.1 of [8] (the tensor power method). Define, for a three-way tensor T and matrices P, Q, R,

$$[T(P,Q,R)]_{ijk} = \sum_r \sum_s \sum_t T_{rst} P_{ri} Q_{sj} R_{tk}.$$

Next, define $\tilde{E}_{W,\theta}^{(123)}$ as in Section 4.3.1 of [8] as follows: Take $W = W_\theta$ as the matrix of left orthonormal singular vectors of $\tilde{E}^{(12)}$ so that $W^T \tilde{E}_\theta^{(12)} W = I$ and define

$$\mu_i^{(W)} = \sqrt{p_i} W^T \mu_i,$$

which are orthonormal. Then take

$$\tilde{E}_{W,\theta}^{(123)} = \tilde{E}_\theta^{(123)}(W,W,W) = \sum_{i=1}^R \frac{1}{\sqrt{p_i}} \mu_i^W \otimes \mu_i^W \otimes \mu_i^W.$$

Now set $\hat{W}$ as the matrix of left orthonormal singular vectors of $\check{E}^{(12)}$, with corresponding $\check{E}_{\hat{W}}^{(123)} = \check{E}^{(123)}(\hat{W}, \hat{W}, \hat{W})$.

At this point, we have at hand an estimate $\check{E}_{\hat{W}}^{(123)}$ of the orthogonal, symmetric tensor $\tilde{E}_{W,\theta}^{(123)}$ so we apply the Tensor power method of [8]. This returns estimates $\hat{\mu}_I^W$, $\hat{p}_i$ of μ_i^W and p_i. Now set $\hat{\mu}_i = \frac{1}{\sqrt{\hat{p}_i}} \hat{W}^\dagger \hat{\mu}_i^W$, which estimates (up to label-swapping) the columns of ΩQ^T.

By repeating the above procedure, but instead symmetrising about Y_2 (so defining a corresponding $\tilde{Y}_1, \tilde{Y}_3$) we produce estimates (up to label-swapping) of the columns of Ω. By permuting these estimates relative to some consistent estimator (such as the MLE, which is consistent without assuming invertibility of the Fisher Information (see e.g. [32]), we finally produce estimates $(\hat{Q}, \hat{\Omega})$ of (Q, Ω).

In the following section, we control the error introduced at each stage of this construction.

2.10.1.2 Control of spectral estimators

The control on the estimators constructed in 2.10.1.1 is ultimately proved in Lemma 2.38; this control is established through the following sequence of technical lemmas. The

estimator we will use are those constructed by the tensor power method in [8]. It is convenient to assume $\kappa = R$ in what follows so that certain matrices are square - the extension to the general case is straightforward using SVD and low rank approximations but is omitted for ease of exposition.

The next lemma makes explicit the sub-Gaussian concentration of empirical tensors, when the data is generated according to a geometrically ergodic Markov chain. It resembles closely Lemma 1 of [2].

Lemma 2.33 (Concentration of empirical tensors)**.** *Define the matrices and tensors*

$$E_\theta^{(12)} = \mathbb{E}_\theta[Y_1 \otimes Y_2], \quad E_\theta^{(13)} = \mathbb{E}_\theta[Y_1 \otimes Y_3], \quad E_\theta^{(123)} = \mathbb{E}_\theta[Y_1 \otimes Y_2 \otimes Y_3].$$

Define further

$$\hat{E}^{(12)} = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_i \otimes Y_{i+1}, \quad \hat{E}^{(13)} = \frac{1}{n-2}\sum_{i=1}^{n-2} Y_i \otimes Y_{i+2},$$
$$\hat{E}^{(123)} = \frac{1}{n-2}\sum_{i=1}^{n-2} Y_i \otimes Y_{i+1} \otimes Y_{i+2}.$$

Then there exists a constant $C = C(R, Q^*)$, such that, for all $x \geq 1$, sufficiently small $\epsilon > 0$, and any superscripts,

$$\sup_{\|\theta - \theta^*\| \leq \epsilon} \mathbb{P}_\theta(\|\hat{E}^{(\cdot)} - E_\theta^{(\cdot)}\|) > \frac{Cx}{\sqrt{n}}) \leq e^{-x^2}.$$

Proof. The proof is essentially the same as the proof of Lemma 1 in [2]. In that paper, the authors use the concentration result of [85] which provides a bound in terms of the pseudo-spectral gap γ_{ps} of the chain

$$Z_n = (X_n, X_{n+1}, X_{n+2}, Y_n, Y_{n+1}, Y_{n+2}),$$

together with Proposition 3.4 of the same reference which relates γ_{ps} to the mixing time t_{mix} of the same chain. Arguing as in [2], we see that t_{mix} is uniformly bounded on a neighbourhood of the true parameter, since the eigenvalues of the transition matrix vary continuously in its entries. For the remainder of the proof, we follow the argument of [2]. □

An important step in the application of the Tensor power method is to pre-process the empirical tensors controlled in Lemma 2.33 so that they are symmetric and orthogonal.

The symmetrisation step involves a linear transformation of the first two observations so that the three-way expectation is symmetric - see Theorem 3.6 of [8]. Note that $E_\theta^{(12)} = \sum_{i=1}^R p_i v_i \otimes w_i$ where $p_i = \mathbb{P}_\theta(X_2 = i)$, $v_i = \mathbb{P}_\theta(Y_1 = \cdot | X_2 = i)$ and $w_i = \mathbb{P}_\theta(Y_2 = \cdot | X_2 = i)$, and by writing v_i, w_i in terms of Q and Ω, we see that $E_\theta^{(12)}$ has rank R and so is indeed invertible.

Lemma 2.34 (Symmetrisation preserves concentration)**.** *Define*

$$\tilde{Y}_1 = \underbrace{E_\theta^{(21)}(E_\theta^{(12)})^{-1}}_{A_\theta} Y_1, \quad \tilde{Y}_2 = \underbrace{E_\theta^{(31)}(E_\theta^{(21)})^{-1}}_{B_\theta} Y_2,$$

where $E_\theta^{(ij)}$ is the transpose of $E_\theta^{(ji)}$. Define $\tilde{E}_\theta^{(ij)}$ and $\tilde{E}_\theta^{(123)}$ in the same way as $E_\theta^{(ij)}, E_\theta^{(123)}$ in Lemma 2.33 but with Y_1, Y_2 replaced by $\tilde{Y}_1, \tilde{Y}_2$ respectively; these matrices are then symmetric. Define estimators $\check{E}^{(12)}$, $\check{E}^{(13)}$ and $\check{E}^{(123)}$ of the symmetric quantities as follows: Split the sample in two and, from the first sample, produce estimates $\hat{E}^{(\cdot)}$. This yields estimates $\hat{A}$, $\hat{B}$ of A_θ, B_θ. Then define

$$\check{E}^{(12)} = \frac{1}{n - \lceil \frac{n}{2} \rceil} \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes \hat{B}Y_{i+1}, \quad \check{E}^{(13)} = \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes Y_{i+2},$$

and

$$\check{E}^{(123)} = \frac{1}{n - \lceil \frac{n}{2} \rceil - 1} \sum_{i=\lceil \frac{n}{2} \rceil}^{n-1} \hat{A}Y_i \otimes \hat{B}Y_{i+1} \otimes Y_{i+2}.$$

Then there exists a constant $\tilde{C} = \tilde{C}(R, W^*)$ such that for all $x \geq 1$, sufficiently small $\epsilon > 0$ and any superscripts,

$$\sup_{\|\theta - \theta^*\| \leq \epsilon} \mathbb{P}_\theta(\|\check{E}^{(\cdot)} - \tilde{E}_\theta^{(\cdot)}\| > \frac{\tilde{C}x}{\sqrt{n}}) \leq e^{-x^2}.$$

Proof. Write $\mathcal{E}$ for the event that $\hat{E}^{(12)}$ is invertible. Choose r such that the ball of radius r around θ^* is contained in the set of θ for which the corresponding $Q = Q(\theta)$ is invertible. Then for all θ in a ball of radius at most $\frac{r}{2}$ inside the ball on which Lemma 2.33 holds, the ball of radius $\frac{r}{2}$ centred at any $Q(\theta)$ contains only those θ' for which $Q'(\theta')$ is invertible. Hence, by choosing $x = c\sqrt{n}$ where $c = c(r)$, we can apply Lemma 2.33 for sufficiently large n to get that $\mathbb{P}_\theta(\mathcal{E}^c) \leq e^{-c^2 n} = o(1)$ as $n \to \infty$.

Working on $\mathcal{E}$, Lipschitz continuity of matrix inversion and multiplication implies that $\hat{A}$ and $\hat{B}$ satisfy an exponential bound of the kind shown for $\hat{E}^{(\cdot)}$, as a corollary of Lemma 2.33. This implies that $\check{E}^{(12)}$ and $\check{E}^{(13)}$ satisfy a corresponding bound when approxi-

mating the analogous quantities with $\hat{A}, \hat{B}$ replaced with A, B. These quantities in turn approximate $\tilde{E}^{(\cdot)}$ by the same arguments as presented in Lemma 2.33.

For the estimates $\hat{A}$ and $\hat{B}$, we can formally take $\hat{A} = \hat{B} = I$ on the event that $\hat{E}^{(12)}$ is not invertible.

We deduce that for $\tilde{C}$ suitably chosen

$$\mathbb{P}_\theta(\|\check{E}^{(\cdot)} - \tilde{E}_\theta^{(\cdot)}\| > \frac{\tilde{C}x}{\sqrt{n}}) \leq e^{-x^2} + o(1),$$

and, by enlarging the constant by some factor greater than one, we get, for sufficiently large n, any $x \geq 1$, and sufficiently small $\epsilon > 0$, that

$$\sup_{\|\theta - \theta^*\| \leq \epsilon} \mathbb{P}_\theta(\|\check{E}^{(\cdot)} - \tilde{E}_\theta^{(\cdot)}\| > \frac{\tilde{C}x}{\sqrt{n}}) \leq e^{-x^2}.$$

□

Lemma 2.34 tells us we can construct an approximation to the symmetrised tensors $\tilde{E}^{(\cdot)}$. In the next result, we show that we can further extend this to an approximation to a symmetrised, orthogonalised tensor, at which point the Tensor power method may be applied to see that the parameter values extracted from this approximate tensor well approximate the parameter values associated to the true orthogonal, symmetric tensor (which are themselves related in an analytic sense to the parameters of interest). The orthogonalisation or 'whitening' step is detailed in Lemma 2.35 and follows closely the ideas of appendix C of [57].

Before we proceed, we remark that Theorem 3.6 of [8] gives us the decomposition of the symmetric tensor $\tilde{E}_\theta^{(123)}$ as

$$\tilde{E}_\theta^{(123)} = \sum_{i=1}^{R} p_i \mu_i \otimes \mu_i \otimes \mu_i,$$

where $p_i = \mathbb{P}_\theta(X_2 = i)$ is taken to be the stationary distribution and $\mu_i = \mathbb{P}_\theta(Y_3 = \cdot | X_2 = i) = (\Omega Q^T)_i$. In a slight abuse of notation, we suppress the dependence on θ in our notation for p_i, μ_i.

This justifies that indeed $\tilde{E}_\theta^{(123)}$ is a symmetric tensor. However, the μ_i are generally not orthonormal, which is required for the application of Theorem 5.1 of [8] (the tensor power

method). First, let us define for a three-way tensor T and matrices P, Q, R,

$$[T(P,Q,R)]_{ijk} = \sum_r \sum_s \sum_t T_{rst} P_{ri} Q_{sj} R_{tk}.$$

We are now ready to show that the concentration of Lemma 2.33 is preserved on further reduction:

Lemma 2.35 (Whitening preserves concentration)**.** *Define $\tilde{E}^{(123)}_{W,\theta}$ as in Section 4.3.1 of [8] as follows: Take $W = W_\theta$ as a linear transformation such that $W^T \tilde{E}^{(12)}_\theta W = I$ (made explicit in the proof) and define*

$$\mu_i^{(W)} = \sqrt{p_i} W^T \mu_i,$$

which are orthonormal. Then take

$$\tilde{E}^{(123)}_{W,\theta} = \tilde{E}^{(123)}_\theta(W,W,W) = \sum_{i=1}^R \frac{1}{\sqrt{p_i}} \mu_i^W \otimes \mu_i^W \otimes \mu_i^W.$$

Then there exists an estimate $\hat{W}$, with corresponding $\check{E}^{(123)}_{\hat{W}} = \check{E}^{(123)}(\hat{W},\hat{W},\hat{W})$ such that, for some $\tilde{C}' = \tilde{C}'(R, Q^*)$, for any $x \geq 1$ and sufficiently small $\epsilon > 0$

$$\sup_{\|\theta - \theta^*\| \leq \epsilon} \mathbb{P}_\theta(\|\check{E}^{(123)}_{\hat{W}} - \tilde{E}^{(123)}_{W,\theta}\| > \frac{\tilde{C}' x}{\sqrt{n}}) \leq e^{-x^2}.$$

Proof. The proof essentially follows Appendix C.6 in [57]. In particular, we take $W, \hat{W}$ to be the matrix of left orthonormal singular vectors of $\tilde{E}^{(12)}$ and $\check{E}^{(12)}$ respectively. Then their Lemma 12 gives that

$$\|\check{E}^{(123)}_{\hat{W}} - \tilde{E}^{(123)}_{W,\theta}\| \leq \|\check{E}^{(123)}(\hat{W},\hat{W},\hat{W}) - \tilde{E}^{(123)}_\theta(\hat{W},\hat{W},\hat{W})\| + \frac{6}{\sqrt{p_{\min}}} \frac{\|\check{E}^{(12)} - \tilde{E}^{(12)}_\theta\|}{\zeta_R(\tilde{E}^{(12)}_\theta)}$$

where $p_{\min}$ is the smallest element of the stationary distribution (which under **Assumption** 2.1 is uniformly bounded below on a ball around the true parameter) and $\zeta_R(\tilde{E}^{(12)})$ is the R^{th} largest singular value of $\tilde{E}^{(12)}_\theta$, which is also positive under this assumption at $\theta = \theta^*$ (and uniformly bounded below on some neighbourhood of the true parameter) as $\tilde{E}^{(12)}_{\theta^*}$ has full rank. Then, on an event $\mathcal{E}$ of arbitrarily large probability, the map $M \mapsto M(\hat{W},\hat{W},\hat{W})$ is Lipschitz with sufficiently large constant $K > 0$; on $\mathcal{E}$ the first term on the right-hand side is then bounded by $K\|\check{E}^{(123)} - \tilde{E}^{(123)}_\theta\|$ and the result follows. □

Having constructed a theoretical orthogonal tensor $\tilde{E}^{(123)}_{W,\theta}$ based on the (unknown) distribution of our observations, and having at hand an estimator $\check{E}^{(123)}_{\hat{W}}$ for this tensor with good concentration properties, Theorem 5.1 of [8] gives guarantees for an algorithm which recovers the μ_i^W (and the w_i). Note that the probability of failure (their η) for the algorithm described may be tuned by the user, and so for a theoretical estimator we may consider the algorithm in the limit $\eta \to 0$ (corresponding to running the algorithm for an arbitrarily long number of iterations). A direct application of the theorem then provides the following result:

Lemma 2.36 (Recovery of orthogonalised parameters)**.** *Suppose Algorithm 1 of [8] is called iteratively, as in the statement of Theorem 5.1 of the same reference. Let $\hat{\mu}_i^W$ be the estimates of μ_i^W returned by this process and $\hat{p}_i$ the estimates of p_i returned (by also applying $x \mapsto x^{-2}$ to the estimated weights). Then there exists $\tilde{C}'' = \tilde{C}''(R, Q)$ and a permutation $\tau \in \mathcal{S}_R$ such that, for all $x \geq 1$ and sufficiently small $\epsilon > 0$,*

$$\sup_{\|\theta-\theta^*\|\leq\epsilon} \mathbb{P}_\theta\Big(\max_{i=1,\dots,R}\|\hat{\mu}_i^W - {}^\tau\mu_i^W\| > \frac{\tilde{C}''x}{\sqrt{n}}, \max_{i=1,\dots,R}\|\hat{p}_i - {}^\tau p_i\| > \frac{\tilde{C}''x}{\sqrt{n}}\Big) \leq e^{-x^2}.$$

The previous result follows easily from the fact that the order of error in the output corresponds to the order of the error in the input tensors, together with the smoothness of $x \mapsto x^{-2}$, so the proof is omitted. It remains to show that these estimates can be 'de-whitened' to produce estimates for the vector of interest μ_i, which we show in the following result.

Lemma 2.37 (Recovery of original parameters)**.** *Let $W^\dagger$ be the Moore-Penrose pseudo-inverse of W^T. Then*

$$\mu_i^{(\theta)} = \mu_i = \frac{1}{\sqrt{p_i}} W^\dagger \mu_i^W,$$

and the $\hat{\mu}_i$ constructed as

$$\hat{\mu}_i = \frac{1}{\sqrt{\hat{\nu}_i}} \hat{W}^\dagger \hat{\mu}_i^W$$

satisfy

$$\sup_{\|\theta-\theta^*\|\leq\epsilon} \mathbb{P}_\theta\Big(\max_{i=1,\dots,R}\|\hat{\mu}_i - {}^\tau\mu_i^{(\theta)}\| > \frac{Cx}{\sqrt{n}}\Big) \leq e^{-x^2},$$

for some $\tau \in \mathcal{S}_R$ and $C = C(R, Q^*)$, for all $x \geq 1$ and for sufficiently small $\epsilon > 0$.

Proof. The first part of the statement is given by Theorem 4.3 of [8]. The control over $\hat{\mu}_i$ then follows from the control over $\check{E}^{(12)}$ established in Lemma 2.34 and the Lipschitz continuity of the singular vectors and Moor-Penrose pseudo-inverse on a neighbourhood of

the true parameter, as well as the control over $\hat{\nu}_i$ and $\hat{\mu}_i^W$ established in Lemma 2.37. □

The estimates produced target (up to label-swapping) the columns of ΩQ^T. By adapting the procedure to symmetrise about the second view (rather than the third), we can produce estimates (up to label-swapping) of Ω. Although the labelling is arbitrary, the permutation of Lemma 2.37 may differ in the two cases.

To alleviate this issue, we use any consistent estimator (such as the maximum-likelihood estimator, which is consistent without assumptions on the Fisher information, see Theorem 1 of [32]). Because the transition matrix Q has full rank, the μ_i (and associated vectors from symmetrisation about the second view) are well-separated, and so we can fix a labelling relative to this estimator.

Lemma 2.38 (Fixing of labels)**.** *Let $\hat{\Xi}_3$ be the matrix whose columns are the $\hat{\mu}_i$ as in Lemma 2.37 (with symmetrisation about the third view as detailed) and $\hat{\Xi}_2$ be the corresponding matrix when symmetrising about the second view. Moreover, let $\check{\Xi}_3$,$\check{\Xi}_2$ be the corresponding estimates formed from the MLE, that is*

$$\check{\Xi}_3 = \check{\Omega}_{MLE}\check{Q}_{MLE}^T, \qquad \check{\Xi}_2 = \check{\Omega}_{MLE}.$$

For $i = 3, 2$, pick $\hat{\tau}_i \in \mathcal{S}_R$ such that

$$\|^{\hat{\tau}_i}\hat{\Xi}_i - \check{\Xi}_i\| = \min_{\tau \in \mathcal{S}_R} \|^{\tau}\hat{\Xi}_i - \check{\Xi}_i\|,$$

where $\|\cdot\|$ is the max of the squared norms of the columns. Define

$$\hat{\Omega} = {}^{\hat{\tau}_2}\hat{\Xi}_2,$$

and

$$\hat{Q}^T = \hat{\Omega}^{-1\,\hat{\tau}_3}\hat{\Xi}_3.$$

Recall $\theta = (Q, \Omega)$ and define $\hat{\theta} = (\hat{Q}, \hat{\Omega})$ .Then there exists a $\tau \in \mathcal{S}_R$ (the permutation accounting for the difference between the truth and the MLE) such that, for all $\epsilon > 0$ small enough,

$$\sup_{\|\theta - \theta^*\| < \epsilon} \mathbb{P}_\theta\Big(\max_{i=1,\dots,R} \|\hat{\theta}_i - {}^{\tau}\theta_i\| > \frac{Cx}{\sqrt{n}}\Big) \leq e^{-x^2},$$

for some $C = C(R, Q)$.

2.10.2 Fisher Information and asymptotic lower bound

The following result gives a lower bound on the asymptotic variance of regular estimators and is used in the proof of Lemma 2.24.

Lemma 2.39. *Let $\hat{\theta}_n$ be a regular estimator in the histogram model $\mathcal{M}(\mathcal{I}_M)$ with M fixed. Let Z denote the random variable whose law is equal to the limit distribution of the scaled and centred estimates $\sqrt{n}(\hat{\theta}_n - \theta^*)$ under $\mathbb{P}_*$ as $n \to \infty$. Then*

$$Cov(Z) \geq J(\theta^*),$$

where $J(\theta)$ is the Fisher information matrix at the parameter θ.

Proof. Consider estimation of the one-dimensional parameter $\lambda^T\theta$. Recall from Lemma 2.40 that for any sequence $\theta_n \to \theta$, we have $J(\theta) = \lim_{n\to\infty} \frac{1}{n} J_n(\theta_n)$, with J_n the joint Fisher information for n observations, as defined in Lemma 2.40.

We follow the arguments of Gill and Levit [50]. Let π be a fixed prior density on $[-1, 1]$ and $J(\pi) = \mathbb{E}[(\log \pi(\theta))'^2]$. For a given $H > 0$, let $\pi(H, n)$ be the rescaling of this prior to the interval $A = [\theta_0 - n^{-\frac{1}{2}}H, \theta_0 + n^{-\frac{1}{2}}H]$ for given $H > 0$. Then, applying the van Trees inequality, we get

$$\mathbb{E}[\lambda^T\hat{\theta}^n - \lambda^T\theta]^2 \geq \frac{1}{\mathbb{E}\lambda^T J_n(\theta)\lambda + nJ(\pi)/H^2},$$

where the left hand expectation is taken over the joint law of the parameter and the data given the parameter (with the parameter distributed according to $\pi(H, n)$) and the expectation in the denominator is over θ having that law also. Dividing through gives

$$\mathbb{E}[\sqrt{n}(\hat{\theta}^n - \theta)]^2 \geq \frac{1}{\frac{1}{n}\mathbb{E}\lambda^T J_n(\theta)\lambda + J(\pi)/H^2}.$$

Taking first $n \to \infty$, then $H \to \infty$, we obtain the result, applying also Lemma 2.40 to get convergence of the expectation term in the denominator to the Fisher information. □

The previous result requires the following convergence property of the Fisher information matrices. We also use it in the proof of Proposition 2.10.1.

Lemma 2.40. *Let $\theta_n \to \theta$. Denote $J_n = J_n^{(M)}$ the Fisher information for n observations*

in the model with M bins given by

$$J_n(\theta) = \mathbb{E}_\theta[-\nabla^2_\theta \log L_n(\theta)],$$

and J the Fisher information for the model. Then

$$\frac{1}{n} J_n(\theta_n) \to J(\theta).$$

When $\theta_n = \theta$ the result is simply the definition of J. The interest is hence in establishing local uniform convergence.

Proof. Theorem 3 of [32] establishes the result in the case of observed information[4],

$$-\frac{1}{n}\nabla^2_\theta l_n(\theta_n) \to J(\theta),$$

a.s. under $\mathbb{P}_\theta$. Moreover, the Fisher and Louis identities [73] show that boundedness of the complete observed information and complete scores (which take simple forms) implies boundedness of the observed information. Write $\mathcal{Y} = [M]^{\mathbb{N}}$ and μ for the counting measure on this space with associated densities $p_\theta(y)$. We have since the observed information is bounded and $p_{\theta_n}(y) \to p_\theta(y)$ pointwise and so in $L^1(\mu)$ by Scheffé's lemma [93],

$$\begin{aligned}
\frac{1}{n} J_n(\theta_n) &= \int_{\mathcal{Y}} -\frac{1}{n}\nabla^2_\theta l_n(\theta_n) p_{\theta_n}(y) d\mu(y) \\
&= \int_{\mathcal{Y}} -\frac{1}{n}\nabla^2_\theta l_n(\theta_n) p_{\theta}(y) d\mu(y) + o(1) \\
&= J(\theta^*) + o(1).
\end{aligned}$$

□

2.10.3 Technical results for the deconvolution argument

Here, we collect the technical results required to prove Lemma 2.18.

Let us recall some notation for what follows. For $r \in \{1, \dots, R\}$, and $g \in L^2(f^*_r)$, define

$$H_{r,M}(g) = g(y_0)\mathbb{P}_*(X_0 = r|\mathcal{G}^0_m) + \sum_{j=-\infty}^{-1} g(y_j)(\mathbb{P}_*(X_j = r|\mathcal{G}^0_M) - \mathbb{P}_*(X_j = r|\mathcal{G}^{-1}_M)),$$

[4]The result as stated in the reference conditions on the value $X_0 = x_0$ - the general statement follows and can be found explicitly in Theorem 13.24 in [33].

and define $H_r(g)$ similarly but with the conditioning being on the sigma algebras $\mathcal{G}^0$ and $\mathcal{G}^{-1}$. Then define, for $g = (g_1, \ldots, g_r)$,

$$H_M(g) = \sum_r H_{r,M}(g_r); \quad H(g) = \sum_r H_r(g_r).$$

These are the score functions in the sub-models with fixed transition matrix, and emission densities varying along the path characterised by g.

Our first technical result allows us to eliminate certain terms when we make the main deconvolution argument. We recall $\mathcal{G}^\ell = \sigma(Y_{-\infty:\ell})$.

Lemma 2.41. *For $g \in \mathcal{H}_r$ and $j < k \leq 0$, we have*

$$\mathbb{E}_*[g(Y_j)\mathbb{P}_*(X_j = r|\mathcal{G}^0)|Y_{k:0}] = 0.$$

Proof. We first note that

$$\mathbb{E}_*[g(Y_j)\mathbb{P}(X_j = r|\mathcal{G}^0)|Y_{k:0}] = \mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|Y_{k:0}] = \mathbb{E}_*[\mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|Y_{k:0}, X_j]|Y_{k:0}],$$

by using usual properties of conditional expectations. Given X_j, Y_j and X_j are conditionally independent of $Y_{k:0}$ and thus $\mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|Y_0, X_j] = \mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|X_j]$. Since X_j has values in the discrete set $[R]$, we can explicitly write the conditional expectation as

$$\mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|X_j] = \sum_{r'=1}^{R} \mathbb{E}_*[g(Y_j)\mathbb{1}_{X_j=r}|X_j = r']\mathbb{1}_{X_j=r'} = \mathbb{E}_*[g(Y_j)|X_j = r]\mathbb{1}_{X_j=r}$$

But, since $g \in L^2(f_r^* dx)$ has zero expectation against the emission distribution for state r, the right hand side vanishes and hence so does its expectation given $Y_{k:0}$. □

In the following section, we relate the score functions H_M in the binned model to score functions H in the full model, where the perturbation g is in the 'direction' of a histogram.

2.10.3.1 On the difference $H_M - H$

Lemma 2.42. *If $h_M = (h_{r,M})_{r \leq R}$ with $\sup_M \sum_r \|h_{r,M}\|_{L^2(f_r^*)} \leq L < \infty$, then*

$$H_M(h_M) = H(h_M) + o_{L^2}(1).$$

Proof. We have, for each r,

$$\begin{aligned} H_{r,M}(h_{r,M}) &= h_{r,M}(y_0)\mathbb{P}_*(X_0 = r|\mathcal{G}_M^0) \\ &\qquad + \sum_{j=-\infty}^{-1} h_{r,M}(y_j)(\mathbb{P}_*(X_j = r|\mathcal{G}_M^0) - \mathbb{P}_*(X_j = r|\mathcal{G}_M^{-1})); \\ H_r(h_{r,M}) &= h_{r,M}(y_0)\mathbb{P}_*(X_0 = r|\mathcal{G}^0) \\ &\qquad + \sum_{j=-\infty}^{-1} h_{r,M}(y_j)(\mathbb{P}_*(X_j = r|\mathcal{G}^0) - \mathbb{P}_*(X_j = r|\mathcal{G}^{-1})). \end{aligned} \tag{2.37}$$

We will show that $H_{r,M}(h_{r,M}) - H_r(h_{r,M}) = o_{L^2}(1)$. Let $\epsilon > 0$ be arbitrary, and $J = J(L,\epsilon) > 0$ be such that $2L\rho^{J-1} < \frac{\epsilon}{3}(1-\rho)$. Then Lemma 2.49 implies that

$$\begin{aligned} \Bigg\| \sum_{j=-\infty}^{-J} & h_{r,M}(y_j)(\mathbb{P}_*(X_j = r|\mathcal{G}^0) - \mathbb{P}_*(X_j = r|\mathcal{G}^{-1})) \\ & + \sum_{j=-\infty}^{-J} h_{r,M}(y_j)(\mathbb{P}_*(X_j = r|\mathcal{G}_M^0) - \mathbb{P}_*(X_j = r|\mathcal{G}_M^{-1}))\Bigg\|_{L^2(\mathbb{P}_*)} \le \frac{2\epsilon}{3}. \end{aligned}$$

We have now reduced the problem to the case where the sums in (2.37) are over $j > -J$. Since $\|h_M\|_{L^2} = O(1)$, to show convergence of the remaining difference between $H_{r,M}$ and H_r, it suffices to show that $\mathbb{P}_*(X_j = r|\mathcal{G}_M^\ell) \to \mathbb{P}_*(X_j = r|\mathcal{G}^\ell)$ in $L^\infty(\mathbb{P}_*)$ for $\ell = 0, 1$ and $-J \le j \le 0$.

We use Lemma 2.48. Let $K > J$. We then obtain, for all $j \ge -J$ and $\ell = 0, -1$,

$$\begin{aligned} \mathbb{P}_*(X_j = r|\mathcal{G}^\ell) &= \sum_{s=1}^R \mathbb{P}_*(X_j = r|Y_{-K:\ell}, X_K = s)\mathbb{P}_*(X_{-K} = s|\mathcal{G}_\infty^\ell) \\ &= \sum_{s=1}^R \mathbb{P}_*(X_j = r|Y_{-K:0})(1 + O(\rho^{j+K}))\mathbb{P}_*(X_{-K} = s|\mathcal{G}_\infty^\ell) \\ &= \big(1 + O(\rho^{j+K})\big) \sum_{s=1}^R \mathbb{P}_*(X_j = r|Y_{-K:0})\mathbb{P}_*(X_{-K} = s|\mathcal{G}_\infty^\ell) \\ &= \big(1 + O(\rho^{j+K})\big)\, \mathbb{P}_*(X_j = r|Y_{-K:0}). \end{aligned}$$

Similarly, $\mathbb{P}_*(X_j = r|\mathcal{G}_M^\ell) = (1 + O(\rho^{j+K}))\mathbb{P}_*(X_j = r|Y_{-K:0}^{(M)})$. Then

$$\begin{aligned} \sup_{Y_{-\infty:0}} |\mathbb{P}_*(X_j = r|\mathcal{G}_M^\ell) - \mathbb{P}_*(X_j = r|\mathcal{G}^\ell)| &\le \sup_{Y_{-\infty:0}} |\mathbb{P}_*(X_j = r|\mathcal{G}_M^\ell) - \mathbb{P}_*(X_j = r|Y_{-K:\ell}^{(M)})| \\ + \sup_{Y_{-\infty:0}} |\mathbb{P}_*(X_j = r|Y_{-K:\ell}^{(M)}) - \mathbb{P}_*(X_j = r|Y_{-K:\ell})| &+ \sup_{Y_{-\infty:0}} |\mathbb{P}_*(X_j = r|Y_{-K:\ell}) - \mathbb{P}_*(X_j = r|\mathcal{G}^\ell)|. \end{aligned}$$

By choosing K sufficiently larger than J, we have seen that the first and third terms are

small, and bounded above by $\frac{\epsilon}{9JL}$ uniformly in $j \geq -J$. It remains to bound the second term. We have

$$\begin{aligned}
&|\mathbb{P}_*(X_j = r|Y^{(M)}_{-K:\ell}) - \mathbb{P}_*(X_j = r|Y_{-K:\ell})| \\
&= \mathbb{P}_*(X_j = r)\left|\frac{\mathbb{P}_*(Y^{(M)}_{-K:\ell}|X_j = r)}{\mathbb{P}_*(Y^{(M)}_{-K:\ell})} - \frac{\mathbb{P}_*(Y_{-K:\ell}|X_j = r)}{\mathbb{P}_*(Y_{-K:\ell})}\right|.
\end{aligned}$$

We write

$$\mathbb{P}_*(Y^{(M)}_{-K:\ell}|X_j = r) = \sum_{x_{-K},\ldots,x_\ell} \mathbb{1}_{x_j=r} \prod_{t=-K}^{\ell} f_{\omega^*_{x_t}}(Y_t)\mathbb{P}_*(X_{-K:\ell} = x_{-K:\ell}|X_j = r),$$

and similarly for $Y_{-K:\ell}$. The true emissions f^*_r are continuous functions on the compact set $[0,1]$, hence they are uniformly continuous and hence are approximated in $L^\infty([0,1])$ by the histograms $f_{\omega^*_r}$ with $\omega^*_r = (\int_{I_m} f^*_r(y)dy : m \in [\kappa_M])$. Since there are a fixed number of such terms, we can choose M large enough that, for each j, the preceding display is bounded above by $\frac{\epsilon}{9JL}$. Putting everything together and using Hölder's inequality, we get that

$$\|H_{r,M}(h_{r,M}) - H_r(h_{r,M})\|_{L^2(\mathbb{P}_*)} \leq \epsilon.$$

The general case follows by summing over r, which concludes the proof. □

2.10.3.2 Deconvolution argument

The next lemma essentially provides a reduction of the deconvolution argument.

Lemma 2.43. *Let h_M be a sequence of step functions on the partition $\mathcal{I}_M$ with $\|h_M\| = \sum_r \|h_{r,M}\|_{L^2} \leq 1$, such that $h_{r,M} \in \mathcal{H}_{r,M}$. For any $k \geq 0$,*

$$H_M(h_M) \overset{L^2(\mathbb{P}_*)}{\to} 0 \implies \sum_{r=1}^{R}\sum_{k'=-k}^{0} h_M(y_{k'})\mathbb{P}_*(X_{k'} = r|Y_{-k:0}) \overset{L^2(\mathbb{P}_*)}{\to} 0.$$

Proof. For notational simplicity, we will write, for any index j and σ-algebra $\mathcal{G}$,

$$h_M(Y_j)\mathbb{P}_*(X_0|\mathcal{G}) = \sum_r h_{r,M}(Y_j)\mathbb{P}_*(X_0 = r|\mathcal{G}).$$

By Lemma 2.42, we have $H_M(h_M) \overset{L^2(\mathbb{P}_*)}{\to} 0 \iff H(h_M) \overset{L^2(\mathbb{P}_*)}{\to} 0$. Denote by Y^k the vector $(Y_{-\infty:k})$. Note that $H(h_M)(Y_{-\infty:0}) = G_M(Y^0)$ has the same distribution as $G_M(Y^j)$

$\forall j \leq 0$, by stationarity of the latent chain. Define

$$D_0^{(M)}(h^{(M)}, J) = \sum_{j=J}^{0} G_M(Y^j) = \sum_{j=J}^{0} h_M(Y_j)\mathbb{P}_*(X_j|\mathcal{G}^0) + \sum_{j=-\infty}^{J-1} h_M(Y_j)(\mathbb{P}_*(X_j|\mathcal{G}^0) - \mathbb{P}_*(X_j|\mathcal{G}^{J-1})) = o_{L^2}(1).$$

Using Lemma 2.41 with $k = 0$ to eliminate the terms which condition on $\mathcal{G}^0$, we obtain

$$\mathbb{E}_*(D_0^{(M)}(h^{(M)}, J)|Y_0) = h_M(Y_0)\mathbb{P}_*(X_0|Y_0) - \sum_{j=-\infty}^{J-1} E(h_M(Y_j)\mathbb{P}_*(X_j|\mathcal{G}^{J-1})|Y_0) = o_{L^2}(1).$$

Using Lemma 2.49, we obtain, for a particular $\rho < 1$ which does not depend on M, that

$$\sum_{j=-\infty}^{J-1} \mathbb{E}_*(h_M(y_j)\mathbb{P}_*(X_j|\mathcal{G}^{J-1})|Y_0) = \underbrace{\sum_{j=-\infty}^{J-1} \mathbb{E}_*(h_M(y_j)\mathbb{P}_*(X_j|\mathcal{G}^{J-1}))}_{=0} + O(\rho^J) = O(\rho^J),$$

with the right-hand sum vanishing because the $\mathbb{E}_*(h_M(y_j)\mathbb{P}_*(X_j|\mathcal{G}^{J-1})) = 0$, by Lemma 2.41. The term $O(\rho^J)$ is uniform in M since $\|h_M\|_{L^2(\mathbb{P}_*)} \leq 1$, and by choosing J large enough this term can be made arbitrarily small. We conclude that

$$h_M(y_0)\mathbb{P}_*(X_0|Y_0) = o_{L^2}(1),$$

as $M \to \infty$. By instead applying Lemma 2.41 with $k < 0$, we can argue similarly that, for all $k \geq 0$,

$$D_k^{(M)}(h_M) := \sum_{r=1}^{R} \sum_{k'=-k}^{0} h_M(y_{k'})\mathbb{P}_*(X_{k'} = r|Y_{-k:0}) = o_{L^2}(1).$$

□

2.10.3.3 On the convergence on $\tilde{\mathcal{A}}_M$

The proof of the following result is very similar to the proof of Lemma 1 in [45], which used that the spaces of score functions for mixture models were nested to show convergence.

Lemma 2.44. *Let $\tilde{\mathcal{A}}_M$ $\tilde{\mathcal{P}}_M$, $\mathcal{A}$, $\mathcal{P}$ be as in the statement of Lemma 2.17. Then $\mathcal{A} = \tilde{\mathcal{A}} :=$*

$\lim_{M\to\infty} \tilde{\mathcal{A}}_M$.

Proof. We have by the arguments of [45] that for any $S \in L^2(\mathbb{P}_*)$ $(\tilde{\mathcal{A}}_M S)_M$ is Cauchy, as the spaces $\tilde{\mathcal{P}}_M$ are nested. By completeness of $L^2(\mathbb{P}_*)$, we then establish convergence to some element $\tilde{\mathcal{A}}S \in L^2(\mathbb{P}_*)$. Since $\tilde{\mathcal{A}}_M$ are projections, $\tilde{\mathcal{A}}$ is also a projection onto its image, which is some subspace

$$\tilde{\mathcal{P}} \subset \mathrm{Cl}\left(\bigcup_{M\in\mathbb{N}} \tilde{\mathcal{P}}_M\right).$$

Recall that $\mathcal{P}$ is defined as the closure of the linear span of the $H(h)$ functions, as in the statement of Lemma 2.17. By definition, $\mathcal{P}$ is closed and, since $\tilde{\mathcal{P}}_M \subset \mathcal{P}$ for all M, we must have $\mathrm{Cl}\left(\bigcup_{M\in\mathbb{N}} \tilde{\mathcal{P}}_M\right) \subset \mathcal{P}$. We show that any element in $H(h) \in \mathcal{P}$ can be written as the L^2 limit of some sequence in $\bigcup_{M\in\mathbb{N}} \tilde{\mathcal{P}}_M$. First, choose a sequence $h_M \to h$ in $L^2(\mathbb{P}_*)$ such that $\|h_M\|_{L^2} \leq 2\|h\|_{L^2}$ for all M, which is possible under **Assumption** 2.5. We now wish to show that, for all $r \in [R]$,

$$\begin{aligned}&(h_{r,M}(y_0) - h_r(y_0))\mathbb{P}_*(X_0 = r|\mathcal{G}^0)\\ &+ \sum_{j=-\infty}^{-1} (h_r(y_j) - h_{r,M}(y_j))(\mathbb{P}_*(X_j = r|\mathcal{G}^0) - \mathbb{P}(X_j = r|\mathcal{G}^{-1}))\end{aligned} \tag{2.38}$$

vanishes in L^2. By Lemma 2.49, we can choose J sufficiently large negative that

$$\sum_{j=-\infty}^{J} (h_r(y_j) - h_{r,M}(y_j))(\mathbb{P}_*(X_j = r|\mathcal{G}^0) - \mathbb{P}_*(X_j = r|\mathcal{G}^{-1})) < \frac{\epsilon}{2}.$$

Since $\|h_{r,M}\|_{L^2} \leq 2\|h_r\|_{L^2}$, J can be chosen so that the above holds uniformly in M.

To control the finitely many remaining terms, it suffices to note the L^∞ boundedness of the probabilities and the L^2 convergence of h_M to h, and so, for M sufficiently large, the remaining finite sum is bounded by $\frac{\epsilon}{2}$, hence (2.38) has L^2 norm at most ϵ. Since ϵ was arbitrary, we conclude that $\mathcal{P} \subset \tilde{\mathcal{P}}$, and so the two spaces coincide and $\mathcal{A} = \tilde{\mathcal{A}}$ as claimed. □

2.10.4 Admissible partitions

The following lemma is well-known in the identifiability literature. We recall it here for completeness.

Lemma 2.45. *Let $F_1^*, \ldots, F_R^*$ be linearly independent distribution functions on $[0, 1]$.*

Then there exists a partition $I_1, \ldots, I_R$ *of* $[0,1]$ *which is admissible in the sense of Definition 2.1.*

2.10.5 Technical results for the proofs of Theorems 2.10 and 2.13

We employ the following lemma (Lemma 2.23 of the main text) in the proof of Theorem 2.10 to eliminate certain terms.

Lemma 2.46. *Consider a sequence of emissions* $\mathbf{f}^{(n)}$ *such that* $g^{(3)}_{Q^*,\mathbf{f}^{(n)}} \to g^{(3)}_{Q^*,\mathbf{f}^*}$ *in* $L^1(dx^3)$. *Then* $\mathbf{f}^{(n)} \xrightarrow{L^1} \mathbf{f}^*$ *along a subsequence.*

Proof. Write $P = \mathrm{diag}(p^*)$. Since $\|g^{(3)}_{Q^*,\mathbf{f}^{(n)}} - g^{(3)}_{Q^*,\mathbf{f}^*}\|_{L^1} = o(1)$, $\|g^{(2)}_{Q^*,\mathbf{f}^{(n)}} - g^{(2)}_{Q^*,\mathbf{f}^*}\|_{L^1} = o(1)$ and

$$\sum_{r,s}(PQ^*)_{rs}(f_r^{(n)}(y_1)f_s^n(y_2) - f_r^*(y_1)f_s^*(y_2)) = o_{L^1}(1).$$

Construct a partition $I_1, \ldots I_R$ of $\mathbb{R}$ such that $F^*_{ir} = \int_{I_i} f_r^* dy$ is of rank R and define $F^{(n)}_{ir} = \int_{I_i} f_r^{(n)}(y)dy$. Then we get, for all i,

$$\sum_{r,s}(PQ^*)_{rs}(F_{ri}^{(n)}f_s^{(n)}(y) - F^*_{ri}f_s^*(y)) = o_{L^1}(1).$$

Since $g^{(3)}_{Q^*,\mathbf{f}^{(n)}} \to g^{(3)}_{Q^*,\mathbf{f}^*}$, Theorem 2.3 of [103] establishes weak convergence of the emissions, which then implies $\mathbf{F}^n \to \mathbf{F}^*$ and so

$$(\mathbf{F}^*)^T(PQ^*)\mathbf{f}^{(n)} = (\mathbf{F}^*)^T(PQ^*)\mathbf{f}^* + o_{L^1}(1),$$

which implies $\mathbf{f}^{(n)} = \mathbf{f}^* + o_{L^1}(1)$. □

In the proof of Theorem 2.13, we use the fact that the approximating distribution for the posterior $\Pi_1(\cdot|Y_{1:n})$ can be chosen to not depend on the observation Y_ℓ. The difficulty in the proof compared to the iid setting arises from the dependence structure, but, by the exponential forgetting (Lemma 2.48), we can prove that the $1/\sqrt{n}$ scaled joint score functions coincide up to $o(1)$ terms, and hence so do the $\sqrt{n}-$scaled MLEs.

Lemma 2.47. *Grant* ***Assumptions*** *2.1-2.5, let* $\hat{\theta} = (\hat{Q}, \hat{\omega})$ *denote the MLE in the model with admissible partition* $\mathcal{I}_M$ *and define* $\hat{\theta}^{-\ell} = (\hat{Q}^{-\ell}, \hat{\omega}^{-\ell})$ *by*

$$\hat{\theta^{-\ell}} = \theta^* + \frac{J_M^{-1}}{n}\nabla_\theta l^{(-\ell)}_{\theta^*}(Y^{-\ell}_{1:n}), \tag{2.39}$$

where $Y_{1:n}^{-\ell} = (Y_1, \ldots, Y_{\ell-1}, Y_{\ell+1}, \ldots, Y_n)$ *and*

$$l_\theta^{-\ell}(Y_{1:n}^{(-\ell)}) = \sum_{i=1}^{\ell-1} l_\theta(Y_i|Y_{1:i-1}) + \sum_{i=\ell+1}^{n} l_\theta(Y_i|Y_{1:i-1}^{(-\ell)}). \tag{2.40}$$

Let $\mathcal{L}_n \subset [n]$ *be an index set whose cardinality satisfies* $|\mathcal{L}_n| \log|\mathcal{L}_n| = o(n)$*. Then, for all* $\epsilon > 0$*,*

$$\mathbb{P}_*(\exists l \in \mathcal{L}_n : \sqrt{n}|\hat{Q}^{(-l)} - Q^*| > \epsilon) \to 0,$$

as $n \to \infty$

Proof. We may write the log-likelihood for the parameter $\theta = (Q, \omega)$ as

$$l_\theta(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} l_\theta(Y_i|Y_{1:i-1}), \tag{2.41}$$

where the $i = 1$ term is the unconditional likelihood. Note that (2.40) corresponds to the log-likelihood based on observing $Y_{1:n}^{(-\ell)}$. Standard arguments for showing asymptotic normality (see e.g. [12]), using Taylor expansions of the derivative of the expression (2.41), give

$$\sqrt{n}(\hat{\theta} - \theta^*) = \frac{J_M^{-1}}{\sqrt{n}} \nabla_{\theta^*} l_{\theta^*}(Y_{1:n}) + R_n,$$

for $R_n = O_{\mathbb{P}_*}(\sup_{\bar{\theta} \in [\theta^*, \hat{\theta}]} \| \frac{1}{n} D_{\bar{\theta}}^2 l_{\bar{\theta}}(Y_{1:n}) + J_M \|)$, the error in approximating the Fisher Information with the negative log-likelihood. The uniform law of large numbers for the observed information as given in Theorem 3 of [32], together with the consistency of the MLE (see e.g. [68]), then implies that $R_n = o_{\mathbb{P}_*}(1)$.

For the estimator $\hat{\theta}^{(-\ell)}$, we have, for each $l \in \mathcal{L}$, by the definition in (2.39), that

$$\sqrt{n}(\hat{\theta}^{-\ell} - \theta^*) = \frac{J_M^{-1}}{\sqrt{n}} \nabla_{\theta^*} l_{\theta^*}^{-\ell}(Y_{1:n}^{-\ell}).$$

Hence

$$\sqrt{n}(\hat{\theta}^{-\ell} - \hat{\theta}) = \frac{J_M^{-1}}{\sqrt{n}} (\nabla_{\theta^*} l_{\theta^*}^{-\ell}(Y_{1:n}^{-\ell}) - \nabla_{\theta^*} l_{\theta^*}(Y_{1:n})) - R_n$$

Note that the R_n term does not depend on ℓ, and vanishes in $\mathbb{P}_*-$probability. We will show that, for each $\ell \in \mathcal{L}_n$, $\mathbb{P}_* \left[\frac{1}{\sqrt{n}} (\nabla_{\theta^*} l_{\theta^*}^{-\ell}(Y_{1:n}^{-\ell}) - \nabla_{\theta^*} l_{\theta^*}(Y_{1:n})) > \epsilon \right] = o(|\mathcal{L}_n|^{-1})$, so that a union bound implies

$$\mathbb{P}_*(\exists l \in \mathcal{L}_n : \frac{1}{\sqrt{n}} (\nabla_{\theta^*} l_{\theta^*}^{-\ell}(Y_{1:n}^{-\ell}) - \nabla_{\theta^*} l_{\theta^*}(Y_{1:n})) > \epsilon) \leq |\mathcal{L}_n| o_{\mathbb{P}_*}(|\mathcal{L}_n|^{-1}) = o_{\mathbb{P}_*}(1),$$

from which the result follows. The argument is based on the exponential forgetting properties, and the expansion of the score considered in [32]. For the full log-likelihood (2.41), we have through the Fisher identity as in [32] that the log-likelihood of θ given $Y_{1:n}$ and $X_0 = x$, denoted $l_n(\theta, X_0 = x)$, has gradient at $\theta = \theta^*$ given by

$$\frac{1}{\sqrt{n}}\nabla_{\theta^*} l_n(\theta, X_0 = x) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} \Delta_{k,0,x},$$

with

$$\begin{aligned}\Delta_{k,0,x} = \mathbb{E}_{\theta^*}\left[\sum_{i=1}^{k} \nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{0:k}, X_0 = x\right] \\ - \mathbb{E}_{\theta^*}\left[\sum_{i=1}^{k-1} \nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{0:k-1}, X_0 = x\right],\end{aligned}$$

which can be seen by writing the log-likelihood as a telescoping sum and using the Fisher identity to write each term as a conditional expectation of the full likelihood. We can do the same for the likelihood without Y^ℓ: define

$$\begin{aligned}\Delta^{(-\ell)}_{k,0,x} = \mathbb{E}_{\theta^*}\left[\sum_{i=1}^{k} \nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y^{-\ell}_{0:k}, X_0 = x\right] \\ - \mathbb{E}_{\theta^*}\left[\sum_{i=1}^{k-1} \nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y^{(-\ell)}_{0:k-1}, X_0 = x\right].\end{aligned}$$

We remark that the term $\Delta^{(-\ell)}_{l,0,x} = 0$ which corresponds to the missing contribution from Y^l in the expansion of the likelihood. Then we can write

$$\frac{1}{\sqrt{n}}\nabla_{\theta^*} l^{(-\ell)}_n(\theta, X_0 = x) = \frac{1}{\sqrt{n}}\sum_{k<\ell-1} \Delta^{(-\ell)}_{k,0,x} + \frac{1}{\sqrt{n}}\sum_{k=\ell+1}^{\ell+1+L_n} \Delta^{(-\ell)}_{k,0,x} + \frac{1}{\sqrt{n}}\sum_{k>L_n} \Delta^{(-\ell)}_{k,0,x}.$$

The first sum remains unchanged compared to the expression without the missing data point. Lemma 8 of [32] shows that we may replace the $\Delta_{k,0,x}$ by $\Delta_{k,0}$ where the latter is defined conditioning only on the $Y_{0:k}$ - the same argument made there shows the analogous result for the $\Delta^{(-\ell)}_{k,0,x}$ with respect to the analogously defined $\Delta^{(-\ell)}_{k,0}$. We note that, under **Assumptions** 2.3 and 2.4, the $\Delta_{k,0}$ and $\Delta^{-\ell}_{k,0}$ are bounded uniformly in $X_{(\cdot)}, Y_{(\cdot)}$, and so, for any $L_n = o(\sqrt{n})$ to be chosen later, we have that for some constant $C > 0$ depending on Q^* and w^*,

$$\mathbb{P}^*\left[\frac{1}{\sqrt{n}}\sum_{k=\ell+1}^{\ell+1+L_n} \Delta^{(-\ell)}_{k,0}(\theta) - \frac{1}{\sqrt{n}}\sum_{k=\ell}^{\ell+1+L_n} \Delta_{k,0}(\theta) > CL_n/\sqrt{n}\right] = 0$$

It remains to control the difference between the remaining contributions for suitably chosen L_n. Similarly to what is done in [32], we define $\Delta_{k,m}$ by conditioning instead on $Y_{m:k}$. Write $\Delta_{k,m}$ as a telescoping sum:

$$\Delta_{k,m} = \mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{k-1},X_k}\omega_{X_k,Y_k})|Y_{m:k}] + \sum_{i=m+1}^{k-1} \mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m:k}] - \mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m:k-1}].$$

Then their equation (19) states that, for $k > m > m'$, [5]

$$\|\mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m:k}] - \mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m':k}]\| \leq 2C\rho_*^{i-m-1},$$

where $\rho^* \in (0,1)$ is a constant which depends on θ^* alone. Furthermore, their equation (20) states that

$$\|\mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m:k}] - \mathbb{E}_{\theta^*}[\nabla_{\theta^*} \log(Q_{X_{i-1},X_i}\omega_{X_i,Y_i})|Y_{m:k-1}]\| \leq 2C\rho_*^{k-i-1}.$$

Combining the above two displays in the same way as Lemma 10 of [32], we have, for a suitable constant $C(Q^*, \omega^*)$ that

$$\|\Delta_{k,0} - \Delta_{k,\ell+1}\|_{L^2(\mathbb{P}_*)} \leq C(Q^*, \omega^*)\frac{\rho_*^{(k-\ell-1)/2-1}}{1-\rho_*}.$$

Arguing similarly for the case of $\Delta_{k,0}^{-\ell}$, we see that

$$\|\Delta_{k,0}^{-\ell} - \Delta_{k,\ell+1}^{-\ell}\|_{L^2(\mathbb{P}_*)} \leq C(Q^*, \omega^*)\frac{\rho_*^{(k-\ell-1)/2-1}}{1-\rho_*}.$$

The $\Delta_{k,\ell+1}^{-\ell} = \Delta_{k,\ell+1}$ as the conditioning excludes Y_ℓ in either case, and so the above displays combine to give

$$\|\Delta_{k,0}^{-\ell} - \Delta_{k,0}\|_{L^2(\mathbb{P}_*)} \leq 2C(Q^*, \omega^*)\frac{\rho_*^{(k-\ell-1)/2-1}}{1-\rho_*}.$$

Applying Minkowski's inequality then gives that

$$\left\| \sum_{k>\ell+1+L_n} (\Delta_{k,0}^{-\ell} - \Delta_{k,0}) \right\|_{L^2(\mathbb{P}_*)} = O(\rho_*^{L_n/2}).$$

[5] Their result is only stated for $m < 0$, but by stationarity extends to all $m \in \mathbb{Z}$

Choosing $L_n = 2M \log|\mathcal{L}_n|$ for sufficiently large $M > 0$ we have $\rho_*^{L_n/2} = e^{(\log|\mathcal{L}_n|)(M \log \rho_*)} = |\mathcal{L}_n|^{M \log \rho_*} = o\left(\frac{1}{n}\right)$. Therefore we finally obtain that for all $l = o(n/\log n)$

$$\mathbb{P}_*\left(\frac{1}{\sqrt{n}}(\nabla_{\theta^*} l_{\theta^*}^{-\ell}(Y_{1:n}^{-\ell}) - \nabla_{\theta^*} l_{\theta^*}(Y_{1:n})) > \epsilon\right) \leq 1/n$$

and since $|\mathcal{L}_n| = o(n)$, so the claim is proved. □

2.11 Assumptions required for the application of Theorem 2.12 to Dirichlet process mixtures

In this section, we detail the assumptions required for Proposition 2.15. We will need the following assumptions on the behaviour of the emissions, which are assumptions (T1)-(T3) of [102].

Assumption 2.8. (i) There exist positive constants M_0, τ_0, γ_0 such that, for all $1 \leq i \leq R$ and all $y \in \mathbb{R}$,

$$f_i^*(y) \leq M_0 \exp(-\tau_0 |y|^{\gamma_0}).$$

(ii) For all $1 \leq i, j \leq R$ there exist constants $T_{i,j}, M_{i,j}, \tau_{i,j}, \gamma_{i,j} < \gamma_0$ such that

$$f_i^*(y) \leq f_j^*(y) M_{i,j} \exp(\tau_{i,j} |y|^{\gamma_{i,j}}), \qquad |y| \geq T_{i,j}.$$

(iii) For all $1 \leq i \leq R$, f_i^* is positive and there exists $c_i > 0$ and $y_i^{low} < y_i^{high}$ such that f_I^* is non-decreasing on $(-\infty, y_i^{low})$, bounded below by $c > 0$ on (y_i^{low}, y_i^{high}), and non-increasing on (y^{high}, ∞).

We will further require the following assumptions on the choice of prior, which are Assumptions (G1) and (S1)-(S3) of [102]. These assumptions are verified by standard choices of Gaussian base measure and inverse gamma prior on the standard deviation.

Assumption 2.9. (i) $\alpha(\mathbb{R} \backslash [-y, y]) \lesssim \exp(-C_1 y^{a_1})$ for y large enough and some $C_1, a_1 > 0$
(ii) **B.** $\sigma \sim \Pi_\sigma$ where Π_σ is an Inverse Gamma distribution.

Condition (ii) of Assumption 2.9 on Π_σ can be relaxed; we essentially Π_σ to admit a Lebesgue density π_σ which behaves like an inverse - Gamma near 0 and has tails bounded by some power of $1/\sigma$ near infinity (see for instance [62]).

2.12 Key results from literature

In this section, we document some of the key results from other works which are used in our contributions. Section 2.12.1 features the forgetfulness properties of [32], while Section 2.12.2 expands upon the presentation of efficiency given in Section 2.3.2.1, recalling the relevant framework of [78].

2.12.1 Forgetting of the hidden Markov chain

The following results are from [32]. Recall that $\bar{q} = \min_{ij} Q^*_{ij} > 0$ under Assumption 2.3. The first result quantifies the exponential forgetting of the hidden chain.

Lemma 2.48 (Douc et al. 2004, [32]). *For all $s \in [R]$, $k, k' \in \mathbb{Z}$, and $j \geq k$, we have*

$$\sup_{Y_{k:k'}} \|\mathbb{P}_*(X_j = \cdot | X_k = s, Y_{k:k'}) - \mathbb{P}_*(X_j = \cdot | Y_{k:k'})\|_{TV} \leq \rho^{j-k},$$

where $\rho = \frac{1-2\bar{q}}{1-\bar{q}}$.

The second of these results follows from applying Equation (20) of [32] to both the known-emission model and the submodels on the emissions. It in particular ensures that the score functions we consider are well-defined.

Lemma 2.49 (Douc et al. 2004, [32]). *For $h \in L^2(\mathbb{P}_*)$, $j \leq 0$, $r, s \in [R]$, and for $\mathbb{P}_*-$almost all $Y_{-\infty:0}$,*

$$|\mathbb{P}_*(X_{j-1} = s, X_j = r | Y_{-\infty:0}) - \mathbb{P}_*(X_{j-1} = s, X_j = r | Y_{-\infty:-1})| \leq \rho^{-j-1}$$

and

$$|h(Y_j)\{\mathbb{P}_*(X_j = r | Y_{-\infty:0}) - \mathbb{P}_*(X_j = r | Y_{-\infty:-1})\}| \leq |h(Y_j)|\rho^{-j-1},$$

where $\rho = \frac{1-2\bar{q}}{1-\bar{q}}$. In particular, by integrating the square of both sides, we have

$$\|h(Y_j)\{\mathbb{P}_*(X_j = r | Y_{-\infty:0}) - \mathbb{P}_*(X_j = r | Y_{-\infty:-1})\}\|_{L^2(\mathbb{P}_*)} \leq \|h\|_{L^2(\mathbb{P}_*)}\rho^{-j-1}.$$

Lemmas 2.48 and 2.49 remain true when replacing Y by $Y^{(M)}$, the coarsened data. The choice of ρ is made independently of M, as it only depends on the transition matrix Q^*, which is the same in all histogram models as it is in the semiparametric model.

2.12.2 Convolution Theorem and Efficiency

The following section pertains to Section 2.3.2.1, in which our model is considered in the framework of [78] to understand the relevant notions of LAN expansion and efficiency. The following sequence of definitions, and the convolution theorem which follows, are as in Section 2 of [78]. In this section, we also give further background to the exposition in Section 2.3.2.1. We do not present the material in its full generality, but rather to the extent required for our application, to which we apply the general framework.

Let Θ denote the parameter space and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures governing the law of $(Y_1, Y_2, \cdots)$ and write $P_{\theta,n}$ for the marginal of $Y_{1:n} = (Y_1, \ldots, Y_n)$. For $\theta_1, \theta_2 \in \Theta$, write

$$\Lambda_n(\theta_1, \theta_2) = \log \frac{dP_{n,\theta_1}}{dP_{n,\theta_2}} = \ell_n(\theta_1) - \ell_n(\theta_2)$$

for the log-likelihood ratio between θ_1 and θ_2, based on observations $Y_{1:n}$.

The first step is to exhibit a LAN expansion, formally defined below, for the semiparametric model we consider. The parameter space is $\Theta = \mathcal{Q} \times \mathcal{F}^R$ and perturbations from the true parameter $\theta = (Q^*, \mathbf{f}^*)$ can be written in the form $(Q, \mathbf{f}) = (Q, f_1, \ldots, f_R) = (Q^* + a^T Q, f_1^*(1+h_1), \ldots f_R^*(1+h_R))$ for appropriate $a \in \mathbb{R}^{R\times(R-1)}$ and $h_r = \frac{f_r}{f_r^*} - 1$, which has finite $L^\infty(f_r^*)$ norm by positivity of f_r^* and satisfies $\int h_r(y) f_r^*(y) d(y) = 0$. This motivates the choice of tangent space at the parameter $(Q^*, \mathbf{f}^*)$ as

$$\mathcal{H} = \{(a, \mathbf{h}), a \in \mathbb{R}^{R\times(R-1)}, \quad \mathbf{h} = (h_r)_{r=1}^R, \quad h_r \in \mathcal{H}_r\}$$

where $\mathcal{H}_r = \mathcal{H}_r(f_r^*) = \{h_r \in L^2(f_r^*) : \int h_r f_r^* = 1, \|h_r\|_\infty < \infty\}$.

The following definition is a specialisation to our context of the corresponding definition found in [78]. We write

$$\theta_n((a, \mathbf{h})) = \left(Q^* + \frac{a}{\sqrt{n}}, f_1^*(1 + \frac{h_1}{\sqrt{n}}), \ldots, f_R^*(1 + \frac{h_R}{\sqrt{n}})\right).$$

Definition 2.50 (Local Asymptotic Normality)**.** Let $\mathcal{H}$ be as above, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\|\cdot\|_{\mathcal{H}}$. We say the model is locally asymptotically normal (LAN) at $\theta_0 \in \Theta$, indexed by the tangent space $\mathcal{H}$, if for each $(a, \mathbf{h}) \in \mathcal{H}$, there exists a sequence

$\{P_{n,\theta_n(a,\mathbf{h})}\}$ of probability measures with

$$\Lambda_n(\theta_n(a,\mathbf{h}),\theta_0) = \Delta_{n,a,\mathbf{h}} - \frac{1}{2}\|(a,\mathbf{h})\|^2_{\mathcal{H}} + o_{\mathbb{P}_*}(1),$$

with $\Delta_{n,a,\mathbf{h}}$ real-valued random variables, which are linear functions of $(a,\mathbf{h})$, such that $\Delta_{n,a,\mathbf{h}} \to N(0,\|(a,\mathbf{h})\|^2_{\mathcal{H}})$ for all $(a,\mathbf{h}) \in \mathcal{H}$.

In essence, the $\Delta_{n,a,\mathbf{h}}$ are score functions in the 'direction' h, since Λ_n is the local change in the log-likelihood when we perturb along the path characterised by h and $\Delta_{n,a,\mathbf{h}}$ is the first order term.

To show the model is LAN with tangent space $\mathcal{H}$, we seek to identify the score functions $\Delta_{n,a,\mathbf{h}}$ note that the log-likelihood ratio $\Lambda(\theta_n(a,\mathbf{h}),\theta_0)$ can be approximated by considering the score in the submodel

$$\Theta_{a,\mathbf{h}} = \{\theta_t = (Q^* + ta, (f^*_r(1+th_r) : r \in [R])),\ |t| \le \epsilon\},$$

which, for fixed $a,\mathbf{h}$ and fixed $\epsilon > 0$ sufficiently small, is a one-dimensional submodel $\Theta_{a,\mathbf{h}} \subset \Theta$. From [32], in which the authors consider score functions in parametric HMMs, we can expand the gradient of the log-likelihood $\ell_n^{(a,\mathbf{h})}(t)$ in the (parametric) model $\Theta_{a,\mathbf{h}}$, at $t_0 = 0$ as

$$\frac{1}{\sqrt{n}}\nabla_t \ell_n^{(a,\mathbf{h})}(t_0) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\Delta_{k,\infty}^{(a,\mathbf{h})} + o_{L^2(\mathbb{P}_*)}(1),$$

where

$$\begin{aligned}\Delta_{k,\infty}^{(a,\mathbf{h})} = \Delta_{k,\infty}^{(a,\mathbf{h})}(Y_{-\infty:k}) &= \mathbb{E}_*[\phi(X_{k-1},X_k,Y_k)|Y_{-\infty}^k] \\ &+ \sum_{i=-\infty}^{k-1}\left(\mathbb{E}_*[\phi(X_{i-1},X_i,Y_i)|Y_{-\infty}^k] - \mathbb{E}_*[\phi(X_{i-1},X_i,Y_i)|Y_{-\infty}^{k-1}]\right),\end{aligned} \tag{2.42}$$

and where, writing $a_{rR} = -\sum_s a_{rs}$ and $Q_{r,R} = 1 - \sum_s Q_{rs}$,

$$\phi(X_{k-1},X_k,Y_k) = \sum_{r=1}^{R}\sum_{s=1}^{R}\mathbb{1}\{X_{k-1}=r, X_k=s\}\frac{a_{r,s}}{Q^*_{r,s}} + \sum_{s=1}^{R}\mathbb{1}_{X_i=s}h_s(Y_i). \tag{2.43}$$

These formulae arise as an application of the results of Section 6.1 of [32] to $\Theta_{a,\mathbf{h}}$.

We note that the contribution of the first term of the right hand side of (2.43), which may

be rewritten as

$$\sum_{r=1}^{R}\sum_{s=1}^{R-1}\left(\frac{\mathbb{1}\{X_{k-1}=r,X_k=s\}}{Q^*_{rs}}-\frac{\mathbb{1}\{X_{k-1}=r,X_k=R\}}{Q^*_{rR}}\right)a_{rs},$$

to the expression defined in (2.42), is precisely the score function for estimation in the model $\Theta_{a,\mathbf{0}}$ in which the emission densities are fixed and known. We see that this is equal to $a^T S_{Q^*}(Y_{-\infty:k})$, for S_{Q^*} the score at Q^* in the $R\times(R-1)$-dimensional parametric model with known emissions and unknown Q, which takes the form

$$\begin{aligned}S_{Q^*}(r,s)(Y_{-\infty:k}) &= \frac{\mathbb{P}_*(X_{k-1}=r,X_k=s|Y_{-\infty:k})}{Q^*_{rs}}-\frac{\mathbb{P}_*(X_{k-1}=r,X_k=R|Y_{-\infty:k})}{Q^*_{rR}}\\ &+\sum_{i=-\infty}^{k-1}\left\{\frac{\mathbb{P}_*(X_{i-1}=r,X_i=s|Y_{-\infty:k})-\mathbb{P}_*(X_{i-1}=r,X_i=s|Y_{-\infty:k-1})}{Q^*_{rs}}\right.\\ &\left.-\frac{\mathbb{P}_*(X_{i-1}=r,X_i=R|Y_{-\infty:k})-\mathbb{P}_*(X_{i-1}=r,X_i=R|Y_{-\infty:k-1})}{Q^*_{rR}}\right\},\end{aligned} \tag{2.44}$$

where we index with $(r,s)\in[R]\times[R-1]$ for convenience sake, but consider S_{Q^*} as a vector of length $R(R-1)$. We then rewrite (2.42), substituting also the expression in (2.43) as

$$\begin{aligned}\Delta^{(a,\mathbf{h})}_{k,\infty} = a^T S_{Q^*} &+\sum_{r=1}^{R}\left\{\mathbb{P}_*(X_k=r|Y^k_{-\infty})h_r(Y_i)\right.\\ &\left.+\sum_{i=-\infty}^{k-1}(\mathbb{P}_*(X_i=r|Y^k_{-\infty})-\mathbb{P}_*(X_i=r|Y^{k-1}_{-\infty}))h_r(Y_i)\right\}.\end{aligned} \tag{2.45}$$

Henceforth we will write $H_r(h_r)(Y_{-\infty:k})$ for the function of Y defined inside the curly brackets. In our notation, we suppress the dependence on k because the law of $H_r(h_r)(Y_{-\infty:k})$ coincides with that of $H_r(h_r)(Y_{-\infty:0})$ under a suitable re-indexing, by stationarity. A Taylor expansion of the log-likelihood in $\Theta_{a,\mathbf{h}}$ then gives us that, for some $t_1\in[0,n^{-1/2}]$,

$$\begin{aligned}\ell_n(\theta_n((a,\mathbf{h})))-\ell_n(\theta_*) &= \frac{1}{\sqrt{n}}\sum_{k=1}^{\infty}\Delta^{(a,\mathbf{h})}_{k,\infty}-\frac{1}{2}\frac{1}{n}\left(-\frac{d^2}{dt^2}|_{t=t_1}\ell^{(a,\mathbf{h})}_n(t)\right)\\ &= \Delta_{n,(a,\mathbf{h})}-\frac{1}{2}\|(a,\mathbf{h})\|^2_{\mathcal{H}}+o_{\mathbb{P}_*}(1).\end{aligned}$$

where $\|(a,\mathbf{h})\|^2_{\mathcal{H}}$ the Fisher information at $t=0$ in the model $\Theta_{a,\mathbf{h}}$, which is defined in [32]

as the $L^2 = L^2(\mathbb{P}_*)$ norm[6] of $\Delta_{0,\infty}^{(a,\mathbf{h})}$, and where $\Delta_{n,(a,\mathbf{h})} = \frac{1}{\sqrt{n}} \sum_{k=1}^{\infty} \Delta_{k,\infty}^{(a,\mathbf{h})}$, which is linear in $(a, \mathbf{h})$ and satisfies $\Delta_{n,(a,\mathbf{h})} \to N(0, \|(a,\mathbf{h})\|_{\mathcal{H}}^2)$, by the discussion directly preceding Theorem 2 in [32]. We also used above the local uniform convergence of the second derivative of the score to the Fisher information matrix at $t = 0$ in $\Theta_{a,\mathbf{h}}$, which is guaranteed by Theorem 3 of [32].

The preceding discussion shows that our model is LAN, which is the first step in understanding efficient estimators. The following definition is adapted from [78] for our specific case.

Definition 2.51 (Differentiability of parameter)**.** Let $v_n(P_{n,\theta})$ be $\mathbb{R}^p-$valued parameters for some $p > 0$. We say the sequence (v_n) is *differentiable* if

$$\sqrt{n}(v_n(P_{n,\theta_n(a,\mathbf{h})}) - v_n(P_{n,0})) \to \dot{v}(a,\mathbf{h}) \quad \forall h \in \mathcal{H}$$

for some continuous linear map $\dot{v} : \mathcal{H} \longrightarrow \mathbb{R}^p$.

In our case, the parameter of interest will be $v_n(P_{n,\theta_n(a,\mathbf{h})}) = Q$, so that $p = R(R-1)$. We then obtain $\dot{v}(a,\mathbf{h}) = a \in \mathbb{R}^p$.

We finally recall the definition of a regular estimator.

Definition 2.52 (Regular estimators)**.** A sequence of maps $T_n = T_n(Y_1, \ldots, Y_n) \in \mathbb{R}^p$ is said to be *locally regular* for v_n if, under $P_{n,\theta_n(a,\mathbf{h})}$, we have

$$\sqrt{n}(T_n - v_n(P_{n,\theta_n(a,\mathbf{h})}) \Rightarrow \mathbb{Z}$$

as $n \to \infty$ for every $(a,\mathbf{h}) \in \mathcal{H}$, where $\mathbb{Z}$ is a Borel measurable tight random element in $\mathbb{R}^p$ which does not depend on $(a,\mathbf{h}) \in \mathcal{H}$.

Following what precedes, we are ready to state the convolution theorem also described in [78], and originally proven in [101]. As with what precedes, we slightly simplify the statement to make clear the application to our context.

Theorem 2.53 (Convolution theorem)**.** *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is LAN at a point θ_* indexed by a linear subspace $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of a Hilbert space. Further suppose (v_n) is a differentiable sequence of parameters. Then, if T_n is locally regular for v_n, there exist tight*

[6] The L^2 norms of all $\Delta_{k,\infty}^{(a,\mathbf{h})}$ coincide by stationarity.

Borel measurable elements $\mathbb{Z}_0$ *and* $\mathbb{W}$ *in* $\mathbb{R}^p$ *with*

$$(A) \quad P(\mathbb{Z}_0 \in \overline{\dot{v}(\mathcal{H})}) = 1$$

$$(B) \quad \mathcal{L}(\mathbb{Z}) = \mathcal{L}(\mathbb{Z}_0 + \mathbb{W})$$

$$(C) \quad \mathbb{Z}_0 \text{ and } \mathbb{W} \text{ are independent}$$

$$(D) \quad \mathcal{L}(b^T\mathbb{Z}_0) = N(0, \|\dot{v}_b^T\|^2) \text{ for every } b \in \mathbb{R}^p$$

Here $\dot{v}_b^T$ *is the unique element of* $\mathcal{H}$ *such that*

$$(\dot{v}^T(b))(a, \mathbf{h}) = \langle \dot{v}_b^T, (a, \mathbf{h})\rangle_{\mathcal{H}}$$

where $\dot{v}^T : (\mathbb{R}^p)^\dagger \to \mathcal{H}^\dagger$ *is the adjoint of* $\dot{v}$*, and where the superscript* $\cdot^\dagger$ *indicates the dual space.*

The convolution theorem essentially states that the limiting law of a regular estimator is the convolution of the laws of a Gaussian (in this case $\mathbb{Z}_0$) and some independent random variable (in this case $\mathbb{W}$), and so in particular the variance is lower bounded by that of $\mathbb{Z}_0$. It also describes the variance of $\mathbb{Z}_0$ in terms of the tangent space, its dual space, and the parameter 'derivative' $\dot{v}$. In our case, we recall that $\dot{v}(a, h) = a$, and $p = R(R-1)$.

We note that we identify the dual of $\mathbb{R}^p$ with itself, through the bijection $b \leftrightarrow \langle b, \cdot\rangle_{\mathbb{R}^p}$, so $\dot{v}^T(b)$ is a shorthand for $\dot{v}^T(\langle b, \cdot\rangle_{\mathbb{R}^p})$. The adjoint $\dot{v}^T$ of $\dot{v}$ then satisfies, by definition, for $(a, \mathbf{h}) \in \mathcal{H}$ and $b \in \mathbb{R}^p$,

$$\dot{v}^T(b)(a, \mathbf{h}) = b^T\dot{v}(a, \mathbf{h}) = \langle b, a\rangle_{\mathbb{R}^p}. \tag{2.46}$$

To find $\dot{v}_b^T$, we need a matching between the usual inner product $\langle b, a\rangle_{\mathbb{R}^p}$ with the inner product $\langle \dot{v}_b^T, (a, \mathbf{h})\rangle_{\mathcal{H}}$, for all $(a, \mathbf{h}) \in \mathcal{H}$. By considering the elements for which $a = 0$, we see that $\dot{v}_b^T$ is in the orthogonal complement of the linear span of the scores in the models $\Theta_{0,\mathbf{h}}$, which is the span of the $H_r(h_r)$ as we vary r and $h_r \in \mathcal{H}_r$. Write $\mathcal{A}$ for the projection onto this space, and write

$$\tilde{S}_Q = S_Q - \mathcal{A}S_Q$$

for the projection of the score function S_Q onto the orthogonal complement of this space. We recall from the form of (2.45) that the inner product on $\mathcal{H}$ takes the form, for $\bar{h}_1 =$

$(a_1, \mathbf{h}_1)$ and $\bar{h}_2 = (a_2, \mathbf{h}_2)$

$$\begin{aligned}\langle \bar{h}_1, \bar{h}_2 \rangle_{\mathcal{H}} &= \langle a_1^T S_{Q^*} + \sum_{r=1}^{R} H_r(h_{1r}), a_2^T S_{Q^*} + \sum_{r=1}^{R} H_r(h_{2r}) \rangle_{L^2} \\ &= \langle a_1^T \tilde{S}_{Q^*} + a_1^T \mathcal{A} S_{Q^*} + \sum_{r=1}^{R} H_r(h_{1r}), a_2^T \tilde{S}_{Q^*} + a_2^T \mathcal{A} S_{Q^*} + \sum_{r=1}^{R} H_r(h_{2r}) \rangle_{L^2} \\ &= \langle a_1^T \tilde{S}_{Q^*}, a_2^T \tilde{S}_{Q_*} \rangle_{L^2} + \langle a_1^T \mathcal{A} S_{Q_*} + \sum_{r=1}^{R} H_r(h_{1r}), a_2^T \mathcal{A} S_{Q_*} + \sum_{r=1}^{R} H_r(h_{2r}) \rangle_{L^2}.\end{aligned} \tag{2.47}$$

Write $\dot{v}_b^T = (a_b, \mathbf{h}_b)$. Then (2.47) gives that

$$\begin{aligned}\langle \dot{v}_b^T, (a, \mathbf{h}) \rangle_{\mathcal{H}} = \langle a_b^T \tilde{S}_{Q^*}, a^T \tilde{S}_{Q^*} \rangle_{L^2} \\ + \langle a_b^T \mathcal{A} S_{Q^*} + \sum_{r=1}^{R} H_r(h_{b,r}), a^T \mathcal{A} S_{Q^*} + \sum_{r=1}^{R} H_r(h_r) \rangle_{L^2}.\end{aligned}$$

We must then choose a_b and $\mathbf{h}_b$ to match with (2.46) as required in the statement of the theorem, at least insofar as to determine the efficient variance $\langle \dot{v}_b^T, \dot{v}_b^T \rangle_{\mathcal{H}}$. By considering inner products with elements in $\mathcal{H}$ of the form $(0, \mathbf{h})$, we see that $\dot{v}_b^T$ must be in the orthogonal complement of the space spanned by the $H_r(h_r)$ considered earlier. By choosing $a = b$, we see that

$$\langle a_b^T \tilde{S}_{Q^*}, b^T \tilde{S}_{Q^*} \rangle_{L^2} = a_b^T \tilde{J} b,$$

where $\tilde{J} = \mathbb{E}_*[S_{Q^*} S_{Q^*}^T]$ is the efficient Fisher information matrix. To get a matching between the above display and the usual L^2 inner product appearing on the right hand side of (2.46), we must choose $a_b = \tilde{J}^{-1} b$. This implies (using also the orthogonality established prior, and symmetry of $\tilde{J}$) that the efficient variance

$$\langle \dot{v}_b^T, \dot{v}_b^T \rangle_{\mathcal{H}} = b^T \tilde{J}^{-1} \tilde{J} (\tilde{J}^{-1} b) = b^T \tilde{J}^{-1} b.$$

By varying b, this characterises the covariance of $\mathbb{Z}_0$ as $\mathrm{Cov}(\mathbb{Z}_0) = \tilde{J}^{-1}$, the inverse efficient information matrix.

2.13 Details on MCMC algorithms used in Section 2.5

2.13.1 MCMC Algorithm for Π_1

In this section we describe the algorithm (Algorithm SA1) used for the simulation study in Section 2.5. S_R denotes the symmetric group of order R and where we use subscripts to refer to algorithm iterations. In our R implementation, used the RHmm package [97] for the forward-backward algorithm in the simulation of the hidden states and computation of the log likelihood, and the gtools package [104] for generation of permutations and Dirichlet draws. In our python implementation, we used the JAX package [14] to write both the forward-backward algorithm and the MCMC[7].

Algorithm SA1: Algorithm for MCMC draws targeting posterior of Q.

Input : Binned data $Y^{(M)} \in [\kappa_M]^n$, number of hidden states R, prior parameters γ,β, iterations I, initial hidden states value $\mathbf{X}_{init}$

Output: List of I draws of Q

Initialise $\mathbf{X}_1 = \mathbf{X}_{init}$

for $i = 1, \ldots, I + b$ **do**

  Draw transition matrix $Q_i \sim P(Q|\mathbf{X}_i)$, histogram weights $\underline{\omega}_i \sim P(\underline{\omega}|\mathbf{X}_i, Y)$, hidden states $\mathbf{X}_{i+1} \sim P(\mathbf{X}|Y, Q_i, \underline{\omega}_i)$

end

Compute $i_{MAP} = \arg\max_i \log \Pi_1(Q_i, \underline{\omega}_i|Y)$

for $i = 1, \ldots, I$ **do**

  $\tau_i = \arg\min_{\tau \in S_R} d\left(({}^{(\tau)}Q_i, {}^{(\tau)}\underline{\omega}_i), (Q_{i_{MAP}}, \underline{\omega}_{i_{MAP}})\right)$

end

return $\left\{{}^{(\tau_i)}Q_i\right\}_{i=1}^{I}$

2.13.2 MCMC Algorithm for Π_2

In Algorithm SA2, we detail the MCMC procedure we implement in R and python/JAX. Once again, we make use of the forward-backward algorithm to simulate latent states, allowing us to exploit the simple structure of the full likelihood. One key difference is that our algorithm requires allocation to a bivariate latent space (with total number of states $RS_{\max}$) involving both the HMM hidden state and the Dirichlet mixture component. Since the forward backward algorithm is $O(NK^2)$ with K the number of states and N the number of samples, and since we take $S_{\max} = \lfloor\sqrt{N}\rfloor$, the overall implementation is $O(N^2)$. The GPU capability of JAX code is of particular help here, as the computations for the forward-backward algorithm with a large number of states involve large matrices.

[7] All code is documented at https://github.com/dm1729/.

2.13.3 MCMC Algorithm for Π'

The MCMC algorithm of Π' is obtained by modifying Algorithm SA2 to remove the interior loop over $c = 1, \ldots, C$ and sampling a transition matrix Q from the conditional distribution given the latent states, as in Algorithm SA1, at the beginning of each outer loop over i. As with Algorithm SA2, the complexity is $O(N^2)$ and so using it to target the marginal posterior on Q is much slower than using Π_1, as discussed in Section 2.5.

Algorithm SA2: Algorithm for MCMC draws targeting cut posterior of emissions. Within the loop over c, we adopt the convention $W_i^0 = W_{i-1}$ and $W_{i-1}^0 = W_0$ etc.

Input : Data $Y \in \mathbb{R}^N$, Thinned transition matrix list $(Q_i)_{i=1,\ldots,\tilde{I}}$, number of hidden states R, prior parameters $M_0, \alpha_\sigma, \beta_\sigma, \mu_c, \sigma_c^2$, interior iterations C, initial hidden state value $\mathbf{X}_{init}$, truncation level S_{max}.

Output: List of $\tilde{I}$ draws of $\{f_r\}_{r=1}^R$

Initialise $\mathbf{X}_0 = \mathbf{X}_{init}$; variances $v^{(r)} \overset{iid}{\sim} \text{InvGamma}(\alpha_\sigma, \beta_\sigma)$; weights $W^{(r)} = (W_1^{(r)}, \ldots, W_{S_{max}}^{(r)}) \sim \text{Dir}(\alpha)$ with $\alpha = \alpha(M_0, S_{max})$ as in (2.23); allocation variables $S_0 = (S_{01}, \ldots, S_{0n})$ as in (2.22)

for $i = 1, \ldots, \tilde{I}$ **do**

  for $c = 1, \ldots, C$ **do**

    Sample locations $(\mu_{i1}^{c,(r)}, \ldots, \mu_{iS_{max}}^{c,(r)})_{r=1,\ldots,R}$ from $P(\mu_i^c | S_i^{c-1}, \mathbf{X}_i^{c-1}, Y)$

    Sample variances $(v_i^{c,(r)})_{r=1,\ldots,R}$ from $P(v_i^c | S_i^{c-1}, \mathbf{X}_i^{c-1}, \mu_i^c, Y)$

    Sample weights $(W_{i,1}^{c,(r)}, \ldots, W_{i,S_{max}}^{c,(r)})_{r=1,\ldots,R}$ from $P(W_i^c | S_i^{c-1}, \mathbf{X}_i^{c-1})$

    Sample latents (S_i^c, X_i^c) from $P(S_i^c, \mathbf{X}_i^c | \mu_i^c, v_i^c, W_i^c, Y, Q_i)$

  end

  Store $(\mu_i, v_i, W_i, S_i) = (\mu_i^C, v_i^C, W_i^C, S_i^C)$

end

for $i = 1, \ldots, \tilde{I}$ **do**

  Compute $f_i^{(r)}$ from (μ_i, v_i, W_i, S_i) by evaluating (an approximation of) (2.21)

end

return $\left\{ f_i^{(r)} \right\}_{i=1}^{\tilde{I}}$

Acknowledgements

The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the CDT StatML.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models
Publication Status	□Published ☑Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Accepted for publication in Electronic Journal of Statistics. Produced in collaboration with Prof. Judith Rousseau (University of Oxford).

Student Confirmation

Student Name:	Daniel Moss		
Contribution to the Paper	I contributed to the development of theory, both for semiparametric estimation and cut posterior, in close collaboration with Prof. Rousseau. It is difficult to separate the contributions each of us had in these parts. I personally wrote code for, and conducted, corresponding numerical experiments, with the directions of these investigations guided by discussions with Prof. Rousseau. I conducted most of the manuscript writing myself, again under guidance from Prof. Rousseau.		
Signature	D.P. Moss	Date	04/01/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Judith Rousseau		
Supervisor comments			
Signature		Date	04/01/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

3. Bayesian Model Selection in Semiparametric Mixtures and Hidden Markov Models

3.1 Introduction

Semiparametric estimation of the mixing weights π of a three component mixture model was considered in [41], as a means to circumvent the usual requirement to specify a parametric family on the emission densities, leveraging the identifiability result of [5]. The authors showed the existence of semiparametric efficient estimators for π and showed that, under suitable prior schemes, that a semiparametric Bernstein-von Mises theorem holds. These results in particular relied on a reduction to a multinomial emission model, which leads to a well-specified parametric model for appropriately binned data. Specifically this involves binning the continuous observed data into some collection $\mathcal{I}_M = (I_m^{(M)})_{m \leq M}$ of intervals which partition the observation space, and considering corresponding histogram priors over the emission densities. Subsequently, [82] considered a similar binning approach in the context of finite state space Hidden markov models, and demonstrated analogous results for the transition matrix Q under mild assumptions on the emissions, as required for identifiability by the results of [4, 44].

However, the works of [45, 82] considered inference when the number of hidden states R is known, in the context of both multidimensional mixtures and HMMs. While there are applications in which R is known (or at least fixed) in advance [79, 84, 107], there are several applications in which estimation of R is itself of interest, for which the BIC or AIC is typically employed [3, 30, 59, 60, 64, 106].

In this work, we seek to address this estimation problem in the context of both nonparametric mixtures and HMMs, by taking a Bayesian approach based on the marginal likelihood

in the multinomial models which arise from the binning techniques used in [45, 82]. In this way, we are able to identify the order of multidimensional mixtures and HMMs with nonparametric emissions. Our approach is quite similar to that of [66], who considered model selection in nonparametric HMMs via penalised likelihood-based and spectral estimators, and to [23], who considered this problem in the setting of *strongly identifiable* mixtures. In our work, we take a Bayesian approach based on the marginal likelihood; our method is therefore linked to the BIC approaches already taken in the applied literature, with the Bayesian Information Criterion being itself motivated by the marginal likelihood. For further background on BIC for singular model selection problems, see [34].

On the theoretical side, we give a simple result extending the underestimation result of [23], and similar also to the underestimation result of [66], to show that a penalised marginal likelihood approach does not underestimate the number of true states asymptotically. Essentially, the penalisation we propose is asymptotically outweighed by an exponential underestimation rate which arises from the misspecification of the model with too few states. It is then straightforward to prove that introducing this penalisation prevents us from overestimating R asymptotically; we prove also a negative result which highlights a difficulty in showing the validity of an unpenalised approach, which we might imagine would be straightforward by making standard arguments based on the posterior concentration results of [43, 92].

On the practical side, we implement the sequential importance sampling (SIS) algorithm of [52] for estimation of the marginal likelihood; the implementation is computationally tractable and allows for a more principled Bayesian approach to model selection than relying on approximations like the BIC. We examine the behaviour of our approach on simulated data, and our experiments indicate the validity of a penalised marginal likelihood approach, although suggest that the extent of the penalisation may not need to be so harsh in practice.

Organisation of the chapter In Section 3.2, we discuss the mixtures and HMMs we consider, explaining how the semiparametric models can be reduced to the categorical models we subsequently study. In Section 3.3, we introduce, and theoretically control, a penalised marginal likelihood approach, before discussing the need for penalisation; proofs for these results are deferred to Section 3.6. In Section 3.4, we implement the method in practice and discuss the penalisation in the context of an application to simulated data.

3.2 Models and Notation

Semiparametric latent variable models Consider latent variables $X_{1:n} = (X_1, \ldots, X_n)$, and observations $Z_{1:n} = (Y_1, \ldots, Y_n)$, for which $Y_{1:n}$ are conditionally independent given $X_{1:n}$ and satisfy

$$Y_i|(X_i = r) \sim F_r \quad 1 \leq r \leq R, \tag{3.1}$$

with $Y_i \in \mathcal{Y}$, and $X_i \in \{1, \ldots, R\} = [R]$. We place no parametric form on the F_r.

One important class of such latent variable models is multidimensional mixtures, considered in [45], in which $\mathcal{Y} = [0,1]^3$ and $F_r = F_{r,1} \otimes F_{r,2} \otimes F_{r,3}$. Here, the X_i are then independent and identically distributed according to some probabilities $(p_1, \ldots, p_R)$ on $[R]$, and the $\{F_{r,c} : r \in [R]\}$ are linearly independent for each $c = 1, 2, 3$. Identifiability is guaranteed under this structure, with minimal assumptions on the emissions, by the identifiability result of [5]. In particular, the $F_{r,c}$ may be modelled nonparametrically, leading to a semiparametric model.

Subsequently, [82] studied semiparametric HMMs, which can be cast in the form of (3.1) with $\mathcal{Y} = [0,1]$, and with the X_i following a stationary Markov chain with nonsingular transition matrix Q, and the $\{F_r : r = 1, \ldots, R\}$ linearly independent. The additional structure, arising from the latent chain, means that identifiability can be achieved without the multidimensional structure which is required in the mixtures (see [42] for details). Hence $\mathcal{Y}$ may be taken to be one-dimensional.

Such semiparametric models can be reduced to multinomial models, via binning of the data, as first described in [45] and subsequently considered in [82]. An interesting feature of both [45] and [82] is that the process of binning the data produces a multinomial model which is not misspecified as far as the mixture weights $p = (p_1, \ldots, p_R)$, or the transition matrix Q, are concerned, provided that a sufficiently fine-grained binning of $[0,1]$ is used to maintain identifiability of the binned model. In particular, the number of states R remains fixed, and so R may be estimated in a semiparametric way by taking a similar binning approach. In what follows, we describe the binning process for the latent variable models of [45] and [82], and the multinomial models which arise from them.

Multidimensional Mixture models By partitioning the interval $[0,1]$ into M bins $I_1, \ldots, I_M$, and considering cartesian products of such bins, we get a partition of $\mathcal{Y} = [0,1]^3$ into M^3 bins. When $Y_i \in I_{m_1} \times I_{m_2} \times I_{m_3}$, we define $\tilde{Y}_i = (m_1, m_2, m_3) \in [M]^3$. The

distribution of $\tilde{Y}_i$ is parameterised by the mixture weights $p \in \{p \in \mathbb{R}^R : p_i \geq 0, \sum_{i=1}^R p_i = 1\}$, and the per-dimension bin probabilities $\boldsymbol{\omega} = (\underline{\omega}_1, \underline{\omega}_2, \underline{\omega}_3) \in \Omega_M^3$, with

$$\Omega_M = \{\underline{\omega} \in [0,1]^{M \times R} \; ; \sum_{m=1}^{M} \omega_{m,r} = 1, \forall r \in [R]\}. \tag{3.2}$$

The data generating process for the $\tilde{Y}_i$ can then be expressed as

$$X_i \overset{\text{iid}}{\sim} p; \quad \tilde{Y}_i | X_i = r \sim \bigotimes_{c=1}^{C} \omega_{cr}. \tag{3.3}$$

The $\tilde{Y}_i \in [M]^3$ are thus independent and identically distributed, with likelihood

$$L(p, \omega | Y_i) = \sum_{r=1}^{R} p_r \prod_{c=1}^{3} \sum_{m=1}^{M} \omega_{c,m,r} \mathbb{1}_{Y_{i,c}=m}, \tag{3.4}$$

Hidden Markov Models Partitioning $[0,1]$ into M bins as before, we now consider HMM observations $Y_i \in \mathcal{Y} = [0,1]$ and write $\tilde{Y}_i = m$ when $Y_i \in I_m$. The distribution of $\tilde{Y}_i$ is now given by

$$X_1 \sim \nu; \quad X_i | X_{i-1} = r \sim (Q_{r,s})_{s=1}^R; \quad \tilde{Y}_i | X_i = r \sim (\omega_{rm})_{m=1}^M, \tag{3.5}$$

with $Q \in \{Q \in \mathbb{R}^{R \times R} : Q_i \geq 0, \sum_{s=1}^R Q_{rs} = 1 \quad \forall r \leq R\}$ and $\omega \in \Omega_M$. Since observations are non-iid, we specify the likelihood over n consecutive observations, which is given by

$$L_n(Q, \omega | Y_1, \ldots, Y_n) = \sum_{r_1, \ldots, r_n}^{R} \nu_{r_1} Q_{r_1, r_2} \ldots Q_{r_{n-1} r_n} \prod_{i=1}^{n} \sum_{m=1}^{M} \omega_{m, r_i} \mathbb{1}_{Y_i = m}. \tag{3.6}$$

Here ν is the stationary distribution of Q which exists under suitable ergodicity assumptions.

Notation We will denote by θ a parameter of either of model (3.3) or (3.5), in cases where the expressions apply for both models or when the model considered is clear, so that $\theta = (p, \omega_1, \omega_2, \omega_3)$ or $\theta = (Q, \omega)$. We denote by Θ_R, the set of all θ corresponding to a model with R states, again abusing notation so that Θ_R refers to both mixtures and HMMs. We note that the parameter spaces $(\Theta_R)_{R \geq 1}$ are nested, in the sense that an R-state mixture of HMM can be viewed as an $R+1$-state model in which one of the states is never visited by the latent variables.

In each case, we assume the data is generated according to a frequentist true parameter

$\theta^* \in \bigcup_{r=1}^{\infty} \Theta_R = \Theta_\infty$, and define

$$R(\theta) = \inf\{R : \theta \in \Theta_R, \theta \notin \Theta_{R-1}\}, \tag{3.7}$$

which exists as the Θ_R are nested. Then set $R^* = R(\theta^*)$. Finally, we will abuse notation and write Y_i for the binned version $\tilde{Y}_i$.

3.3 Estimation of R^*

In Section 3.3.1, we show the validity of a penalised sequential marginal likelihood approach, in the three-dimensional mixture and HMM contexts. The arguments are rather straightforward adaptations of those presented in [23]. In Theorems 3.1 and 3.2, we show consistency of this approach for the mixture and HMM case respectively.

In Section 3.3.2, we discuss the unpenalised version of the marginal likelihood approach. Since the marginal likelihood has inherent penalisation of the parameter dimension, it is not necessarily appealing to penalise further if not required. We present, in Proposition 3.4, a negative result which demonstrates, in our context, that the typical approach to showing consistency is flawed.

3.3.1 Consistency of Penalised Sequential Marginal Likelihood Approach

Consider a Bayesian model defined over Θ_∞, and with prior π over (R, θ) defined hierarchically so that

$$\pi(d\theta) = \sum_{R=1}^{\infty} \pi(R)\pi_R(d\theta). \tag{3.8}$$

Write $\ell_n^{(R)}(\theta)$ for the log-likelihood of θ given data $Y_1, \ldots, Y_n$, and write $\ell_n^* = \ell_n^{(R^*)}(\theta^*)$. The marginal likelihood of Θ_R, given observations $Y_1, \ldots, Y_n$, is then given by

$$\mathbb{B}_n(R+1) = \pi(R) \int_{\Theta_R} e^{\ell_n^{(R)}(\theta) - \ell_n^*} \pi_R(d\theta). \tag{3.9}$$

We now define the penalised sequential marginal likelihood estimator by

$$\hat{R}_n = \inf\{R : \log \mathbb{B}_n(R+1) \leq \log \mathbb{B}_n(R) + \mathrm{pen}_n(R)\} \tag{3.10}$$

where $\mathrm{pen} : \mathbb{N} \to \mathbb{R}^+$ is a non-negative penalty function. This estimate is similar to the $\hat{R}_n^L$ of [23], however we now introduce the penalty term. The estimator can be interpreted as

choosing the smallest value of R for which the log marginal likelihood does not subsequently grow too much, as specified by the penalty. The penalty can alternatively be seen as enforcing a particular choice of priors $\pi(R)$, which vary with the sample size n. We discuss the use of the penalty in more detail in Section 3.3.2.

The idea is based on the the notion that the log-marginal likelihood grows sharply in the range $R \in \{1, \ldots, R^*\}$ owing to the misspecifcation of the model, but in the overspecified case $R > R^*$ we do not have such a harsh drop off as the marginal likelihood in this case is for a well-specified model.

We first prove, in Theorem 3.1, consistency of $\hat{R}_n$ in the mixtures context. We require the following assumption on the emission weights.

Assumption 3.1. The emission weights satisfy $\omega^*_{c,m,r} > 0$ in the case of the mixture model (3.3), and $\omega^*_{m,r} > 0$ in the case of the HMM 3.5. Write $\omega^*_{c,r} = (\omega^*_{c,m,r})_{m=1}^M \in [0,1]^M$. Then the set $\{\omega^*_{c,r} : r \in [R]\}$ is linearly independent for each $c = 1, 2, 3$, in the mixture case. The same holds for the $\omega^*_r \in [0,1]^M$ defined similarly, in the HMM case.

Assumption 3.1 is a necessary condition for identifiability, and invertibility of the Fisher information matrix. It is a consequence of assumptions made in [45, 82], in the semiparametric context. It does not place any requirement of linear independence as we vary the dimension, c, in the mixture context. We note that $M \geq R$ is necessary for Assumption 3.1 to hold. We make also the following assumption on the prior.

Assumption 3.2. The prior $\pi_R(d\mathbf{p}, d\underline{\omega})$ factorises as $\pi_A(d\mathbf{p}) \prod_{r=1}^R \pi_B(d\underline{\omega}_r)$, where π_A, π_B have positive and continuous density with respect to Lebesgue measure, and the prior mass $\pi(R)$ of the $R-$state model is positive for all R.

Adapting slightly the proof of Theorem 1 of [23], we then obtain the following result:

Theorem 3.1. *Let $\mathbb{P}^*_n$ be the law of n observations from the mixture model (3.3). Let $\hat{R}_n$ be as in (3.10) and grant Assumptions 3.1 and 3.2. Then, for each $R = 1, \ldots, R^* - 1$, we have, for any penalty $\mathrm{pen}_n(R) \lesssim \log n$,*

$$\mathbb{P}^*_n(\hat{R}_n = R) \to 0 \quad \text{as } n \to \infty.$$

Assume also that $\mathrm{pen}_n(R^*) \geq \rho \log n$ for $\rho > \dfrac{D(R^*)}{2}$, with

$$D(R^*) = R^* - 1 + 3(M-1)R^*.$$

Then

$$\mathbb{P}_n^*(\hat{R}_n > R^*) \to 0 \quad \text{as } n \to \infty.$$

In the mixtures context of Theorem 3.1, the true mixture weights automatically satisfy $p_r^* > 0$ by definition of R^*. In the HMM context, we require the following additional regularity assumption on the transition matrix.

Assumption 3.3. Consider the HMM (3.5). We assume the transition matrix Q^* satisfies $Q_{ij}^* > 0$ for all i, j, and $\det Q^* \neq 0$.

Assumption 3.4. The prior $\pi_R(dQ, d\underline{\omega})$ factorises as $\prod_{s=1}^R \pi_A(Q_s) \prod_{r=1}^R \pi_B(d\underline{\omega}_r)$, where π_A, π_B have positive and continuous density with respect to Lebesgue measure. Here, Q_s denotes row s of the transition matrix, and the prior mass $\pi(R)$ of the $R-$state model is positive for all R.

Theorem 3.2. *Let $\mathbb{P}_n^*$ be the law of n observations from the hidden Markov model (3.5). Let $\hat{R}_n$ be as in (3.10) and grant Assumptions 3.1, 3.3 and 3.4. Then, for each $R = 1, \ldots, R^* - 1$, we have, for any penalty $\mathrm{pen}_n(R) \lesssim \log n$,*

$$\mathbb{P}_n^*(\hat{R}_n = R) \to 0 \quad \text{as } n \to \infty.$$

Assume also that $\mathrm{pen}_n(R^*) \geq \rho \log n$ for $\rho > \dfrac{D(R^*)}{2}$, with

$$D(R^*) = R^*(R^* - 1) + (M - 1)R^*$$

Then

$$\mathbb{P}_n^*(\hat{R}_n > R^*) \to 0 \quad \text{as } n \to \infty.$$

The proofs of Theorems 3.1 and 3.2 are presented in Section 3.6.

Remark 3.3. Theorems 3.1 and 3.2 each comprise an underestimation and overestimation result. The underestimation result is a straightforward adaptation of the underestimation result of [23], to the context of multidimensional mixtures and HMMs, who prove the unpenalised case under assumptions which we verify for our model. Their result shows that an exponential rate is achieved for the underestimation probability, owing to the Kullback-Leibler separation between Θ_{R^*} and Θ_R when $R < R^*$. Theorems 3.1 and 3.2 strengthen slightly this underestimation result by showing that, under penalisation of the log marginal likelihood which is at most logarithmic, we still do not underestimate

R asymptotically. The overestimation bound is rather crude and, in Section 3.3.2, we elaborate on the difficulties encountered when attempting to relax this penalisation.

3.3.2 Regarding the penalisation

It may be possible to weaken, or eliminate entirely, the need to penalise, if we can control the quantity

$$\mathbb{P}_*(\log \mathbb{B}_n(R^*+1) - \mathrm{pen}(R^*) > -\frac{D(R^*)}{2}\log n + C), \tag{3.11}$$

potentially with $\mathrm{pen}(R^*) = 0$, in a more sophisticated way then our approach in Theorems 3.1 and 3.2, which directly applies Markov's inequality. This is because we expect that the quantity $\mathbb{B}_n(R^*+1)$ is highly concentrated, in the sense that there exists a sequence $b_n \to 0$ such that $\mathbb{P}_*(\mathbb{B}_n(R^*+1) > b_n) \to 0$ despite its $\mathbb{E}_*$-expectation being constant.

Starting from a posterior concentration result, let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets on which the posterior over Θ_{R^*+1} concentrates. Then

$$\mathbb{B}_n(R^*+1) \leq 2\pi(R^*+1)\int_{A_n} e^{\ell_n^{(R^*+1)}(\theta)-\ell_n^*}\pi_{R^*+1}(d\theta)$$

as soon as $\pi_{R^*+1}(A_n|x) > \frac{1}{2}$. Now Markov's inequality and Fubini's theorem imply that

$$\begin{aligned}\mathbb{P}_*(\mathbb{B}_n > b_n) &\leq \frac{2\pi(R^*+1)}{b_n}\mathbb{E}_*\int_{A_n} e^{\ell_n^{(R^*+1)}(\theta)-\ell_n^*}\pi_{R^*+1}(d\theta)\\ &\leq \frac{2\pi(R^*+1)}{b_n}\pi_{R^*+1}(A_n)\end{aligned} \tag{3.12}$$

and so for any sequence $b_n \to 0$ such that $\pi_{R^*+1}(A_n) = o(b_n)$, we get $\mathbb{P}_*(\mathbb{B}_n > b_n) \to 0$. The key point is that we wish to restrict to the set of high posterior probability (and low prior mass) before applying Markov's inequality.

Typically A_n will take the form $A_n = \{\theta \in \Theta_{R^*+1} : \|g_\theta^{(R^*+1)} - g_{\theta^*}\|_1 < \epsilon_n\}$. For ease of exposition, we restrict our discussion in what follows to the mixtures setting. However, the ideas translate to HMMs by considering the likelihood on three consecutive observations in place of g_θ. Such sets are identified as having high posterior mass through an application of results such as [92] in the mixtures setting, or [43] in the HMM setting.

However, the prior is typically understood in terms of the parameters, being the emission weights and, depending on context, the mixture weights or transition matrix. We thus wish to relate the A_n, defined in terms of concentrations of marginal densities, to neighbourhoods characterised by distances in the parameter space, to bound its prior mass.

Such an approach is taken by [23], who prove a lower bound of the form

$$\|g_\theta^{(R^*+1)} - g_{\theta^*}\|_1 \geq C(\theta^*)L(\theta, \theta^*) \tag{3.13}$$

for g_θ the likelihood, which for our context would be given by (3.4), and a function $L(\theta, \theta^*)$, for which we can bound the prior mass of θ such that $L(\theta, \theta^*)$ is small. Then, when $\tilde{\epsilon}_n = C\epsilon_n$ for large enough $M > 0$, $A_n \subset \{\theta : L(\theta, \theta^*) < \tilde{\epsilon}_n\}$, and we get a tighter control in (3.12), with the penalty required ultimately depending on the posterior rate ϵ_n, through the associated value of b_n. In the case of mixtures, where a rate $\epsilon_n = \frac{\sqrt{\log n}}{n}$ is achieved in [92], penalisation of the order $\epsilon \log n$ would suffice for any $\epsilon > 0$; more penalisation may be required in the HMM case, even under 3.13, when employing the posterior concentration established in [43].

We can think of $L(\theta, \theta^*)$ as an expression of distance, which should account for the directions within θ_{R^*+1} from which we may approximate the parameter $\theta_* \in \Theta_{R^*}$, corresponding to either "merging" or "emptying" of the redundant component, as discussed in [92] (see also the formulation of the mixtures proofs in Appendix C of [23]). For our setting, the analogous expression L is detailed below in the statement of Proposition 3.4.

In [23], the authors prove a result of the form (3.13) under the *strong identifiability* condition, but this condition does not hold in our setting. We might have anyway hoped that the bound (3.13) would hold in our context of multinomial mixtures, for a natural choice of L in the multinomial context. However, in Proposition 3.4, we show that this is not the case.

To set notation for what follows, denote, for $\theta \in \Theta_{R^*+1}$, $\omega_s = (\omega_{c,m,s} : c \leq 3, m \leq M)$ for $s \leq R^* + 1$, and $p = (p_s : s \leq R^* + 1)$. Similarly for $\theta^* \in \Theta_{R^*}$, define ω_r^*, p_r^*, for $r \leq R^*$. Now set, for $\epsilon > 0$ and for $r \in \{1, \ldots, R^*\}$, $A(r) = \{s \leq R^* + 1 : \|\omega_s - \omega_r^*\| \leq \epsilon\}$, and set $A(0) = \{1, \ldots, R^* + 1\} \setminus \bigcup_{r=1}^{R^*} A(r)$. When $\epsilon > 0$ is sufficiently small, the sets $A(\cdot)$ are thus non-overlapping provided the ω_r^* are distinct, which is guarenteed whenever Assumption 3.1 holds. Finally, define $\bar{p}_r = \sum_{s \in A(r)} p_s$, $\bar{p}_0 = \sum_{s \in A(0)} p_s$, $q_s = p_s/\bar{p}_s$ $(s \leq R^* + 1)$, $d_s = \omega_s - \omega_r^*$ (when $s \in A(r)$) and $\bar{d}_r = \sum_{s \in A(r)} q_s d_s$. These definitions mirror those made in [23], and are used to define the lower bound $L(\theta, \theta^*)$. The $A(r)$ correspond to components which are *merging* into the state r, while $A(0)$ corresponds to the *emptying* components.

Proposition 3.4. *Let $\mathbb{P}_n^*$ be the law of n observations from the mixture model (3.3), for*

a choice of θ^* satisfying Assumption 3.1 which we detail in the proof. Then there exists a sequence of parameters $(\theta_j)_{j\geq 1}$, with $\theta_j \in \Theta_{R^*+1}$, for which

$$\lim_{j\to\infty} \frac{\|g_{\theta_j}^{(R^*+1)} - g_{\theta^*}\|}{L(\theta_j, \theta^*)} = 0, \tag{3.14}$$

with $L(\theta, \theta^*)$ given by

$$L(\theta, \theta^*) = \bar{p}_0 + \sum_{r=1}^{R^*} |\bar{p}_r - p_r^*| + \sum_{r=1}^{R^*} \|\bar{d}_r\| + \sum_{s=1}^{R^*+1} q_s \|d_s\|^2. \tag{3.15}$$

Remark 3.5. While Proposition 3.4 identifies an issue with the approach we detail above, we do not claim that it is impossible to weaken the penalisation - only that something goes wrong with this typical approach. One could imagine carefully choosing sequences of priors which restrict directions along which the approximation of $\|g_\theta^{(R^*+1)} - g_{\theta^*}\|$ by $L(\theta, \theta^*)$ is poorly behaved. Alternatively, one could may seek to prove a weakened bound, in which the constant $C(\theta^*)$ in (3.13) is relaxed to a (slowly) convergent sequence $C_n(\theta^*) \to 0$. In Section 3.4, we present simulations that suggest a weaker penalty would suffice, although it remains unclear whether penalisation may be entirely avoided, especially in the HMM setting where there is the additional complexity involved in approximating $Q^* \in \Theta_{R^*}$ by singular matrices (see [43], for instance),

3.4 Experiments

In this section, we consider the practical implementation of the Bayesian model selection methodology. In Section 3.4.1, we discuss the computational approach used for estimating HMMs in multinomial mixtures. In Sections 3.4.2 and 3.4.3, we demonstrate this approach on simulated data, both directly on categorical models, and also for $\mathbb{R}$-valued data which we reduce via binning.

3.4.1 Algorithm Description

For the marginal likelihood estimation, we implement the Sequential Importance Sampling algorithm (SIS) of [52]. This provides an efficient, parallelisable method of estimating marginal likelihoods in latent variable models (including mixtures and HMMs) in which the posterior given both the observable data *and* the unobserved latents is conjugate. The algorithm sequentially samples the latent variables X_m conditional on the observations $Y_{1:t-1}$ and the already-sampled latent variables $X_{1:t-1}$, based on how each sample in

sequence would affect the joint marginal likelihood of $X_{1:t}, Y_{1:t}$, which is available in closed form.

This process provides an unbiased estimate of the marginal likelihood of $Y_{1:n}$; the process is then repeated independently I times, generating a different sequence $X_{1:n}$ in each instance, with the results averaged to give a final estimate. The algorithm is $O(n^2)$ in the size of the data, and $O(I)$ in the number of independent loops.

In the mixtures settings, we consider a Dirichlet prior π_α with parameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ across the weights p; for the HMM case the rows of the transition matrix are independent and identically distributed as π_α, under the prior. The per-dimension, per-state emission probabilities $\omega_{c,r} \in [0,1]^M$ (or ω_r for the HMM), are distributed with Dirichlet prior π_β with parameter $\beta = (\beta_1 \ldots, \beta_M)$. The prior hyperparameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ and $\beta = (\beta_1, \ldots, \beta_M)$; we set all co-ordinates of α and β to be one.

To set notation, first write

$$C_r(X_{1:t}) = \{X_i : i \leq t, X_i = r\}; \quad t \leq n. \tag{3.16}$$

Now write, for $r \leq R$, and for $C \in \{1, 3\}$ (for HMM or mixture),

$$m_r(Y_{1:t}; C_{r,t}) = \prod_{c=1}^{C} \int_{[0,1]^M} \prod_{i \in C_{r,t}} \omega_{c,y_{i,c},r} \pi_\beta(d\omega_{c,r}). \tag{3.17}$$

Also, define $n_r(X_{1:t}) = |C_r(X_{1:t})|$. We detail the SIS procedure of [52] in Algorithm 3, for the mixtures setting[1].

The expression (3.17) is straightforward to evaluate in closed form by conjugacy of the Dirichlet-multinomial. Since repeats of the outer loop in Algorithm 3 are independent, parallelisation via GPU is straightforward, as noted in [52]. We leverage this this in our implementation[2] through the use of the JAX library [14]; computations were run using an NVIDIA Ampere A40.

3.4.2 Categorical Simulations

We first examine data simulated according to categorical distributions. We consider both multidimensional mixtures and HMMs. We consider, for each model, six parameter con-

[1] The only modification in the HMM setting is in the definition of γ_r, where the factor involving n_r and α_r is adapted to the posterior distribution of $Q_{X_{t-1},\cdot}$ given $X_{1:t-1}$.

[2] Our implementation can be found at www.github.com/dm1729/HMMs

Algorithm 1: SIS Algorithm of [52] (Mixtures Case)

Input : Binned data $Y_{1:n}$, number of hidden states R, Dirichlet prior parameters α, β, independent loops I.
Output: Estimate of marginal likelihood $m_n(R)$
Sample initial latent $X_1 \sim \alpha$.
Initialize $w^{(i)} = m_r(Y_1|C_{r,1})$ with $r = X_1$
for $i = 1, \ldots, I$ **do**
  for $t = 2, \ldots, n$ **do**
    for $r = 1, \ldots, R$ **do**
      Set $C_{r,t-1} = C_r(X_1, \ldots, X_{t-1})$
      Set $\tilde{X}_t = r$, $\tilde{C}_{r,t} = C_r(X_1, \ldots, X_{t-1}, \tilde{X}_t)$
      Compute $\gamma_r = \dfrac{m_r(Y_{1:t}; \tilde{C}_{r,t})}{m_r(Y_{1:t-1}; C_{r,t-1})} \times \dfrac{n_r(X_{1:t-1}) + \alpha_r}{\sum_{r=1}^{R} n_r(X_{1:t-1}) + \alpha_r}$
    end
    Sample $X_t \sim \left(\dfrac{\gamma_1}{\sum_{r=1}^{R} \gamma_r}, \ldots, \dfrac{\gamma_R}{\sum_{r=1}^{R} \gamma_r} \right)$
    Update $w^{(i)} = w^{(i)} \sum_{r=1}^{R} \gamma_r$
  end
end
return $\hat{m}_n(R) = I^{-1} \sum_{i=1}^{I} w^{(i)}$

figurations. The configurations cover $R^* \in \{2, 3, 4\}$, and for each R^* two configurations for the emission probabilities are considered. The observations take values in $\{1, \ldots, M\}$ where $M = R^* + 1$.

Parameter configurations Consider probabilities over $\{1, \ldots, R^*\}$, with $R^* \in \{2, 3, 4\}$, given by

$$p_2 = (0.6, 0.4) \quad p_3 = (0.3, 0.3, 0.4) \quad p_4 = (0.3, 0.25, 0.2, 0.25)\,. \tag{3.18}$$

Consider also the transition matrices defined as

$$Q_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.05 & 0.1 & 0.85 \end{pmatrix} \quad Q_4 = \begin{pmatrix} 0.8 & 0.1 & 0.05 & 0.05 \\ 0.1 & 0.7 & 0.15 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \\ 0.05 & 0.025 & 0.025 & 0.9 \end{pmatrix}. \tag{3.19}$$

Next, consider parameters $\bar{\omega}_2^{(A)}, \bar{\omega}_3^{(A)}, \bar{\omega}_4^{(A)}, \in \Omega_M$, by

$$\bar{\omega}_2^{(A)} = \begin{pmatrix} 0.6 & 0.2 \\ 0.2 & 0.6 \\ 0.2 & 0.2 \end{pmatrix} \quad \bar{\omega}_3^{(A)} = \begin{pmatrix} 0.4 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 \end{pmatrix} \quad \bar{\omega}_4^{(A)} = \begin{pmatrix} 0.4 & 0.15 & 0.15 & 0.15 \\ 0.15 & 0.4 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.4 & 0.15 \\ 0.15 & 0.15 & 0.15 & 0.4 \\ 0.15 & 0.15 & 0.15 & 0.15 \end{pmatrix}. \tag{3.20}$$

Finally, consider parameters $\bar{\omega}_2^{(B)}, \bar{\omega}_3^{(B)}, \bar{\omega}_4^{(B)}, \in \Omega_M$, by

$$\bar{\omega}_2^{(B)} = \begin{pmatrix} 0.8 & 0.1 \\ 0.1 & 0.8 \\ 0.1 & 0.1 \end{pmatrix} \quad \bar{\omega}_3^{(B)} = \begin{pmatrix} 0.7 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.7 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \quad \bar{\omega}_4^{(B)} = \begin{pmatrix} 0.8 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.8 \\ 0.05 & 0.05 & 0.05 & 0.05 \end{pmatrix}. \tag{3.21}$$

For the mixture model simulations, we then consider three-dimensional data generated according to

$$X_i \sim p_{R^*} \quad (Y_{i1}, Y_{i2}, Y_{i3}) | (X_i = r) \sim \underline{\omega}_r^{(\cdot)} \otimes \underline{\omega}_r^{(\cdot)} \otimes \underline{\omega}_r^{(\cdot)}, \tag{3.22}$$

where we vary $R^* \in \{2, 3, 4\}$ and consider emissions governed both the A and B superscript parameters. Thus, in our simulations, the three components of the observation are conditionally iid given the hidden state, though we nevertheless do not enforce this in the prior as detailed below.

For the HMM simulations, we consider one-dimensional data generated according to

$$X_i \sim MC(Q_{R^*}) \quad Y_i | X_i = r \sim \underline{\omega}_r, \tag{3.23}$$

again varying R^* and the superscript of $\underline{\omega}$. The X_i are initialized in their stationary distribution.

The idea of these parameter configurations is, in each case, to ensure the R^*-state latent structure is well represented in the data, but to adjust the relative difficulty of inference based on the separation in the emission distributions.

In each case, we fit to a dataset of size $N \in \{1600, 3200, 6400, 12800\}$. The datasets of size

$(\underline{\omega}^{(A)})$	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	2	2	3	1	3	1
3200	3	2	3	1	2	1
6400	3	2	3	1	4	1
12800	2	2	4	1	3	2
$(\underline{\omega}^{(B)})$	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	2	2	3	3	5	4
3200	2	2	3	3	5	4
6400	2	2	4	3	5	4
12800	2	2	4	3	5	4

Table 3.1: Values of $\hat{R}_n$ for unpenalised estimator ($\mathrm{pen}_n(R) = 0$) and penalised estimator ($\mathrm{pen}_n(R) = \frac{D(R)}{2} \log n$), categorical multidimensional mixtures. Above: emission configuration A. Below: emission configuration B.

$(\underline{\omega}^{(A)})$	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	2	1	1	1	5	1
3200	2	2	2	1	2	1
6400	2	2	2	1	2	1
12800	2	2	2	1	2	1
$(\underline{\omega}^{(B)})$	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	3	2	2	2	3	3
3200	3	2	3	2	3	3
6400	3	2	3	2	3	3
12800	2	2	3	3	4	3

Table 3.2: Values of $\hat{R}_n$ for unpenalised estimator ($\mathrm{pen}_n(R) = 0$) and penalised estimator ($\mathrm{pen}_n(R) = \frac{D(R)}{2} \log n$), categorical HMMs. Above: emission configuration A. Below: emission configuration B.

$N < 12800$ are restrictions of the full dataset of size $N = 12800$.

Prior distributions For the mixtures, we equip each $(\omega_{r,c,m} : 1 \leq m \leq M) \in [0,1]^M$ with a Dirichlet prior with equal, unit weights (so uniform over the simplex), independently for each $r \in [R]$ and $c \in \{1,2,3\}$. We equip the mixture weights $(p_r)_{r=1}^R$ with a uniform prior over the simplex. Analogously, for the HMM, we equip each $(\omega_{r,m} : 1 \leq m \leq M) \in [0,1]^M$ with the same prior as for the mixture emission weights, independent across $r \in [R]$. For the transition matrix, we equip each row $(Q_{rs})_{1 \leq s \leq R}$ with a uniform prior over the simplex, independently across $1 \leq r \leq R$.

Results and Discussion In both the HMMs and mixtures, we see that the proecedures typically perform more poorly when using emission configuration A, as to be expected.

In configuration B, for the mixtures case, we see that the penalised version outperforms the unpenalised version for all values of and R^*. We plot in Figure 3.1 the cumulative output of the SIS algorithm for the $R^* = 4, N = 12800$ subcase, to understand the role of the penalisation.

Contrarywise, the unpenalised version generally outperforms the penalised version for inference in the categorical HMMs, for both emission configurations. In Figure 3.2, we plot the HMM analogue of Figure 3.1, i.e. the $R^* = 4, N = 12800$ subcase of emission configuration B.

In Figure 3.1, we see that there is a large gap between the underspecified and well-specified cases, and then the overspecified evidence estimate ($R = 5$) is almost identical to the well-specified evidence estimate ($R = 4$). This is congruent with what we expect from the theory, which shows that the evidence decays much more rapidly when we are in the $R < R^*$ regime as the model is misspecified. In this case, the penalisation allows us to choose the correct number of states. In Figure 3.2, we see that the penalty causes us to incorrectly reject the $R = 4$ case, despite a relatively large increase in evidence between $R = 3$ and $R = 4$.

Jointly, Figures 3.1 and 3.2 suggest that a small degree of penalisation may be preferable to a completely unpenalised estimator, but that this penalisation should be made relatively small so as to not overlook relatively large increases in the marginal likelihood.

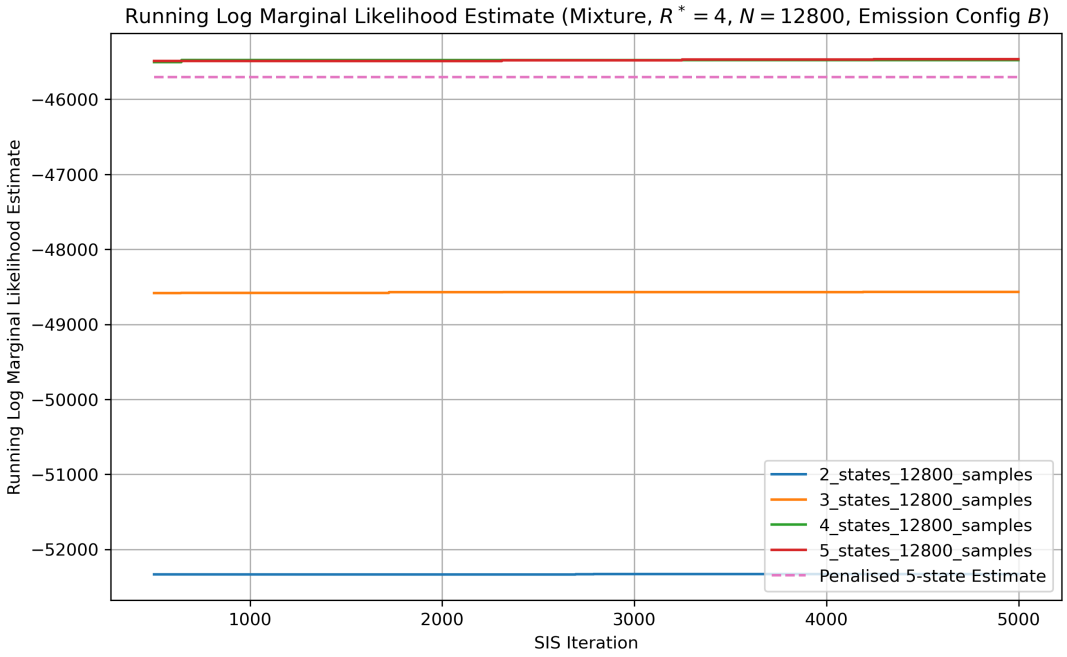


Figure 3.1: Running log-marginal likelihood estimate (sequential binning) as we increase the number of SIS iterations, for the categorical mixture case, with $R^* = 4$, $N = 12,800$, and emission configuration B. Plotted in pink, we see the penalised threshold to choose $R = 5$ (red line) over $R = 4$ (green line). The increase in evidence from $R = 4$ states to $R = 5$ states does not overcome our penalisation and we correctly select the four state model in the penalised setting, while incorrectly choosing five states in the unpenalised setting.

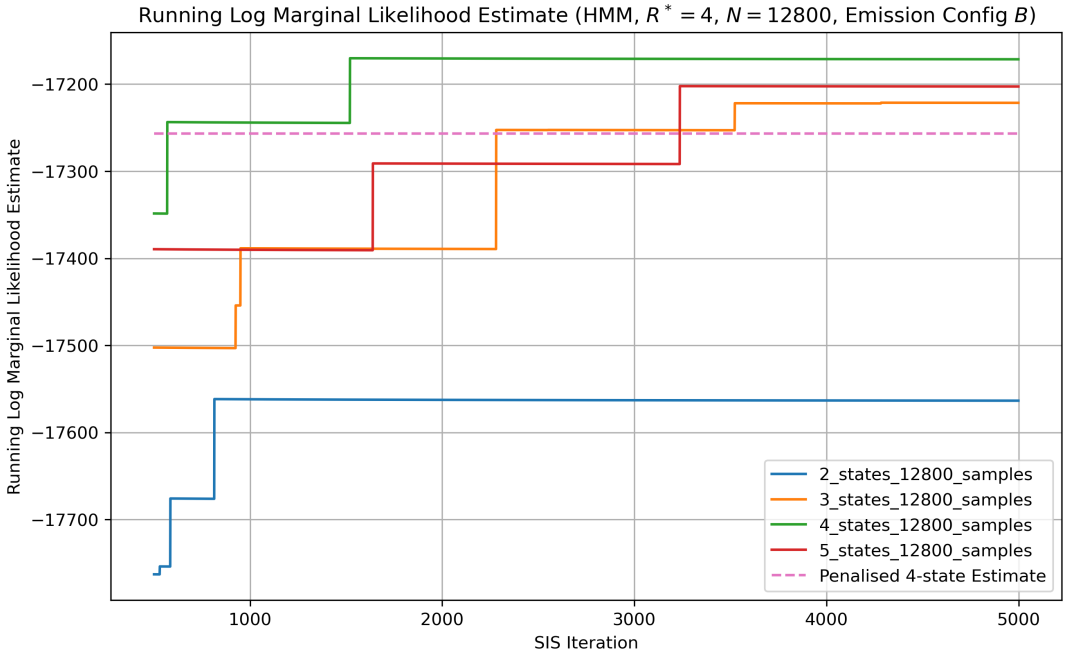


Figure 3.2: Running log-marginal likelihood estimate (sequential binning) as we increase the number of SIS iterations, for the categorical HMM case, with $R^* = 4$, $N = 12,800$, and emission configuration B. Plotted in pink, we see the penalised threshold to choose $R = 4$ (green line) over $R = 3$ (orange line). The increase in evidence from $R = 3$ states to $R = 4$ states does not overcome our penalisation, and we incorrectly select the three state model in the penalised setting, while correctly choosing the four state model in the unpenalised setting.

3.4.3 Semiparametric Simulations

We now examine data generated according to HMMs, with emission distributions over $\mathbb{R}$, with varying tail behaviour. We demonstrate the marginal likelihood estimation in the semiparametric context - first binning the data, then estimating the evidence in the corresponding multinomial model.

Parameter configurations We consider HMMs with transition matrices Q_2, Q_3 and Q_4 as defined in the previous section, but now with emission distributions take the form

$$F^*_{r,\nu} = \mu^*_r + F_\nu \tag{3.24}$$

where F_ν is a t-distribution with ν degrees of freedom. Hence, for a fixed choice of ν, the data is generated as

$$X_i \sim MC(Q_{R^*}) \quad Y_i | X_i = r \sim F^*_{r,\nu}. \tag{3.25}$$

We consider the $\nu = \infty$ case (Gaussian), and the case $\nu = 4$. These are centred, depending on R^*, at the means given by

$$\mu_2 = (-1, 1) \quad \mu_3 = (-2, 0, 2) \quad \mu_4 = (-3, -1, 1, 3). \tag{3.26}$$

Binning and Prior distributions We adopt a binning approach similar to that of [82], though we adapt it to the data, by transforming the $\mathbb{R}$-valued data to $[0, 1]$ by a link function of the form

$$G_0(y) = \begin{cases} (1 + e^{-y})^{-1} & y < a_1, y > a_2 \\ \zeta + \eta y & y \in (a_1, a_2) \end{cases}$$

with a_1, a_2 the 5th and 95th percentile of the full dataset of size $n = 12800$, and $\zeta, \eta \in \mathbb{R}$ chosen so that G_0 is continuous. The unit interval is then binned into intervals of equal length.

We consider two binning schemes:

- *Sequential Binning:* When comparing the model with R states to the model with $R + 1$ states, we bin the data with $M = R + 1$ bins. We then rebin the data for the next sequential comparison.
- *Global Binning:* We set an a-priori upper bound $R_{\max}$ on R, and bin our data with $M = R_{\max}$ before comparing the different values in R in sequence to determine $\hat{R}_n$.

Student's-t	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	3	1	2	2	2	2
3200	2	2	2	2	2	2
6400	2	2	2	2	2	2
12800	2	2	2	2	2	2
Gaussian	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	3	2	2	2	2	2
3200	3	2	2	2	2	2
6400	3	2	3	2	3	2
12800	2	2	3	2	4	4

Table 3.3: Values of $\hat{R}_n$ for unpenalised estimator ($\text{pen}_n(R) = 0$) and penalised estimator ($\text{pen}_n(R) = \frac{D(R)}{2} \log n$), semiparametric HMMs, global binning (five bins). Top: t distribution with $\nu = 4$ degrees of freedom. Bottom: Gaussian distribution.

Student's-t	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	2	2	3	2	4	3
3200	2	2	3	3	4	3
6400	2	2	3	2	4	3
12800	3	2	3	3	4	3
Gaussian	$R^* = 2$		$R^* = 3$		$R^* = 4$	
N	Unpenalised	Penalised	Unpenalised	Penalised	Unpenalised	Penalised
1600	3	2	3	2	3	3
3200	3	2	4	3	5	3
6400	2	2	4	3	5	3
12800	3	2	4	4	4	3

Table 3.4: Values of $\hat{R}_n$ for unpenalised estimator ($\text{pen}_n(R) = 0$) and penalised estimator ($\text{pen}_n(R) = \frac{D(R)}{2} \log n$), semiparametric HMMs, sequential binning. Top: t distribution with $\nu = 4$ degrees of freedom. Bottom: Gaussian distribution

In our case, we set $R_{\max} = 5$.

Finally, given the binned data, we place prior distributions identical to those considered for the categorical HMMs of Section 3.4.2.

Results and Discussion Tables 3.3 and 3.4 give the values of the estimators based on the global binning procedure, and the sequential binning procedure, respectively. We generally found that the penalisation was too harsh when binning globally, since the model dimension $D(R)$, which drives the penalisation, depends on M which in this case was quite high even for small values of R. As in the categorical simulations, the unpenalised version would at times overestimate R^*, while underestimation was much more common for the penalised version (even when binning sequentially). The results suggest that some penalisation may be necessary, although the extent of our penalisation is possibly too

severe.

In Figures 3.3 and 3.4, we examine the output of the sequentially binned algorithm, for the Gaussian and Student's t emissions case respectively. Each figure shows the evolution of the log-marginal likelihood estimate, as we increase the number of SIS iterations, in the case $N = 12800$ and $R^* = 4$. In each case, the penalisation does not prevent us from choosing $R = 3$ compared to $R = 2$, but the penalisation does prevent us from choosing $R = 4$ compared to $R = 3$, where the unpenalised version is able to do so. The Figures demonstrate, once again, that a smaller level of penalisation, as discussed in Section 3.4.2, would be preferable in order to not reject these values while still controlling the overestimates we sometimes see when using the fully unpenalised estimator.

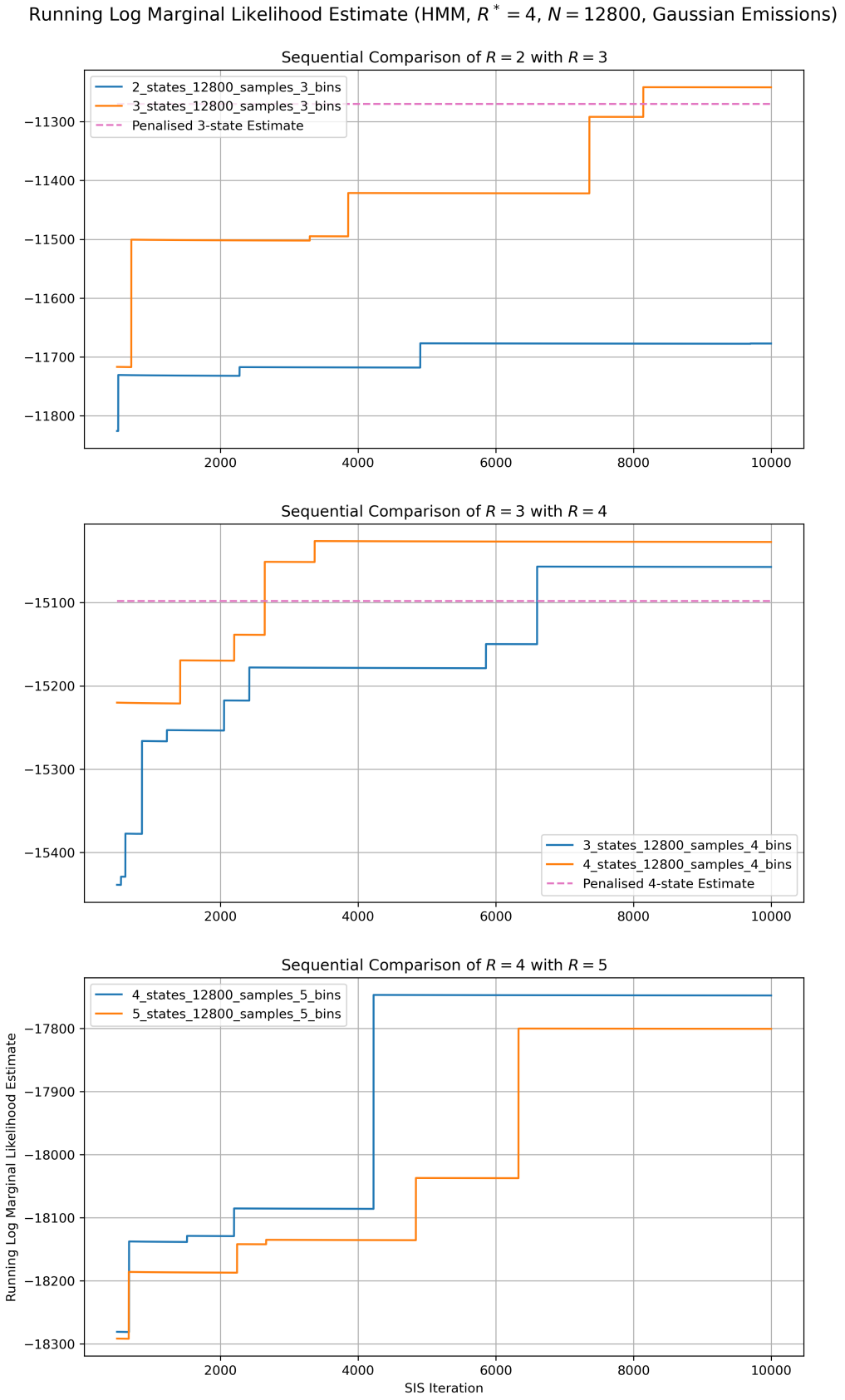


Figure 3.3: Running log-marginal likelihood estimate (sequential binning) as we increase the number of SIS iterations, for the Gaussian emission case with $R^* = 4$ and $N = 12,800$. Plotted in pink, we see the penalised threshold to choose the higher R value (orange line). The increase in evidence from $R = 3$ states to $R = 4$ states does not overcome our penalisation.

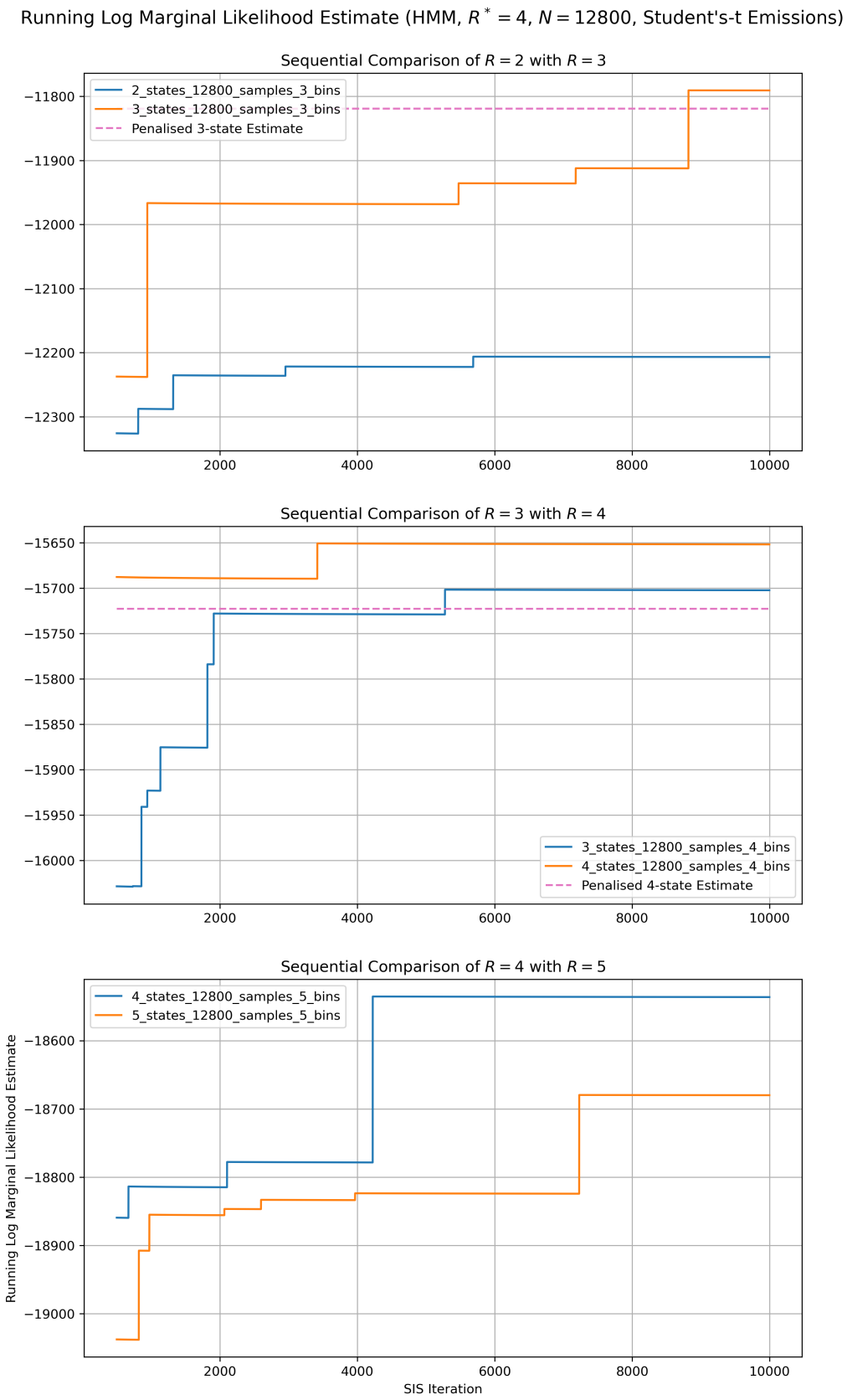


Figure 3.4: Running log-marginal likelihood estimate (sequential binning) as we increase the number of SIS iterations, for the Student's t ($\nu = 4$ degrees of freedom) emission case with $R^* = 4$ and $N = 12,800$. Plotted in pink, we see the penalised threshold to choose the higher R value (orange line). The increase in evidence from $R = 3$ states to $R = 4$ states does not overcome our penalisation.

3.5 Conclusion

In our work, we have demonstrated the possibility to estimate the number of hidden states R using the marginal likelihood, for application in semiparametric mixture models and HMMs of the kind studied in [45, 82].

On the theoretical side, we have shown the validity of a penalised marginal likelihood approach, which leverages the exponential underestimation rate identified in [23], in the mixtures setting, and [66] in the HMM setting beforehand. In our investigations of an unpenalised approach, we identify a shortcoming with establishing a bound of the form (3.13), which would have permitted control of high-posterior regions under the prior in an approach similar to [23]. We believe this method has particular interest as a substitute for the BIC-based model selection approaches often employed in practice, whose justification is ultimately based on marginal likelihood based approaches.

We then implement the method in practice, using the computational approach to calculating marginal likelihoods of [52]. Our computations suggest that some penalisation may indeed be useful, although indicate that the penalty under which our theoretical guarantees are established may, in practice, be too harsh.

An interesting future direction of study could pursue questions of optimal penalisation. Theoretically, this could be addressed by studying priors which place little mass on sets for which the quantity $\|g_\theta^{(R^*+1} - g_{\theta^*}\|$ is difficult to control, or through an explicitly weaker global lower bound to achieve a less harsh penalisation, as suggested by experiments. An extended computational investigation into such penalties could also shed light on how best to tune the method in practice.

3.6 Proofs

3.6.1 Proof of Theorems 3.1 and 3.2

We present the proof in detail for Theorem 3.1, which primarily involves adapting the work of [23]. The proof of Theorem 3.2 then follows by straightforward adaptations of the arguments of Theorem 3.1 to the multinomial HMM setting, accounting for the dependence between observations using the same ideas as in the Bernstein-von Mises result of [28].

The following result mirrors Lemma 2 of [23]. We show in the proof that, although the assumptions under which their lemma is derived do not hold in our setting, we may

nevertheless reach the same conclusion.

Lemma 3.6. *Let $R \leq R^*$ and $\delta \in (0, \alpha M \wedge \delta_0)$, with α, M, δ_0 constants defined in the proof. We define H, H_{R^*} as*

$$H(\theta) = \mathbb{E}_* \left[\frac{\log g_{\theta^*}}{\log g_\theta} \right], \quad H_R^* = \inf_{\theta \in \Theta_R} H(\theta), \tag{3.27}$$

with g_θ the mixtures likelihood. Then, under the assumptions of Theorem 3.1, we have with probability at least $1 - 2\exp(-n\delta^2/8M)$,

$$\mathbb{B}_n(R) \geq \frac{\pi(R)\pi_R(S_R(\delta))}{2} e^{-n(H_R^* + \delta)}. \tag{3.28}$$

Proof. We focus on the key differences between our setting and that of [23], who consider mixture models whose emission densities f_γ are parameterised by $\gamma \in \Gamma$. In their paper, the assumptions of the lemma are verified by assuming that there exists $\alpha > 0$ such that, for all $\gamma^{(1)}, \gamma^{(2)} \in \Gamma$, with $\gamma^{(1)} \neq \gamma^{(2)}$,

$$\sup_{\gamma \in \Gamma} \mathbb{E}_{x \sim f_{\gamma^{(1)}}} \left[\left(\frac{f_{\gamma^{(2)}}(x)}{f_\gamma(x)} \right)^\alpha \right] < \infty. \tag{3.29}$$

Here f_γ is the emission density corresponding to the parameter $\gamma = (\gamma_1, \ldots, \gamma_M) \in \Gamma$.

In our context, the parameters γ correspond to the weights $\omega = \omega_{r,c,m}$, where r is fixed and c, m vary, and Γ corresponds to the set of such parameters. This assumption (3.29) then does not hold, since we may always choose ω such that $\omega_{r,c,m} = 0$ for some m with $\omega^{(1)}_{r,c,m} > 0$, to see the supremum is not finite.

The assumption (3.29) is sufficient, but not necessary, and indeed it suffices to verify directly that, for the likelihood g_θ and some $\alpha > 0$,

$$\varphi(t|\theta) := \mathbb{E}_* \left[\left(\frac{\log g_{\theta^*}}{\log g_\theta} \right)^t \right] \tag{3.30}$$

is infinitely differentiable over $t \in [0, \alpha]$, with bounded second derivative in t, for any fixed $\delta_0 > 0$ and $\theta \in S_R(\delta)$, with

$$S_R(\delta) = \left\{ \theta \in \Theta_R : H(\theta) \leq H_R^* + \frac{\delta}{2} \right\}, \quad \delta \in (0, \delta_0). \tag{3.31}$$

It is easy to see that $0 < H_R^* < \infty$ and so, in order for $H(\theta) < H_R^* + \frac{\delta}{2}$, θ must be such that g_θ is everywhere bounded away from zero, since g_{θ^*} is atomic. Hence, there exists

$c^*(\delta_0)$ for which

$$S_R(\delta) \subset S_R(\delta_0) \subset \{\theta \in \Theta_R : g_\theta > c^*(\delta_0)\}. \tag{3.32}$$

Then $\phi(t|\theta)$, as a finite sum of terms with numerator bounded from above and denominator bounded from below, is smooth with bounded second derivative for all $t \in [0, \alpha]$ whenever $\theta \in \{\theta \in \Theta_R : g_\theta > c^*(\delta_0)\}$. We write

$$M = \sup_{\delta \in (0,\delta_0)} \sup_{\theta \in S_R(\delta)} \sup_{t \in [0,\alpha]} \phi''(t|\theta) < \infty.$$

The bound which precedes then replaces Assumption **A1** in the proof of Lemma 2 of [23], and we reach their same conclusion. □

Proof of Theorem 3.1. The proof of Theorem 3.1 is split into an underestimation step, which is an adaptation of Theorem 1 of [23], and an overestimation step, which is a consequence of elementary arguments, due to the penalisation.

Underestimation step: We first show $\mathbb{P}_*(\hat{R}_n = R) \to 0$ for $R < R^*$. Elementary calculation shows that for any X, Y real-valued random variables and $a \in \mathbb{R}$,

$$\mathbb{P}(X < Y) \leq \mathbb{P}(X < a) + \mathbb{P}(Y > a). \tag{3.33}$$

We have that $\{\hat{R}_n = R\} \subset \{\log \mathbb{B}_n(R+1) \leq \log \mathbb{B}_n(R) + \text{pen}(R)\}$ and so it suffices to bound, for a to be chosen later,

$$\begin{aligned} \mathbb{P}_*\{\log \mathbb{B}_n(R+1) - \text{pen}(R) \leq \log \mathbb{B}_n(R)\} \leq \mathbb{P}_*(\log \mathbb{B}_n(R) > a) \\ + \mathbb{P}_*(\log \mathbb{B}_n(R+1) - \text{pen}(R) < a). \end{aligned} \tag{3.34}$$

From Lemma 3.6 we have that, for $R = 1, \ldots, R^* - 1$,

$$\mathbb{B}_n(R+1) \geq \frac{\pi(R+1)\pi_{R+1}(S_{R+1}(\delta))}{2} e^{-n(H_{R+1}+\delta)} \tag{3.35}$$

with probability at least $1 - 2\exp\{-\frac{n\delta^2}{8M}\}$. To control the second term in (3.34), we wish to show that, with high probability,

$$\log \mathbb{B}_n(R+1) - \text{pen}(R) \geq \log \pi(R+1) + \log(\pi_{R+1}(S_{R+1}(\delta))) - \log(2) - n(H^*_{R+1} + \delta) \tag{3.36}$$

If $\text{pen}(R) = o(n)$, then for $n > n_0$ for n_0 large enough, $\text{pen}(R) < \frac{\delta}{2}n$ and so (3.36) holds

as soon as

$$\log \mathbb{B}_n(R+1) \geq \log \pi(R+1) + \log(\pi_{R+1}(S_{R+1}(\delta))) - \log(2) - n\left(H^*_{R+1} + \frac{\delta}{2}\right), \quad (3.37)$$

and hence (3.36) holds with probability at least $1 - 2\exp\{-\frac{n\delta^2}{32M}\}$. Choosing a equal to the right hand side of (3.36), the proof of Theorem 1 in [23] controls the remaining term of (3.35) and the proof for the underestimation is complete.

Overestimation step We now show that $\mathbb{P}^*(\hat{R}_n > R^*) \to 0$. First note that $\{\hat{R}_n > R^*\} \subset \{\log \mathbb{B}_n(R^*+1) \geq \log \mathbb{B}_n(R^*) + \text{pen}(R^*)$ and so it suffices to bound

$$\begin{aligned}\mathbb{P}_*\{\log \mathbb{B}_n(R^*+1) \geq \log \mathbb{B}_n(R^*) + \text{pen}(R^*)\} \leq \mathbb{P}_*(\log \mathbb{B}_n(R^*) < a) \\ + \mathbb{P}_*(\log \mathbb{B}_n(R^*+1) - \text{pen}(R^*) > a)\end{aligned} \quad (3.38)$$

The first term is bounded when we choose $a = -\frac{D(R^*)}{2}\log n + C$, by standard arguments involving a Taylor expansion of the log likelihood to second order, which are valid as we consider in this first term the behaviour of the well-specified evidence $\mathbb{B}_n(R^*)$. For the second term, we apply Markov's inequality

$$\begin{aligned}&\mathbb{P}_*(\log \mathbb{B}_n(R^*+1) - \text{pen}(R^*) > -\frac{D(R^*)}{2}\log n + C) \\ &\mathbb{P}_*\left(\mathbb{B}_n(R^*+1) > \exp\{-\frac{D(R^*)}{2}\log n + C + \text{pen}(R^*)\}\right) \\ &\leq \exp\left\{\frac{D(R^*)}{2}\log n - C - \text{pen}(R^*)\right\} \underbrace{\mathbb{E}_*\left[\mathbb{B}_n(R^*+1)\right]}_{=\pi(R^*)}\end{aligned} \quad (3.39)$$

which vanishes in the limit $n \to \infty$, under the condition that $\text{pen}(R^*) \geq \rho \log n$ for $\rho > \frac{D(R^*)}{2}$ and n sufficiently large. □

3.6.2 Proof of Proposition 3.4

Proof of Proposition 3.4. To exhibit such a sequence, it will suffice to restrict to parameters $\theta_j \in \Theta_{R^*+1}$, of the form $\theta_j = (p^j, \underline{\omega}^j)$ with $\underline{\omega} = (\omega_1, \omega_2, \omega_3)$ such that $\omega_1 = \omega_2 = \omega_3$. These correspond to mixtures where the three dimensions are conditionally iid. We assume θ^* takes this form also. These restrictions are not fundamental to the argument, but rather simplify notation. We then abuse notation slightly by replacing the expressions for $d_j, \bar{d}_j$ in the statement (which had dimension $3M$), with versions corresponding to one co-ordinate (thus of dimension M); the lower bound we consider here thus differs in the

constants from that of the statement, but is nevertheless of the same order as we take $j \to \infty$.

For such models, we wish to exhibit a specific choice of θ_j for which

$$\lim_{j\to\infty} \frac{\|g_{\theta_j}^{(R^*+1)} - g_{\theta^*}^{(R^*)}\|}{L(\theta_j, \theta^*)} = 0. \tag{3.40}$$

For a generic parameter, we write $\omega_r = (\omega_{c,m,r})_{m=1}^M \in [0,1]^M$, since by assumption the weights do not depend on the dimension index c. Then the numerator in Equation (3.40), when using the L^1 norm[3], can be written as

$$\begin{aligned}\|g_{\theta_j}^{(R^*+1)} - g_{\theta^*}^{(R^*)}\|_1 &= \sum_{\underline{m}\in[M]^3} |\sum_{r=1}^{R^*+1} p_r^j \omega_{r,m_1}^j \omega_{r,m_2}^j \omega_{r,m_3}^j - \sum_{r=1}^{R^*} p_r^* \omega_{r,m_1}^* \omega_{r,m_2}^* \omega_{r,m_3}^*| \\ &= \underbrace{\|\sum_{r=1}^{R^*+1} p_r^j \omega_r^j \otimes \omega_r^j \otimes \omega_r^j - \sum_{r=1}^{R^*} p_r^* \omega_r^* \otimes \omega_r^* \otimes \omega_r^*\|_1}_{\tilde{T}_j}. \end{aligned} \tag{3.41}$$

By choosing $p_r^j = p_r^*$, and $\omega_r^j = \omega_r^*$, for $r \leq R^* - 2$, we restrict to the case $R^* = 2$. Let $p_1^j \to p_1^*$, so that $p_2^j + p_3^j \to p_2^*$, and write $\bar{p}_1^j = p_1^j$, $\bar{p}_2^j = p_2^j + p_3^j$. Write $q_2^j = p_2^j/\bar{p}_2^j$ and $q_3^j = p_3^j/\bar{p}_3^j$ so $q_2^j + q_3^j = 1$. Write $\bar{d}_1^j = \omega_1^j - \omega_1^*$, $d_2^j = \omega_2^j - \omega_2^*$, $d_3^j = \omega_3^j - \omega_2^*$, and $\bar{d}_2^j = q_1^j d_2^j + q_2^j d_3^j$. Now define

$$\begin{aligned} T_j &= \sum_{r=1}^{2} (\bar{p}_r^j - p_r^*) \omega_r^* \otimes \omega_r^* \otimes \omega_r^* \\ &\quad + \sum_{r=1}^{2} p_r^* (\bar{d}_r^j \otimes \omega_r^* \otimes \omega_r^* + \omega_r^* \otimes \bar{d}_r^j \otimes \omega_r^* + \omega_r^* \otimes \omega_r^* \otimes \bar{d}_r^j) \\ &+ \bar{p}_2^j \Big\{ q_2^j (d_2^j \otimes d_2^j \otimes \omega_2^* + d_2^j \otimes \omega_2^* \otimes d_2^j + \omega_2^* \otimes d_2^j \otimes d_2^j) \\ &\quad + q_3^j (d_3^j \otimes d_3^j \otimes \omega_2^* + d_3^j \otimes \omega_2^* \otimes d_3^j + \omega_2^* \otimes d_3^j \otimes d_3^j) \Big\}. \end{aligned} \tag{3.42}$$

Given this choice of parameters, our lower bound $L_j = L(\theta_j, \theta^*)$ simplifies (up to constants) to

$$L_j = \sum_{r=1}^{2} |\bar{p}_r^j - p_r^*| + \sum_{r=1}^{2} \|\bar{d}_r^j\| + \left(q_2 \|d_2^j\|^2 + q_3 \|d_3^j\|^2 \right). \tag{3.43}$$

[3]Since the emissions are multinomial probabilities, the marginal densities $g_\theta^{(R)}$ are finite dimensional object and all norms are equivlent.

Then $\tilde{T}_j = T_j + o(L_j)$ as $j \to \infty$, so it suffices to exhibit a sequence for which $\lim_j T_j/L_j = 0$. Next, write

$$\bar{\beta}_r^j = \frac{\bar{d}_r^j}{\|\bar{d}_r^j\|} \quad (r \in \{1,2\}), \quad \beta_2^j = \frac{d_2^j}{\|d_2^j\|}, \beta_3^j = \frac{d_3^j}{\|d_3^j\|}, \tag{3.44}$$

and write also

$$\alpha_r^j = \frac{\bar{p}_r^j - p_r^*}{L_j}, s_r^j = \frac{\|\bar{d}_r^j\|}{L_j} \quad (r \in \{1,2\}), \tag{3.45}$$

and

$$t_2^j = \frac{\|d_2^j\|^2}{L_j}, \quad t_3^j = \frac{\|d_3^j\|^2}{L_j}. \tag{3.46}$$

Every sequence of parameters p^j, ω^j corresponds to a sequence of the above quantities, which we detail in what follows. By compactness arguments, they have limits along a subsequence which we pass to, and we denote their limits by dropping the superscript j. For the quotient T_j/L_j to have limit zero, we must have

$$\begin{aligned} 0 &= \sum_{r=1}^{2} \alpha_r \omega_r^* \otimes \omega_r^* \otimes \omega_r^* \\ &+ \sum_{r=1}^{2} p_r^* s_r (\bar{\beta}_r \otimes \omega_r^* \otimes \omega_r^* + \omega_r^* \otimes \bar{\beta}_r \otimes \omega_r^* + \omega_r^* \otimes \omega_r^* \otimes \bar{\beta}_r) \\ &+ \bar{p}_2 (q_2 t_2 + q_3 t_3)(\bar{\beta}_2 \otimes \bar{\beta}_2 \otimes \omega_2^* + \bar{\beta}_2 \otimes \omega_2^* \otimes \bar{\beta}_2 + \omega_2^* \otimes \bar{\beta}_2 \otimes \bar{\beta}_2) \end{aligned} \tag{3.47}$$

We will show that it is possible to choose θ_j such that

$$\bar{\beta}_1 = \bar{\beta}_2 = \beta_2 = \beta_3 = \frac{\omega_1^* - \omega_2^*}{\|\omega_1^* - \omega_2^*\|}. \tag{3.48}$$

Set $N = \|\omega_1^* - \omega_2^*\|$, and note also by their definitions that $\alpha_2 = -\alpha_1$. Then (3.47) holds if and only if

$$\begin{aligned} 0 &= \alpha_1 (\omega_1^* \otimes \omega_1^* \otimes \omega_1^* - \omega_2^* \otimes \omega_2^* \otimes \omega_2^*) \\ &+ \frac{p_1^* s_1}{N} \{(\omega_1^* - \omega_2^*) \otimes \omega_1^* \otimes \omega_1^* + \omega_1^* \otimes (\omega_1^* - \omega_2^*) \otimes \omega_1^* + \omega_1^* \otimes \omega_1^* \otimes (\omega_1^* - \omega_2^*)\} \\ &+ \frac{p_2^* s_2}{N} \{(\omega_1^* - \omega_2^*) \otimes \omega_2^* \otimes \omega_2^* + \omega_2^* \otimes (\omega_1^* - \omega_2^*) \otimes \omega_2^* + \omega_2^* \otimes \omega_2^* \otimes (\omega_1^* - \omega_2^*)\} \\ &+ \frac{\bar{p}_2 (q_2 t_2 + q_3 t_3)}{N^2} \Big\{ \omega_2^* \otimes (\omega_1^* - \omega_2^*) \otimes (\omega_1^* - \omega_2^*) \\ &\qquad + (\omega_1^* - \omega_2^*) \otimes \omega_2^* \otimes (\omega_1^* - \omega_2^*) + (\omega_1^* - \omega_2^*) \otimes (\omega_1^* - \omega_2^*) \otimes \omega_2^* \Big\}. \end{aligned} \tag{3.49}$$

Suppose we exhibit parameters for which the relations

$$\begin{aligned} \alpha_1 &= -3p_1^* s_1/N = -\alpha_2 \\ \frac{\bar{p}_2(q_2 t_2 + q_3 t_3)}{N^2} &= \frac{p_1 s_1}{N} = \frac{\alpha_2}{3}. \\ \frac{p_2^* s_2}{N} &= 2\frac{p_1^* s_1}{N} = \frac{2\alpha_2}{3} \end{aligned} \tag{3.50}$$

hold. Then, substituting the relations (3.50), we see that (3.49) holds.

It thus suffices to exhibit parameters for which the relationships of (3.50) hold, and for which we have the required relationship (3.48). Define

$$\begin{aligned} p_1^j &= p_1^* - \epsilon_j \\ \bar{p}_2^j &= p_2^* + \epsilon_j \\ \omega_1^j &= \omega_1^* + (\omega_1^* - \omega_2^*)\eta_{1,j} \\ \omega_2^j &= \omega_2^* + (\omega_1^* - \omega_2^*)\eta_{2,j} \\ \omega_3^j &= \omega_2^* + (\omega_1^* - \omega_2^*)\eta_{3,j}, \end{aligned} \tag{3.51}$$

for $\epsilon_j, \eta_{1,j}, \eta_{2,j}, \eta_{3,j} > 0$, and $q_j \in (0,1)$ with $p_2^j = \bar{p}_2^j q_j$, $p_3^j = \bar{p}_2^j(1-q_j)$. Since $(\omega_1^* - \omega_2^*)$ has zero sum, and since $p_1^j + \bar{p}_2^j = 1$, these choices determine a valid parameter θ_j whenever the $\eta_1^j, \eta_2^j, \eta_3^j$ are sufficiently small. Writing $N = \|\omega_1^* - \omega_2^*\|$ as before, we see that these choices lead to $\beta_1^j = \beta_2^j = \bar{\beta}_2^j = \bar{\beta}_3^j = \dfrac{\omega_1^* - \omega_2^*}{N}$, for all j, and hence in the limit as we require. Finally, we have that

$$L_j = 2\epsilon_j + N\left(\eta_{1,j} + q_j\eta_{2,j} + (1-q_j)\eta_{3,j}\right) + N^2\left(q_j\eta_{2,j}^2 + (1-q_j)\eta_{3,j}^2\right). \tag{3.52}$$

This then implies expressions for $\alpha_1^j, \alpha_2^j, s_1^j, s_2^j, t_2^j, t_3^j$ in terms of the $\epsilon_j, \eta_{1,j}, \eta_{2,j}, \eta_{3,j}, q_j$. By choosing suitable values of the $\epsilon_j, \eta_{1,j}, \eta_{2,j}, \eta_{3,j}, q_j$, as required by the discussion which precedes, we conclude that $\lim_j T_j/L_j = 0$ as required.

□

Acknowledgements

The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the

CDT StatML.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

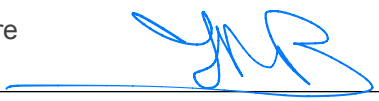
Title of Paper	Bayesian Model Selection in Semiparametric Mixtures and Hidden Markov Models
Publication Status	□Published □ Accepted for Publication □Submitted for Publication ☑Unpublished and unsubmitted work written in a manuscript style
Publication Details	Unpublished work written in a manuscript style. Produced in collaboration with Prof. Judith Rousseau (University of Oxford).

Student Confirmation

Student Name:	Daniel Moss		
Contribution to the Paper	I contributed to the development of theory for model selection results, in close collaboration Prof. Rousseau. It is difficult to separate the contributions each of us had in these parts. I personally wrote code for, and conducted, corresponding numerical experiments, with the directions of these experiments guided by discussions with Prof. Rousseau. I conducted most of the manuscript writing myself, again under guidance from Prof. Rousseau.		
Signature	D.P. Moss.	Date	04/01/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Judith Rousseau		
Supervisor comments			
Signature		Date	04/01/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

4. Bayesian Analysis of Conflict Risk with Multivariate Discrete-time Hawkes Processes

4.1 Introduction

Statistical modelling of the occurrence of conflict events dates back to [90], who demonstrated the power-law nature of the scale of conflicts. More recently, [25] reexamined this notion in the context of terrorism. In [15], nonlinear spatial models were employed to understand the geographic contagion of civil unrest.

Understanding such events is of key importance for both the short-term assessment of risk, and the identification of longer-term trends. Such information is of interest to various actors, including both governmental organisations and nonprofits, and can inform policy and crisis response. One such actor is the *Armed Conflict Location & Event Data Project* (ACLED), who curate a global collection of conflict data and produce corresponding analysis, with both aspects being open to the public. The data is provided at the daily level, with details about both the type of event (e.g. Battles, Riots) and the location (e.g. administrative district), among various other features. The rich nature of this data, and the dependencies therein, motivates the use of statistical tools capable of modelling the evolution of conflict over space and time.

In this work, we focus on the use of Hawkes processes as a means to model such datasets. First introduced in [53,54], Hawkes processes are self-exciting point processes, in which the intensity at a given time varies according to the events which precede it; for a review paper see [65]. As such, they have enjoyed use in a variety of contexts in which activity tends to evoke reaction, including seismology [83], finance [55], neurology [13, 46], epidemiology [17, 24], and social harm [81].

Univariate, discrete-time, Hawkes processes were proposed as means to model conflict in [16], who also incorporated a spatial triggering kernel in conjunction with the tempo-

ral kernel to identify spatiotemporal relationships in conflict events. The focus of their application was at the intra-national level, in particular looking at countries in southern Asia, and they too used data provided by ACLED. The use of *discrete-time* processes is motivated by the fact that data from ACLED arrives in daily summaries, and so lacks the temporal precision associated with a continuous process. Univariate, discretised approaches have been considered in a variety of other applications, see for instance [17, 80].

Our work seeks to complement and advance the methodology of [16], which was limited in its ability to capture spatial interactions in certain countries. Departing from their approach, which employed a spatial triggering kernel to incoporate spatial dependencies, we instead propose a Bayesian multivariate Hawkes process across geographical regions. Moreover, we employ a more flexible likelihood for the discrete counts given the intensity, with the discretised setting allowing us to move beyond conditionally Poisson likelihoods while still maintaining a coherent, computationally tractable model.

There is an established body of literature on multivariate Hawkes processes for continuous time, both on the theoretical and practical side, see for instance [31, 35, 67, 70, 95, 96]. In the discrete setting, [69] considered a discrete-time *network* Hawkes model which is similar to the model which we propose, although the priors we consider are closer to those of [31].

Despite the computational burden of the multivariate model, we demonstrate that inference may be straightforwardly parallelised, which is not only a computational benefit but also allows the practitioner to model in isolation the impact of a group of regions on a single region of interest, in a statistically coherent way. Moreover, the model is easily tuned by an expert with a priori knowledge of the interaction matrix between regions and types of event. As a case study, we examine conflict in Bangladesh; we first demonstrate our model's capabilities on simulated data, before applying our techniques to the ACLED dataset.

Structure of the chapter In Section 4.2, we summarise the background to the work, first discussing the analysis of ACLED, before reviewing the limitations imposed by the spatial kernel approach of [16] and motivating a multivariate approach. In Section 4.3, we introduce the multivariate methodology; we motivate also the need for more flexible likelihoods, and describe our choices of priors for the Bayesian modelling. In Section 4.4, we assess the performance of our model in a simulated setting, before applying the methodology to data provided by ACLED.

4.2 Existing methodology for modelling the ACLED dataset

In Section 4.2.1, we introduce the ACLED dataset in more depth and give an overview of their own analytical tools. In Section 4.2.2, we introduce discrete time Hawkes processes as used in [16], upon which the methodology of this paper will build.

4.2.1 Data and Analysis from ACLED

The Armed Conflict Location & Event Data Project (ACLED[1]) is a non-profit providing data and analysis for political violence and protest around the world. ACLED curates a dataset which details events falling into six categories or *event types*: Battles; Violence against civillians; Explosions/Remote violence; Riots; Protests; and Strategic Developments. The data is quite rich, and includes details about the event type (as well as various *event subtypes* which refine further the aforementioned categories) the date of the event, the location of the event (as indicated by longitude and latitude[2]), information about the actors involved in the event (e.g. paramilitary group name), the source of information around the event (e.g. newspaper name), number of fatalities, and further natural language description of details of the event.

Currently, ACLED display relatively crude measures of risk through their website across all event types on their dashboard[3], which simply aggregate counts for the different event types. There are also some more refined tools in the "Early Warning & Forecasting" section of their website[4], the most sophisticated of which is the ACLED **C**onflict **A**lert **S**ys**T**em (CAST). The CAST methodology is based on time series models, tuned by cross validation, which aggregate spatially at the Level 1 (coarsest subnational level) administrative region, and which provide point estimates for future event counts. CAST is also limited to predictions for three of the six event types. A full report describing this methodology is available at the ACLED website.

Besides CAST, ACLED provides various other more basic tools such as the "Volatility and Risk Predictablity" index which, as in CAST, spatially aggregates to the Level 1 ad-

[1] https://acleddata.com/

[2] The longitude and latitude is the most granular description (although is still a slight coarsening of the locations, seemingly due to rounding). The country, and three levels of administrative region within the country, are also given, with Level 1 being the highest level of aggregation (fewest regions) and Level 3 being the most refined aggregation (most regions) besides the actual longitude/latitude

[3] https://acleddata.com/dashboard

[4] https://acleddata.com/early-warning-research-hub/

ministrative region. The methodology is outlined in Table 4.1, which has been reproduced from ACLED's website[4].

Volatility/Baseline	Low baseline†	High baseline††
Low volatility*	Low risk	Consistent risk
High volatility**	Growing risk	Extreme risk

Table 4.1: Methodology for ACLED's Volatility & Risk Predictability Index.
* Fewer than 6 weeks out of the past year see violence spiking 2 or more standard deviations above the baseline for the given Level 1 administrative region. Baseline given by two-year weekly average.
** 6 or more weeks out of the past year, violence spiked 2 or more standard deviations above the baseline for the given Level 1 administrative region.
† Less than 1.5 events per week
†† More than 1.5 events per week

The work of [16] sought to strengthen the existing methodology by using (discretised) spatiotemporal Hawkes processes, aggregating spatially at the Level 2 administrative region (which is more finely grained than the Level 1 administrative region) and temporally by month. These were used to model all event types, and a Bayesian approach was taken and implemented in STAN [98]. This approach allowed for uncertainty quantification, finer grained spatial estimates, and risk assessment for all event types.

4.2.2 Discrete time Hawkes processes and spatial dependency

A *discrete time Hawkes process* is a data-generating process for a discrete time series $Z_1, Z_2, \dots$, with $Z_i \in \{1, 2, \dots\}$, which satisfies

$$Z_t | Z_1, Z_2, \dots Z_{t-1} \sim \text{Po}(\lambda_t) \tag{4.1}$$

where

$$\lambda_t = \mu_t + \alpha \sum_{t' < t} Z_{t'} g(t - t'). \tag{4.2}$$

Here μ_t is a deterministic rate at which events occur sporadically, α controls the average number of 'offspring' from an event (viewing Hawkes processes as a superposition of births caused by different events) and h is an excitation kernel with the property that $\sum_{i=1}^{\infty} g(i) = 1$, which ensures identifiability of α. Such models were considered in the context of the ACLED dataset in [16], with Z_t representing the number of events occuring in a fixed window of time.

The authors then extended their analysis to a *spatiotemporal discrete time Hawkes processes*, using a spatial kernel. Such a process models data $S_{t,x,y}$, for which we have a count

at location (x, y) which takes values in a discrete co-ordinate grid $\mathcal{C}$. The data generating process is this case is then

$$S_{t,x,y}|\{S_{t',x',y'}\}_{t'<t,(x,y)\in\mathcal{C}} \sim \text{Po}(\lambda_{t,x,y}) \tag{4.3}$$

$$\lambda_{t,x,y} = \mu_{t,x,y} + \alpha \sum_{t'<t} \sum_{(x',y')\in\mathcal{C}} S_{t,x,y} g(t-t') h(x-x', y-y') \tag{4.4}$$

with g as before and h a spatial excitation kernel. In the setting of [16], $\mathcal{C}$ contains the longitude and latitude of the centroids of Level 2 administrative regions. For the kernels g and h, [16] proposed a geometric decay in time given by

$$g_\beta(t) = \beta(1-\beta)^{t-1} \quad t \geq 1, \tag{4.5}$$

and a squared exponential decay in space given by

$$h_\sigma\left(\|(x,y)-(x',y')\|\right) = \exp\left(-\frac{\|(x,y)-(x',y')\|^2}{\sigma^2}\right). \tag{4.6}$$

It was found in [16] that the suitability of this approach was strongly dependent on the nature of the data being modelled, given the strict form of (4.6). In particular, the model was unable to capture meaningful spatial dependency in Bangladesh. Figures 4.1 and 4.2 show respectively the true distribution in the data, across second level administrative regions, of the aggregate events across a five-year period, and a simulated five-year dataset from a spatial kernel model of the form considered in [16].

This motivates an alternative approach to considering spatial dependency in this data, in order to capture structure which is not well expressed by a rigid choice of spatial kernel. In Section 4.3, we present our methodology for modelling such dependency, which we believe provides an alternative, complementary approach for those datasets which were not well-captured by the model structure of [16].

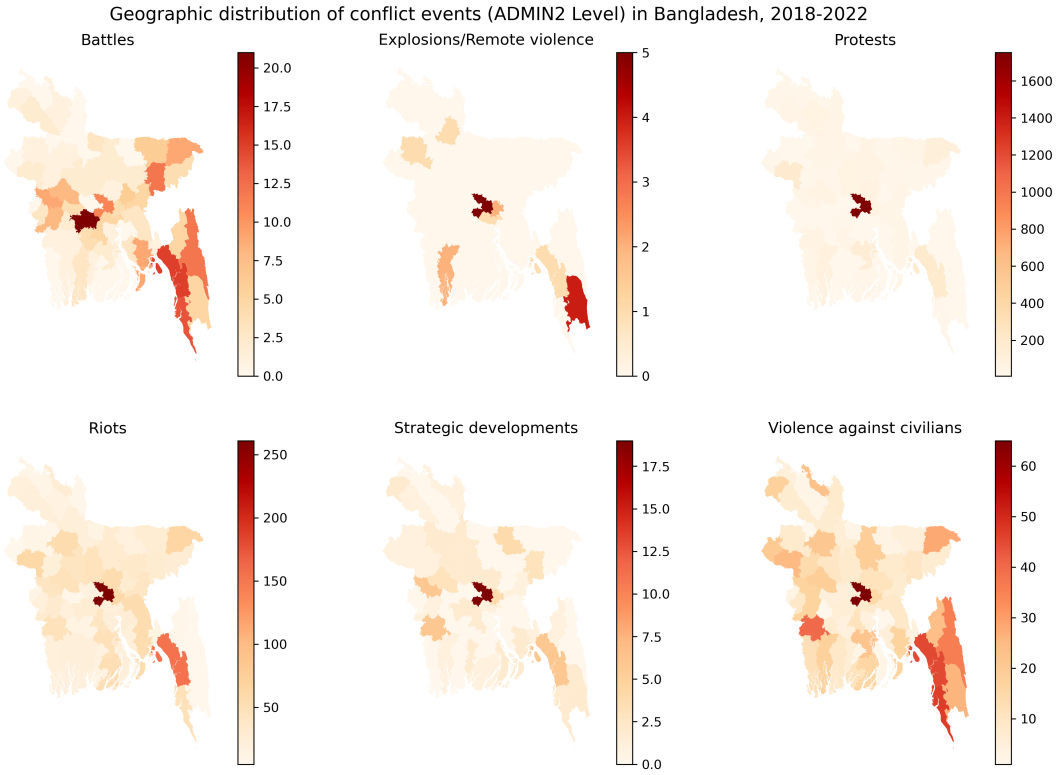


Figure 4.1: Real counts (totalled across 2018-2022) of conflict events of different types, by second-level administrative region of Bangladesh.

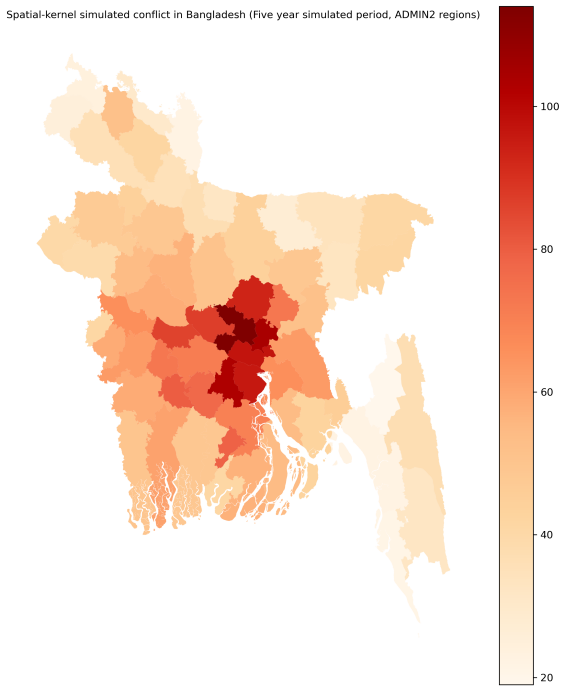


Figure 4.2: Simulated five-year data from a spatial kernel model, with squared exponential kernel (4.6). The simulated data has a characteristic shape owing to the kernel choice, which does not well capture the distributions of observed events shown in Figure 4.1

4.3 Structure of proposed models

In the following section, we aim to develop a flexible model capable of capturing arbitrary spatial dependencies. In Section 4.3.1, we introduce the multivariate approach to modelling spatial relation. In Section 4.3.1.1, we demonstrate how this model, and the models considered by [16], can be adapted to include covariates for added flexibility. In Section 4.3.2, we motivate and introduce more flexible likelihoods given the intensity λ_t. Finally in Section 4.3.3, we discuss the practical details of implementing these models with a Bayesian approach, and in particular we disucss our prior choices.

4.3.1 Multivariate model

In order to better express the dependency between spatial regions, we consider in our approach a *multivariate* discrete time Hawkes process. This involves modelling jointly the event counts $\{S_t^{(j)}\}_{t\leq T}^{j\leq k}$ of k geographic regions $C_1, C_2, \ldots, C_k$ by the relationship

$$S_t^{(j)} | \{S_{t'}^{(i)}\}_{t'<t}^{1\leq i\leq k} \sim \text{Po}(\lambda_t^{(i)}), \tag{4.7}$$

where the rates $\lambda_t^{(k)}$ are given as

$$\begin{aligned} \lambda_t^{(j)} &= \mu_j + \sum_{i=1}^{k} \lambda_{t,i}^{(j)}; \qquad\qquad 1 \leq j \leq k \\ \lambda_{t,i}^{(j)} &= \alpha_{ij} \sum_{t'<t} S_{t',i} g_{\beta_{ij}}(t-t'). \end{aligned} \tag{4.8}$$

We interpret the rate $\lambda_t^{(j)}$ as the intensity of events at time t in region j, with those events occurring sporadically at rate μ_j, and as a result of other events in region i at rate $\lambda_{ti}^{(j)}$. Although counts from ACLED are received daily, we consider counts on a weekly basis - there are a few motivations for doing so:

- Much of the analysis provided on ACLED's own website is at the weekly level (for instance, the Volatility & Risk Predictability Index described in Table 4.1 describes risk in terms of weekly counts).
- Data from ACLED is typically released on a weekly basis.
- Inference in a model with daily counts is challenging, both computationally due to the number of days, but also analytically due to the large number of days with zero counts.

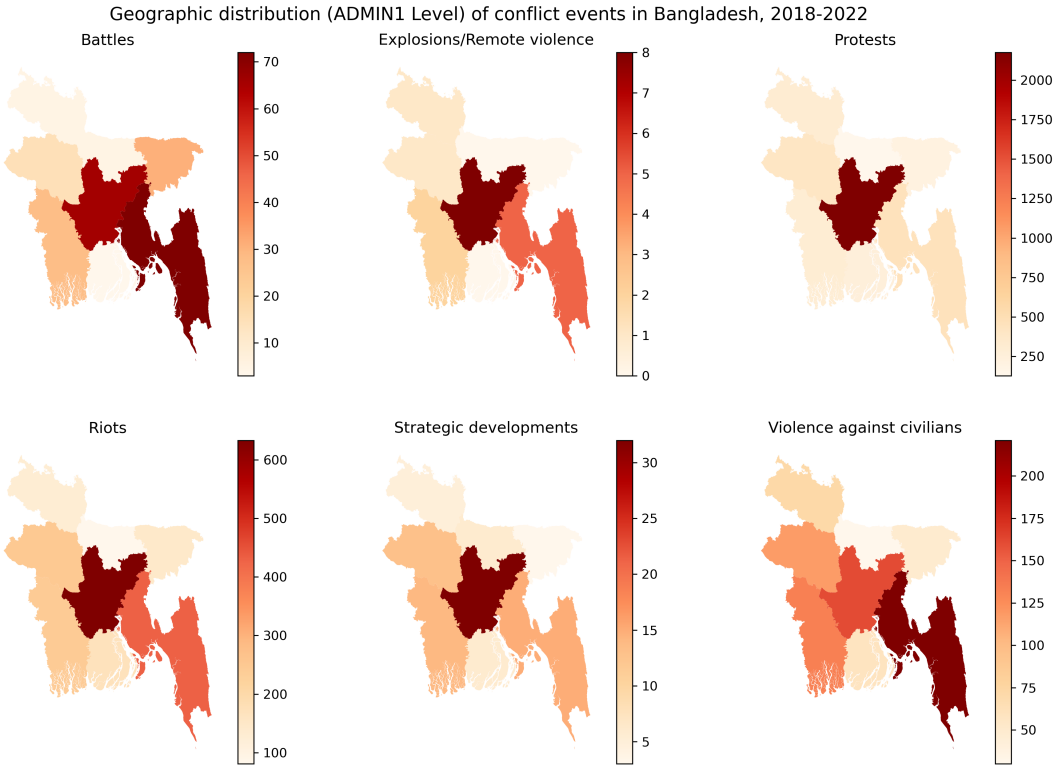


Figure 4.3: Real counts (totalled across 2018-2022) of conflict events of different types, by Level 1 administrative region of Bangladesh.

Even though we formulate this model primarily as a means to capture spatial interaction, it is not inherently spatial and can equally model interaction between events of different types (e.g. Riots and Battles) in a fixed spatial region. The parameters of the model are then

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1k} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kk} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kk} \end{pmatrix} \tag{4.9}$$

Such models allow us to express a much richer set of spatial interactions, and capture potentially complex spatial dependencies in the data which are not well-explained by geographic distance alone (even if a more flexible spatial kernel was chosen). In Figure 4.3, we make a plot of observed events by type, similar to Figure 4.1, but at the coarser (Level 1, rather than Level 2) administrative region spatial aggregation.

The price for this flexibility is that the parameter space is of signifcantly higher dimension than the model considered in [16]; the practioner has a number of options relating to this increased statistical complexity:

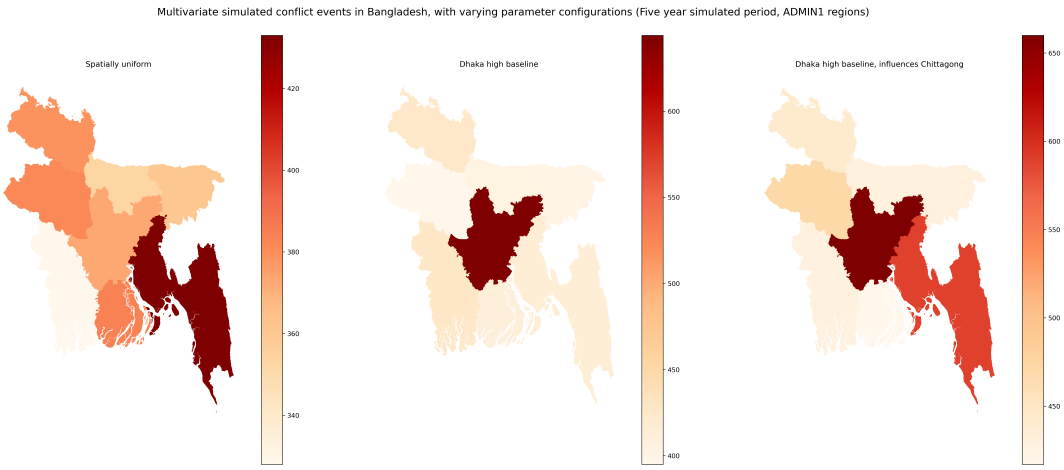


Figure 4.4: Simulated five-year data from multivariate models, using three distinct parameter choices. The flexibility of the multivariate model means that such varying parameter choices can lead to visually very different distributions, which are better-positioned to capture the true data of Figure 4.3. Left: Higher spatial uniformity, intended to capture a shape similar to the data for Violence against civillians. Middle: Primary focus on Dhaka, intended to capture a shape similar to the data for Protests. Right: Model in which Dhaka influences Chittagong, intended to capture a shape similar to the data for Explosions/Remote violence.

- Consider sparsity inducing priors on the interactions α_{ij}, similar to what was considered in [31] who proposed a Horseshoe prior on the α parameter in a similar model. Certain parameters can also be set exactly to zero a priori, if the signal is considered too weak to be captured. We discuss these approaches in Section 4.3.3

- Use a coarser spatial binning. The idea is that meaningful inferences aggregated over larger areas of space provide more insight than unreliable inferences over a more fine-grained spatial partition. In our approach, we consider the multivariate model for Bangladesh over only the first-level administrative regions (of which Bangladesh comprises eight, as shown in Figures 4.3 and 4.4), in contrast to the second-level administrative regions considered in [16] (of which Bangladesh comprises sixty-four).

On the computational side, the structure of the posterior in multivariate Hawkes processes means that this approach parallelises straightforwardly across spatial regions in the sense we describe in Section 4.3.3.1. With sufficient multithreading, the computational cost is then comparable to that of the spatial-RBF approach.

4.3.1.1 Inclusion of Covariates

A natural approach for reducing the complexity of the parameter space would be to account for the differences between the μ_j and α_{ij}, along the lines of what is done in [105] who

considered covariates in univariate Hawkes processes. In this case, we could model

$$\mu_j = f(z_j, \vartheta) \tag{4.10}$$

with f some regression function, z_i some covariate associated with the sporadic rate μ_i, and ϑ some parametrisation common to all μ. The disadvantage of such an approach computationally is that inference on the regions $C_1, \ldots, C_k$ can then not be done in parallel, as described in Section 4.3.3.1; on the statistical side it may also be challenging to select appropriate explanatory variables z_j.

We could also extend our model to include covariates, by replacing μ_j with

$$\mu_{j,t} = g(z_{j,t}, \vartheta_j) \tag{4.11}$$

where ϑ_j is a region-specific parameter for the global regression function g, and $z_{j,t}$ is a time-dependent covariate. We do not pursue this extension here, but may be of interest for instance if the practioner has a priori knowledge of seasonal or other exogenous effects.

4.3.2 Likelihood models

In Section 4.1, we introduced the discrete time Hawkes process as a necessary discretisation of the continuous time Hawkes process, owing to the daily binning of the data as provided by ACLED. One limitation of the discretised version is that events within the same discretised unit of time may not excite each other; in practice it is intuitive however that events on the same day, and certainly within the same week, may be associated to earlier events in that period, rather than strictly events in previous periods. As such, the observed distribution in each period of time is likely to have tails heavier than Poisson, owing to the possibility of a 'cluster' of events occurring within it. Since the Poisson distribution imposes that the probability p_0 of getting a zero is related to the mean λ by the relationship

$$\lambda = -\log p_0 \tag{4.12}$$

and so the heavy tailed phenomenon can from another perspective be viewed as a zero-inflation phenomenon, whereby the proportion $\hat{p}_0$ of counts in the data is excessively high compared to the mean count λ, in violation of Equation (4.12). In Figure, 4.5, we see two examples where the proportion of zeros in the data differs greatly from what would be achieved by a Poisson model, by matching simulated Poisson paths of varying μ and α

with the observed number of events.

In this section, we suggest various alternatives to the Poisson likelihood, which still respect the dependency structure of the discrete time Hawkes process while better capturing the heavy-tailed (or zero-inflated) characteristics of the data: the negative binomial likelihood, and zero-inflated versions of both Poisson and negative binomial. In each case we replace (4.7) with the described alternative.

Negative Binomial We consider first the negative binomial as an overdispersed version of the Poisson. The dispersion parameter varies between regions. We replace Equation (4.7) with

$$S_t^{(j)} | \{S_{t'}^{(j)}\}_{t'<t}^{1\leq j\leq k} \sim \text{NegBin}(\lambda_t^{(j)}, \phi^{(j)}), \tag{4.13}$$

where our parametrisation defines the probabilities as

$$\mathbb{P}_{\text{NegBin}}(S = s; \lambda, \phi) = \binom{s+\phi-1}{s} \left(\frac{\lambda}{\lambda+\phi}\right)^s \left(\frac{\phi}{\lambda+\phi}\right)^\phi,$$

as used in the alternative negative binomial parametrisation in STAN [98]. The mean and variance in this formulation then take the convenient form

$$\mathbb{E}_{\text{NegBin}}[S; \lambda, \phi] = \lambda; \quad \mathbb{V}_{\text{NegBin}}[S; \lambda, \phi] = \lambda + \frac{\lambda^2}{\phi}$$

so that λ retains its intepretation as the mean, while a lower value of ϕ indicates a greater overdispersion relative to the Poisson.

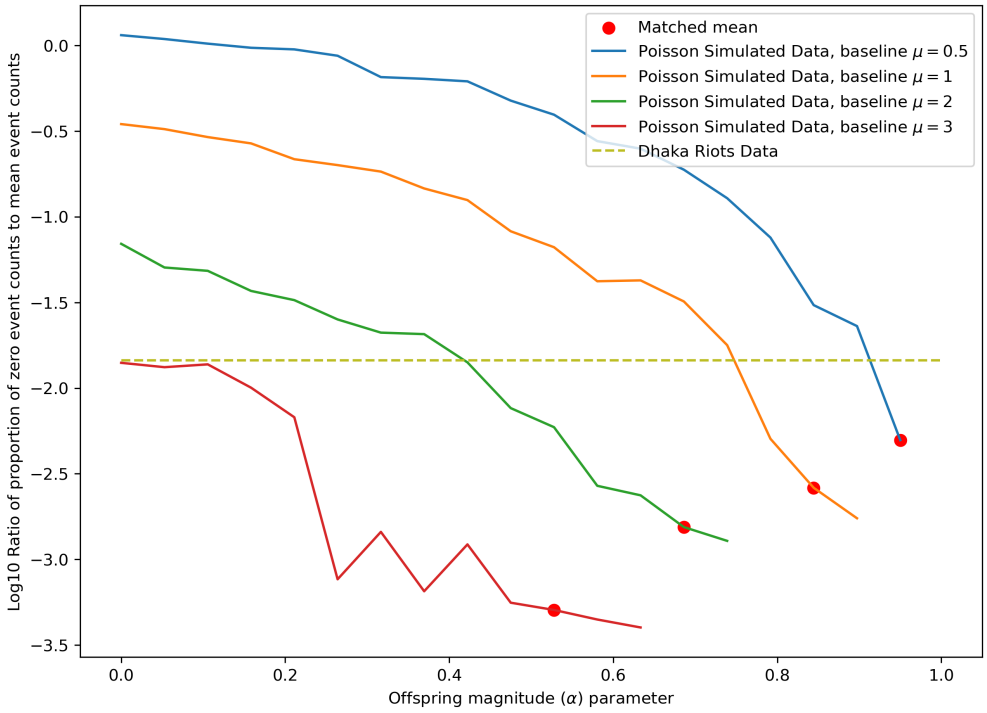
Poisson Hurdle Model The Poisson hurdle model introduces an additional parameter θ, which varies between regions, and replaces Equation (4.7) with

$$S_t^{(j)} | \{S_{t'}^{(i)}\}_{t'<t}^{1\leq i\leq k} \sim \text{PoHurdle}(\lambda_t^{(j)}, \theta^{(j)}), \tag{4.14}$$

where

$$\mathbb{P}_{\text{PoHurdle}}(S = s; \lambda, \phi) = \begin{cases} e^{-\lambda\theta} & : s = 0 \\ (1 - e^{-\lambda\theta}) \mathbb{P}_{\text{Po}}^{(\lambda)}(S = s | S \geq 1) & : s \geq 1 \end{cases}.$$

We thus recover the Poisson in the special case $\theta = 1$ but generally allow for an adjusted weight at zero, with the distribution of mass on $\{1, 2, \dots\}$ behaving like a $\text{Po}(\lambda)$ on the event $\{S > 0\}$. This departs from the typical formulation of the Poisson hurdle model in



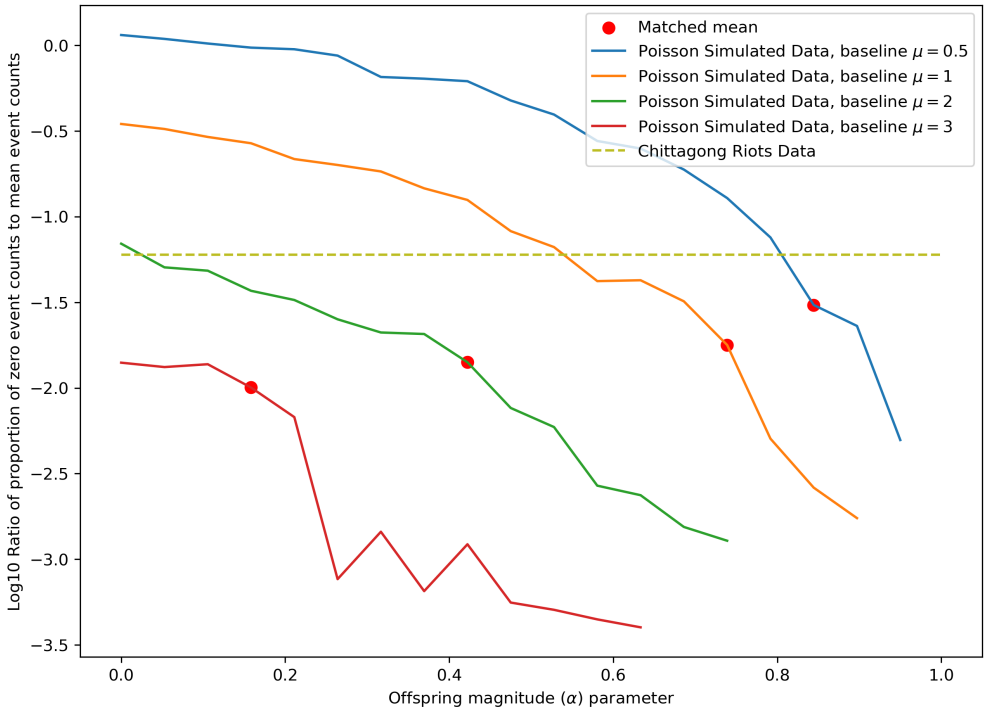


Figure 4.5: Proportion of zeros relative to mean, across various baseline rates μ and varying α. In red, we choose the α for which the simulated number of events most closely matches the region of interest (Dhaka above, Chittagong below), approximating a best fit for the given μ. The relative zero proportion observed in the actual data (2010-2015, Riots) is plotted in the horizontal dashed line. For all μ, Poisson simulated paths have too small a proportion of zero counts, and can differ by an order of magnitude.

its use of θ in the exponent - this is done so we can still have our zero weight be inflated *relative to the particular* λ which is varying between different data points. The associated mean is

$$\mathbb{E}_{\text{PoHurdle}}[S;\lambda,\theta] = \frac{(1-e^{-\lambda\theta})\lambda}{1-e^{-\lambda}}, \tag{4.15}$$

and so the interpretation of λ changes slightly relative to the parameter θ.

Negative Binomial Hurdle We take a similar approach as with the Poisson hurdle, but include the overdispersion ϕ as an additional parameter as in the negative binomial case. As previously both parameters vary by region, and we replace Equation (4.7) with

$$S_t^{(j)} | \{S_{t'}^{(i)}\}_{t'<t}^{1\leq i\leq k} \sim \text{NBHurdle}(\lambda_t^{(j)}, \phi^{(j)}, \theta^{(j)}), \tag{4.16}$$

where

$$\mathbb{P}_{\text{NBHurdle}}(S=s;\lambda,\phi,\theta) = \begin{cases} \left(\frac{\phi}{\lambda\theta+\phi}\right)^{\phi} & : s=0 \\ \left(1-\left(\frac{\phi}{\lambda\theta+\phi}\right)^{\phi}\right)\mathbb{P}_{\text{NegBin}}^{(\lambda,\phi)}(S=s|S\geq 1) & : s\geq 1 \end{cases}.$$

The associated mean is

$$\mathbb{E}_{\text{NBHurdle}}[S;\lambda,\phi,\theta] = \frac{1-(\phi/(\lambda\theta+\phi))^{\phi}}{1-(\phi/(\lambda+\phi))^{\phi}}\lambda, \tag{4.17}$$

again adjusting the interpetation of λ relative to θ, ϕ.

4.3.3 Implementation of Bayesian Approach

4.3.3.1 Posterior independence of regions

For the multivariate model, we can make the following observation about the likelihood for the model of Section 4.3.1. Set $S_t = \{S_{t'}^{(j)}\}_{t'\leq t}^{j\leq k}$ the collection of event data up to time t inclusive. Then the likelihood, possibly with additional parameters $\phi = (\phi^{(1)}, \ldots, \phi^{(k)})$ and $\theta = (\theta^{(1)}, \ldots, \theta^{(k)})$, satisfies

$$p(S_T|\mu,\alpha,\beta,\phi,\theta) = \prod_{t=1}^{T} p(S_t|S_{t-1},\mu,\alpha,\beta,\phi,\theta) = \prod_{t=1}^{T}\prod_{i=1}^{k} p(S_t^{(i)}|S_{t-1},\mu_i,\alpha_i,\beta_i,\phi^{(i)},\theta^{(i)}) \tag{4.18}$$

where α_i, β_i and σ_i denote the i^{th} *column* of the respective matrix. Thus, provided we choose a prior of the form

$$\Pi(\mu, \alpha, \beta, \sigma) = \prod_{i=1}^{k} \Pi_i(\mu_i, \alpha_i, \beta_i, \sigma_i), \tag{4.19}$$

we will have independence of $(\mu_i, \alpha_i, \beta_i, \phi^{(i)}, \theta^{(i)})$ and $(\mu_j, \alpha_j, \beta_j, \phi^{(j)}, \theta^{(j)})$ for $i \neq j$ in the posterior. In particular this is satisfied for any product prior, and also the prior we consider in Section 4.3.3.2.

Alongside the previously discussed computational advantage, this statistically principled ability to parallelise our inference also means that parameters measuring the impact of a large group of regions on a smaller group of regions of interest can be learnt, without the need to learn the model across all the regions involved.

4.3.3.2 Prior Choice

We detail in this section our prior choices and possible modifications thereof, which respect the posterior independence discussed in Section 4.3.3.1

Prior on sporadic rates μ_j The prior on μ_j is in some sense the most sensitive to the scale of our data, as the order of magnitude ought to match the order of magnitude of the observed data. To ensure the prior for μ_j is supported over a reasonable range of values, we propose an empirical Bayes Gamma prior with a data-driven choice of parameters to help weakly guide the choice of prior mean. For the parametrisation with density

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{4.20}$$

the mean is α/β and the variance α/β^2. We construct a prior by taking α a constant (we used $\alpha = 2$) and choosing β such that the mean is half of the average weekly event count. The standard deviation is then of the same order as the mean, and the precise choice of α controls the precise scaling.

Should the practitioner prefer to avoid an empirical Bayes approach, it would be valid to inform one's prior using data from other time periods, or geographies, for which the prior belief is that the order of magnitude should be roughly the same.

Prior on temporal excitation parameter β_{ij} Since we have $\beta_{ij} \in [0, 1]$, we adopt the uniform prior $\beta_{ij} \sim U[0, 1]$. A practitioner seeking a truly subjective prior might prefer the density to vanish at zero and one, however we believe this to be a reasonable objective approach and found our posteriors to be well-supported away from this boundary; this prior was also employed by [16].

As a means of simplifying the parameter space, we further constrain $\beta_{ij} = \beta_{i'j}$ for all i, i'. The rationale for this choice is that the parameters α_{ij} may be considered of greater individual interest, given that our modification of the univariate, or spatial kernel, approach of [16] is primarily intended to better understand the *magnitude* of spatial interactions. Since the spatial kernel approach scales only the magnitude of interactions between regions without affecting the temporal kernel, it is natural that we focus primarily on these parameters. However, constraining β_{ij} to be equal for all i, j would be in violation of the form of Equation (4.18).

Of course, if the practitioner has reason to believe that the β_{ij} differ signifcantly when varying i for fixed j, then the model is more strongly misspecified which could lead to erroneous inference for the corresponding α_{ij}. In this case, the prior can be straightforwardly adjusted to allow disagreement between these β_{ij}, or agreement for particular indices.

Prior on offspring magnitude α_{ij} In [31], the authors successfully used a Horseshoe prior on the α_{ij} parameters to introduce additional regularisation, given a prior belief that several interactions are likely to be small. Since their model allowed for $\alpha_{ij} < 0$ (representing inhibition), we adapt their choice slightly to a 'one-sided' horseshoe prior, which is constrained to take positive values.

As an additional means of reducing model complexity, we also consider in our modelling the a priori assumption that there is no (meaningful) impact of one region on another, that is to say $\alpha_{ij} = 0$. In this way, we can significantly reduce the number of parameters we estimate, leading to improved inference on the parameters which we believe to have a significant effect.

Prior on remaining parameters For the likelihood parameters $\theta^{(j)}, \phi^{(j)}$, we chose independent standard half-normal priors, being the law of $|Z|$ when $Z \sim N(0, 1)$. These were chosen as uninformative priors, and we would expect them to be well supported on a reasonable range of values for these parameters. In particular, the probability that ϕ is very small does not vanish too quickly, allowing for possibly large levels of overdispersion.

4.4 Empirical Analysis and Model Application

4.4.1 Simulated Data

To understand the ability of our Bayesian model to capture the ground truth and correctly quantify uncertainty, we simulate data from a fixed ground truth. The data is meant to mirror the region-by-region averages found in the Bangladesh data (specifically, weekly counts for Riots between 2010 and 2015), in order to understand what is feasible given a certain amount of data, although the actual parameters and associated structure may be very different.

The purpose of our simulation study is to understand the following aspects of the model

- The efficacy and influence of our prior choices, in particular the behaviour of the empirical Bayes prior on μ_j, and the Horseshoe-type prior on $(\alpha_{ij})_{i=1}^{k}$
- The potential for improved inference by setting *a priori* $\alpha_{ij} = 0$ for particular regions C_i thought not to affect C_j in any meaningful way.
- The ability of the Bayesian model to describe uncertainty in future events, in particular the coverage of one-step ahead posterior predictive bands, where given the events up to time t and the posterior distribution, we examine the implied distribution over the events at time $t + 1$.

Simulation details We considered eight simulated regions $C_1, C_2, \ldots, C_8$; these were identified respectively with the first-level administrative regions of Bangladesh, Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur, and Sylhet. We then simulated our data from the following underlying parameter configurations:

$$\mu = \begin{pmatrix} 0.15 \\ 0.75 \\ 1.5 \\ 0.1 \\ 0.1 \\ 0.2 \\ 0.15 \\ 0.25 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.3 & 0.2 & 0 & 0.2 & 0 & 0 \\ 0.15 & 0.3 & 0.6 & 0.2 & 0.05 & 0.2 & 0.15 & 0.1 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 \end{pmatrix} \tag{4.21}$$

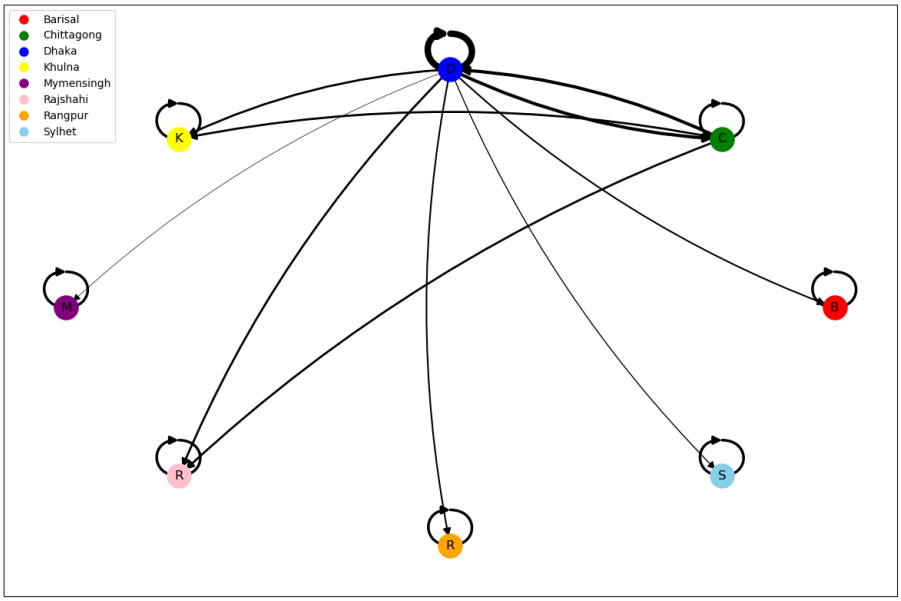


Figure 4.6: Graph of interactions between regions, as described by ground truth parameters of simulations. Thickness of line corresponds to strength of interaction.

and $\beta_{ij} = 0.7$ for all i, j. For instance, $\mu_3 = 1.5$ corresponds to a sporadic rate of 1.5 riots per week in C_3 (Dhaka). We can interpret the α matrix as a directed graph as in Figure 4.6; when inferring α we are seeking to recover the structure of this graph.

Ten datasets were simulated and then fitted to from this ground truth; each dataset was simulated for a time period equivalent to weekly counts between 2010 and 2015. For each simulated dataset, we considered modelling the impact on two of the regions: C_3 (Dhaka), which had the highest sporadic rate and was impacted only by Dhaka and Chittagong in the ground truth; and C_6 (Rajshahi), which has a relatively low sporadic rate relative to the amount of activity in the region, with many of its events being downstream of excitement from C_2 (Chittagong) and C_3 (Dhaka).

In each of these cases, we then fit two models: a saturated model, in which all regions were considered to have a possible impact, and a 'pre-trimmed' model, intended to mimic expert input who can understand *a priori* which edges of the graph in Figure 4.6 would likely be neglibible, and thus associated to nuisance parameters which could make inference more difficult for more relevant effects. To simulate this situation, we modelled only the effect from C_1, C_2 and C_6: all of these were nonzero when modelling C_6 (Rajshahi), while this contained a single zero-value when modelling C_2 (Dhaka) owing to the lack of directed edge from C_6 to C_2.

Analysis of the posterior In Figures 4.9 and 4.10, we examine the posterior distribution of the sporadic rates μ under the empirical Bayes prior. In our simulation, Rajshahi had a low sporadic rate, with most of its events arising downstream of those in Dhaka. We can see that the posterior captures well the sporadic rate in Dhaka; for Rajshahi the posterior shifts considerably towards the lower value, which is an extreme outlier in the prior.

In Figures 4.7 and 4.8, we plot percentiles (10th, 50th and 90th) of each of the posteriors over the respective Regions, with each posterior corresponding to one of the ten datasets and either the saturated or pre-trimmed model. The figures demonstrate the possibility to identify the magnitudes of interactions from the saturated model, but also the improvement in inference for the nonzero parameters in the pre-trimmed model.

Since the prior on α_{ij} is supported on the positive reals, we necessarily overestimate these parameters, but we nevertheless see a characteristic accumulation of the 10th percentiles around zero and the medians below 0.1. We investigate this further in Figures 4.11 and 4.12, where we see that the zero α_{ij} values have a characteristic posterior shape where the mode is around zero.

We note in the case of the Rajshahi simulated dataset (Figure 4.12) that this shape is also present for the nonzero parameter of Rajshahi's self-excitation $\alpha_{6,6}$; we expect that this comes as a result of the slight overestimation of the baseline rate as seen in Figure 4.10, given where the mass of the posterior is concentrated. These parameters play relatively similar roles, in governing the overall rate of events which are not caused by outside regions.

We investigate this further by plotting the posterior distribution of the *internal rate* which we define as $\bar{\lambda}_{int,i} = \mu_i/(1-\alpha_{i,i})$, which would be the average number of events in the region in the absence of influence from other regions. In Figure 4.13, we see that this internal rate distribution captured the ground truth slightly better than the distribution of Figure 4.10, suggesting that part of our underestimation in α is offsetting our overestimation in μ as suggested. Overall, we conclude that inference for all of these quantities is challenging when events are more often triggered externally.

Despite occasional challenges in individual parameter inference, specifically with regards to the simulated Rajshahi data, one-step-ahead predictive coverage is nevertheless good, as described in Table 4.2. This justifies the use of the posterior as a predictive tool for risk monitoring; inference for individual parameters is more prone to error but may still provide meaningful insights.

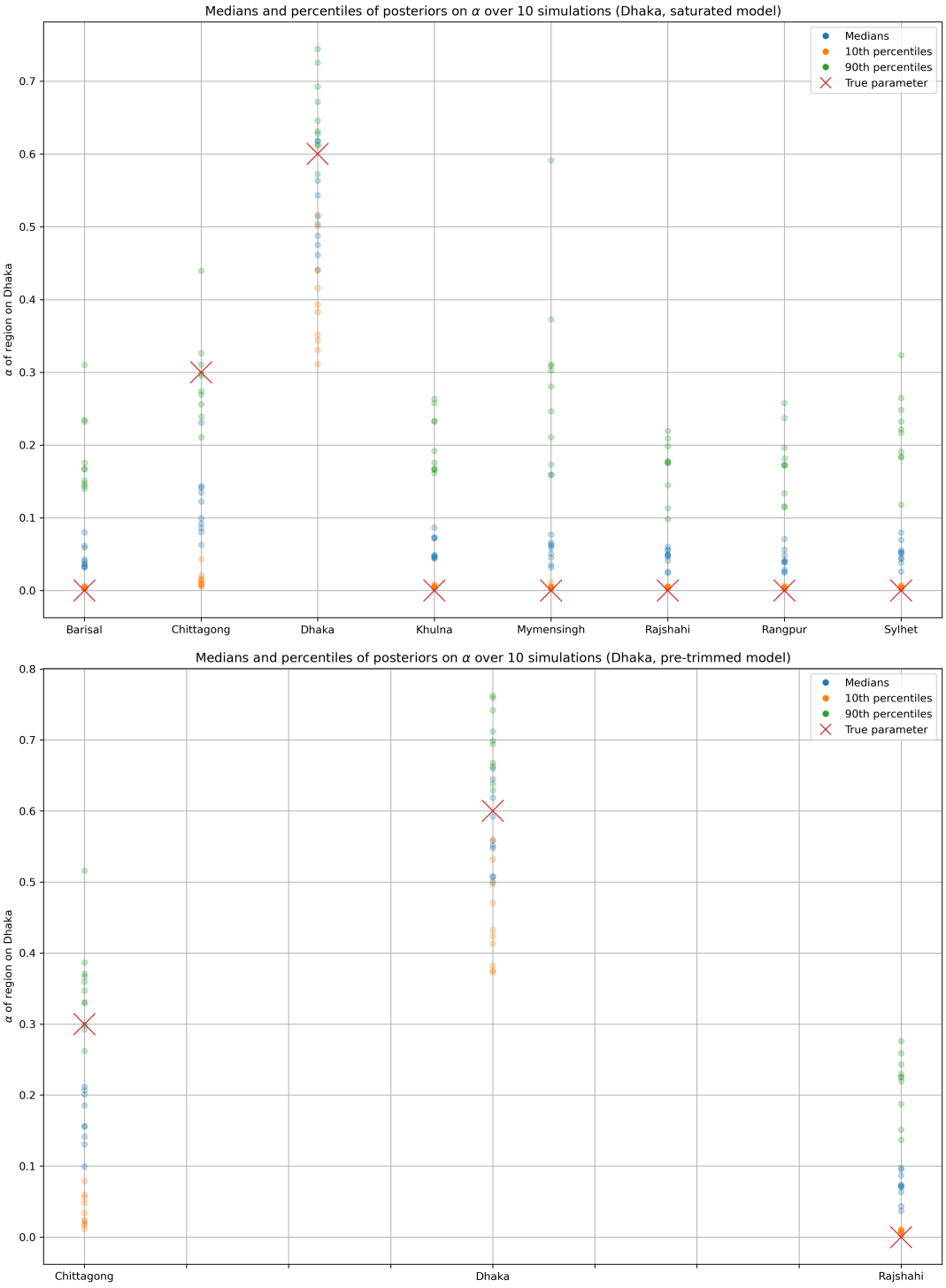


Figure 4.7: Comparison of saturated model (above) and pre-trimmed model (below) for inference on α_{ij}, for (simulated) Dhaka data. Plotted medians and percentiles of parameters across ten posteriors fit to ten simulated datasets.

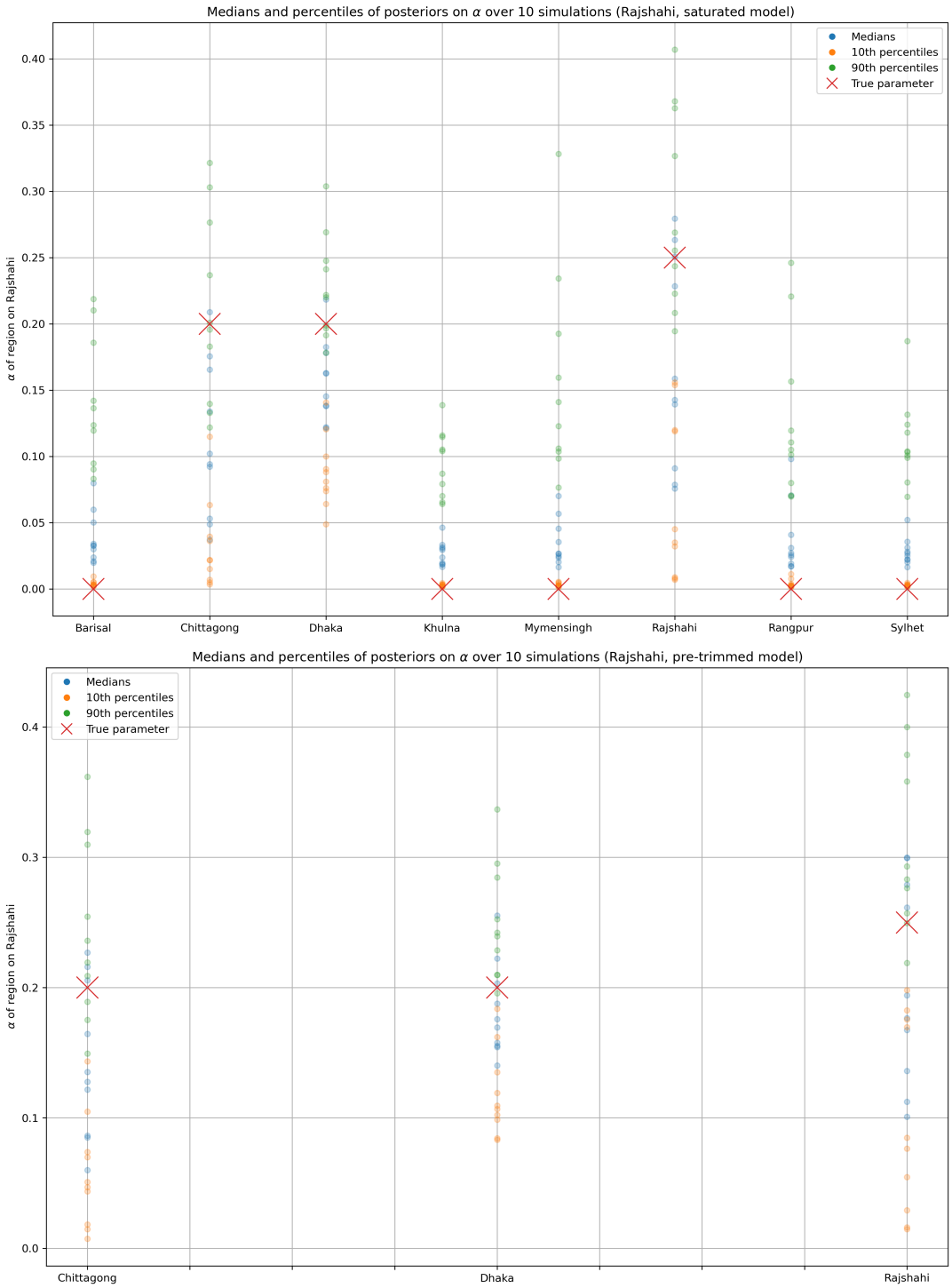


Figure 4.8: Comparison of saturated model (above) and pre-trimmed model (below) for inference on α_{ij}, for (simulated) Rajshahi data. Plotted medians and percentiles of parameters across ten posteriors fit to ten simulated datasets.

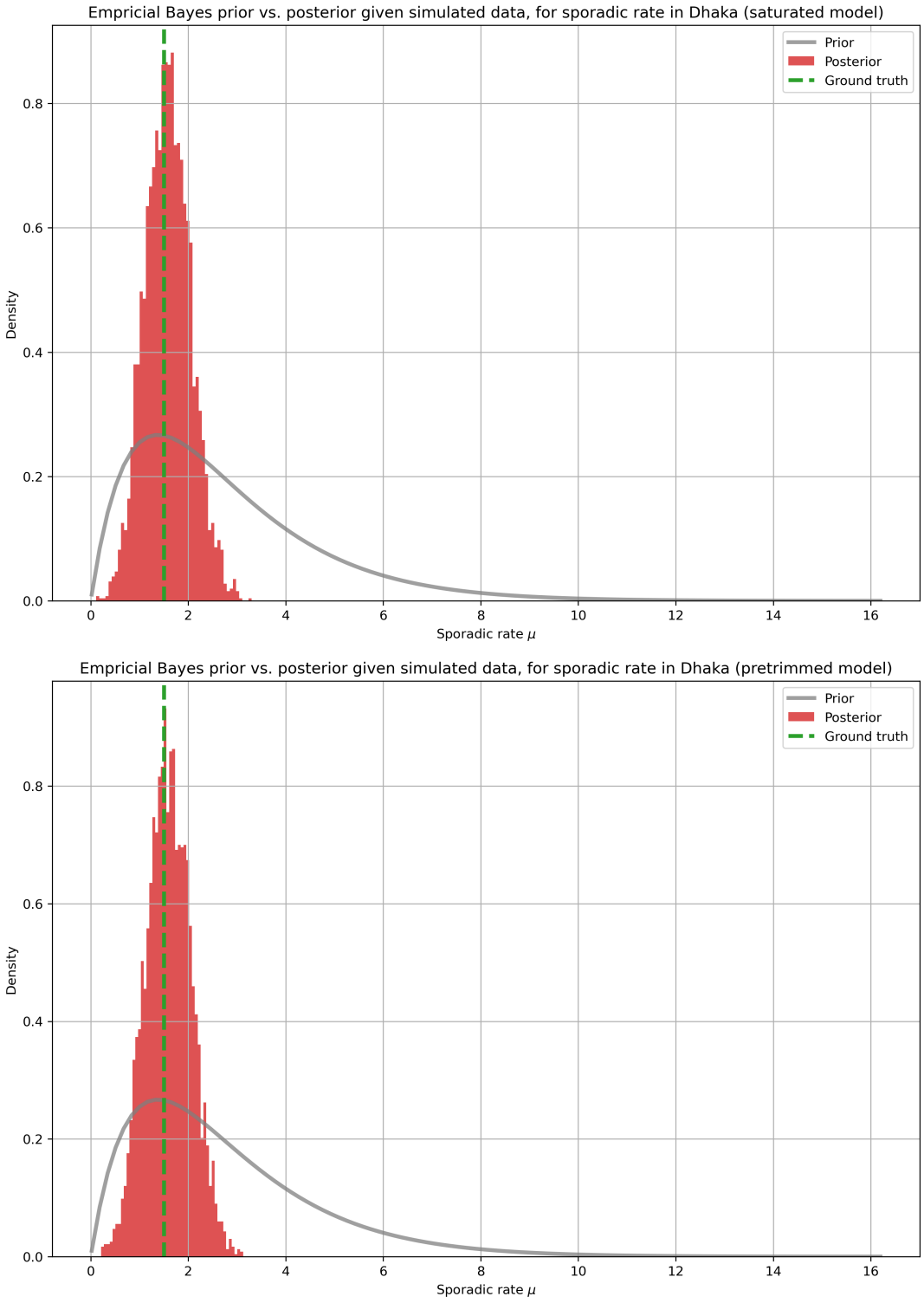


Figure 4.9: Prior/posterior plots for parameter μ_3 (Dhaka sporadic rate), given single simulated dataset. The empirical Bayes prior parameters are as described in Section 4.3.3.2. Saturated model (above) and pre-trimmed model (below).

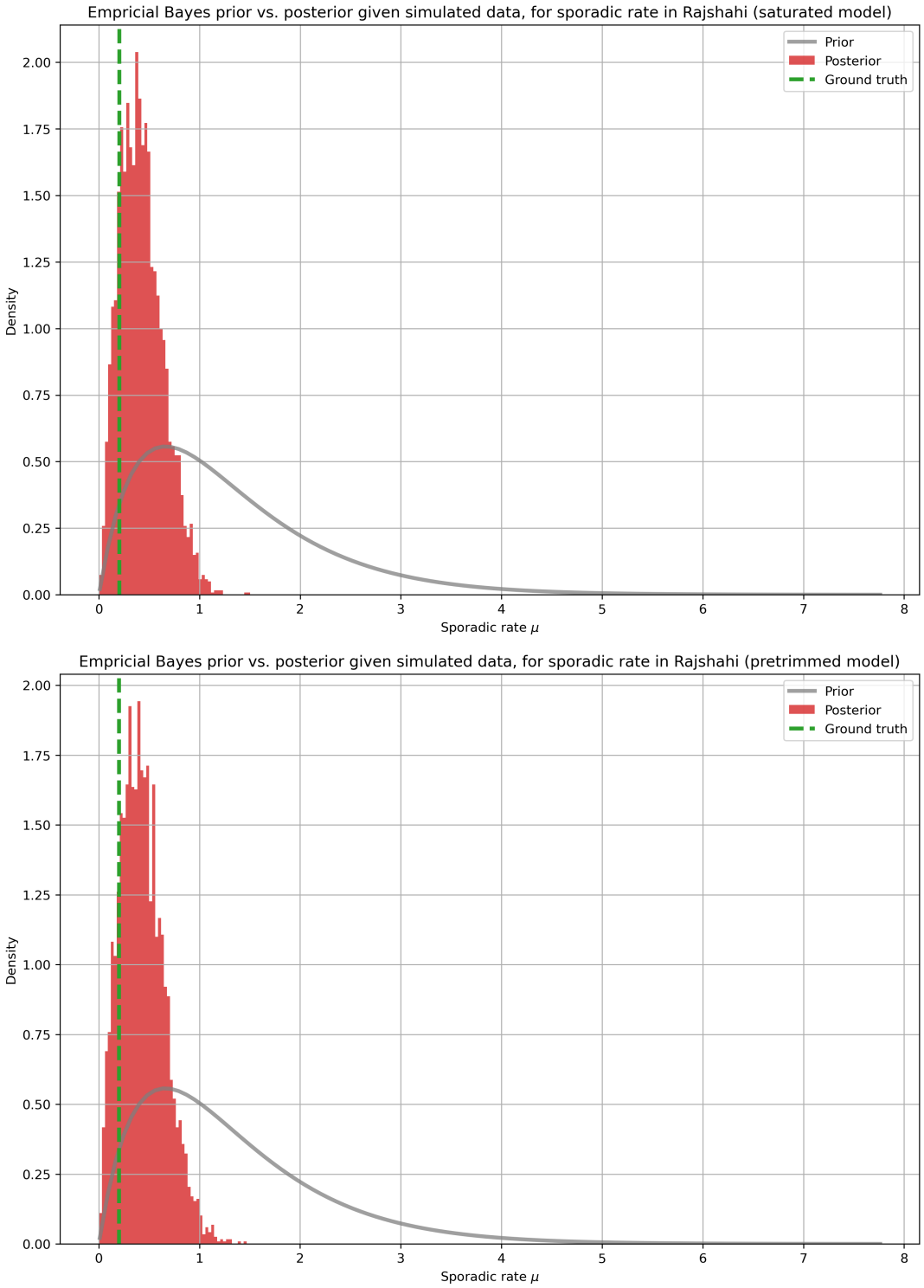


Figure 4.10: Prior/posterior plots for parameter μ_6 (Rajshahi sporadic rate), given single simulated dataset. The empirical Bayes prior parameters are as described in Section 4.3.3.2. Saturated model (above) and pre-trimmed model (below).

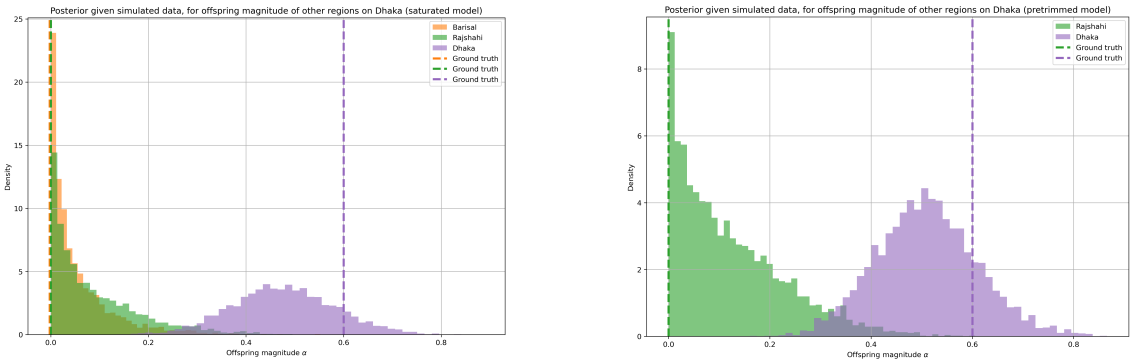


Figure 4.11: Posterior plots for parameters $\alpha_{i,3}$ with $i \in \{1, 3, 6\}$ (Impacts on Dhaka of Barisal, Dhaka and Rajshahi resp.), given single simulated dataset. Saturated model (left) and pre-trimmed model (right, excludes $\alpha_{1,3}$).

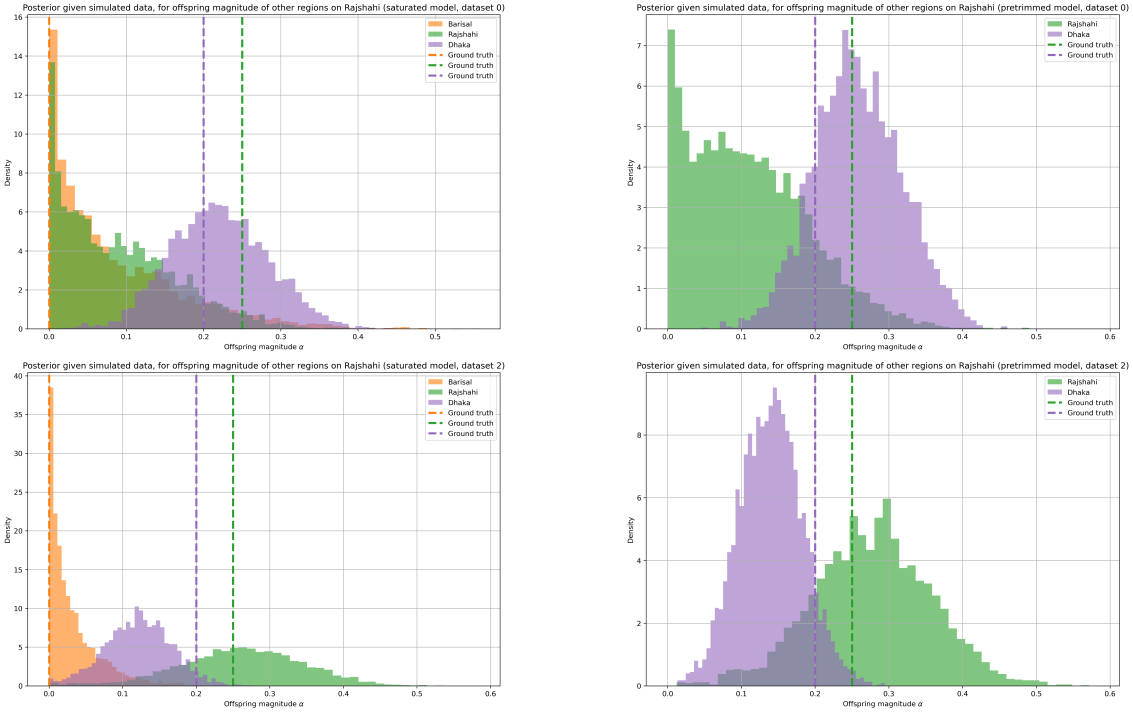


Figure 4.12: Posterior plots for parameters $\alpha_{i,6}$ with $i \in \{1, 3, 6\}$ (Impacts on Rajshahi of Barisal, Dhaka and Rajshahi resp.), given two simulated datasets (dataset 0 on top, and dataset 2 on bottom). Saturated model (left) and pre-trimmed model (right, excludes $\alpha_{1,6}$).

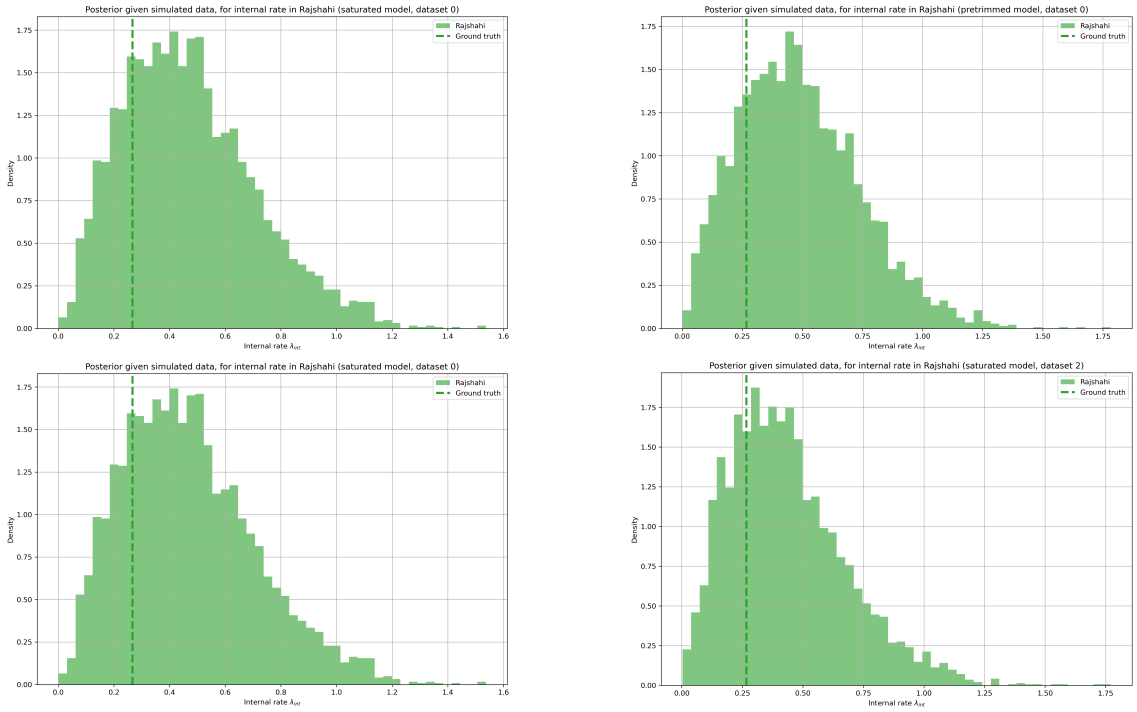


Figure 4.13: Posterior plot for parameter $\mu_6/(1-\alpha_{6,6})$ (internal rate for Rajshahi), given two simulated datasets (dataset 0 on top, and dataset 2 on bottom). Saturated model (left) and pre-trimmed model (right, excludes $\alpha_{1,6}$).

	Saturated Model			Pre-Trimmed Model		
Impacted Region	Lowest	Mean	Highest	Lowest	Mean	Highest
C_3 (Dhaka)	0.891	0.935	0.971	0.920	0.936	0.949
C_6 (Rajshahi)	0.933	0.947	0.962	0.904	0.949	0.974

Table 4.2: Coverage of one-step ahead 90% predictive bands in simulated data. Independent datasets used for posterior fit and coverage computation. Lowest and highest refer to the lowest and highest coverages across the ten simulated dataset, mean refers to the mean coverage across all simulated datasets.

4.4.2 Application to Bangladesh data from ACLED

Having developed an understanding of the model in Section 4.4.1, we present a selection of results from running the model on real data from ACLED, for Riots and Protests. Given the complexity of the model, we restrict our focus in this section to the posterior predictive offered by the Bayesian approach, and on the ability of the model to capture spatial interactions in the multivariate paradigm we consider.

Datasets and model configurations We consider two disjoint datasets, 2010-2015 and 2016-2021, to assess performance in two disjoint (and relatively independent) periods. In each case, we consider a saturated model, a pre-trimmed model in which incoming α parameters not corresponding to Dhaka, Chittagong, or the region itself are set to zero. We also fit a univariate model where no spatial interactions are permitted.

Spatial interactions In Figures 4.14, 4.15, 4.16, 4.17, 4.18 and 4.19, we present the 'incoming' α parameters for a given impacted region. These are the parameters $(\alpha_{ij})_i$ as we vary i, for fixed j. Here we describe α_{ij} as the excitation from i on j, with j the impacted region. In each plot, j is fixed and each violin corresponds to the marginal posterior for the described choice of i. For ease of exposition, we focus on the regions j being Dhaka, Chittagong and Rajshahi, and consider only the fit for the model for the more recent years 2016-2021. We plot the line $\alpha_{ij} = 0.1$, to mark the region where most spurious α_{ij} values accumulated in the simulation study (see Figures 4.7 and 4.8).

Coverage of one-step-ahead predictive bands For each training period, each impacted region, and each model, we compute the coverage of one-step-ahead predictive bands on hold-out data which is the year after training. We also quote the median size of these predictive bands, as there is interest in having good coverage with minimal size. We list these values in Tables 4.3 and 4.4.

Interestingly, the univariate model typically offers similar predictive performance as the pre-trimmed and saturated models. We remark that there is a slight decline in the case of Dhaka; this region has the most events and thus is likely to benefit most from a more complex model. In this case, the pre-trimmed and saturated models offer similar performances, but the pre-trimmed model allows for greater reliability in interpreting the α_{ij} parameters, as we discuss. In the case of Rajshahi, we note that the saturated model has considerably larger predictive intervals; the large number of supurious α_{ij} included in the posterior lead to greater variability in posterior predictions.

As such, our recommendation is for practitioner's to carefully select regions for which they have a prior belief that interaction is likely to be meaningful, as including too many parameters is more harmful than including too few parameters, especially when the event counts are low. We demonstrated our model with a naive pre-trimming based on the largest regions, but expert guidance can help practitioners to optimally tune this choice.

Visualisation of one-step-ahead predictive bands As a means to visualise the one-step-ahead prediction process, we plot the excitation and one-step-ahead bands against the counts, by time, in Figures 4.20, 4.21 and 4.22. Here we restrict to the same handful of experiments for which we plotted the incoming α parameters, for the purposes of presentation. The plots also demonstrate how the intensity of the process adapts in response to events.

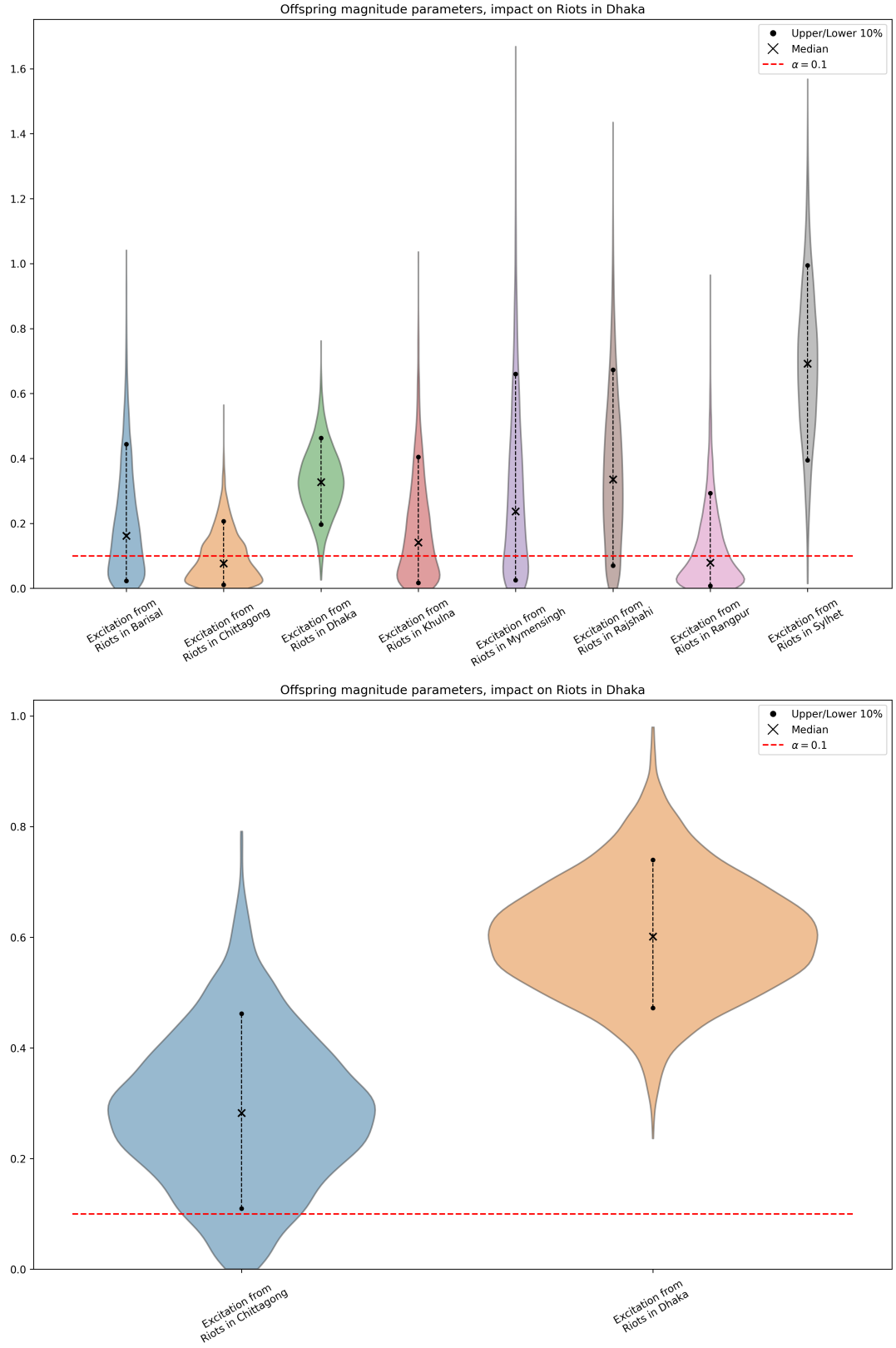


Figure 4.14: α parameters influencing Riots in Dhaka, for saturated and pre-trimmed model trained from 2016-2021. Several distributions in the saturated models are diffuse, and may obfuscate the meaningful parameters. In the pre-trimmed model, the impact of Chittagong and Dhaka is seen more clearly.

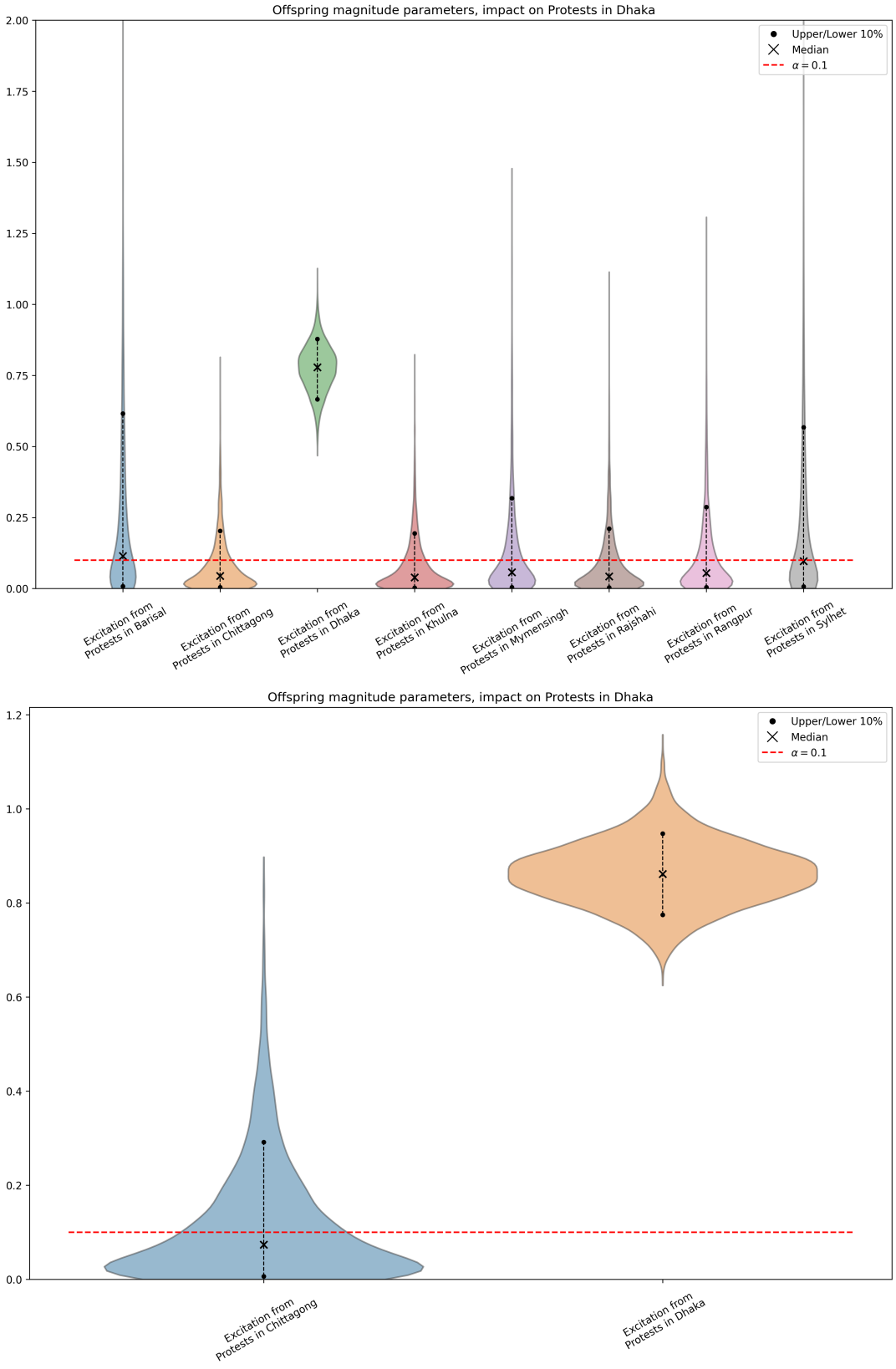


Figure 4.15: α parameters influencing Protests in Dhaka, for saturated and pre-trimmed model trained from 2016-2021. In both the saturated and pre-trimmed models, the only meaningful excitation seems to be coming from Dhaka onto itself; the other parameters have the bulk of the mass around zero as discussed in Section 4.4.1. This reflects the inbalance in the real counts of protests (see Figure 4.3)

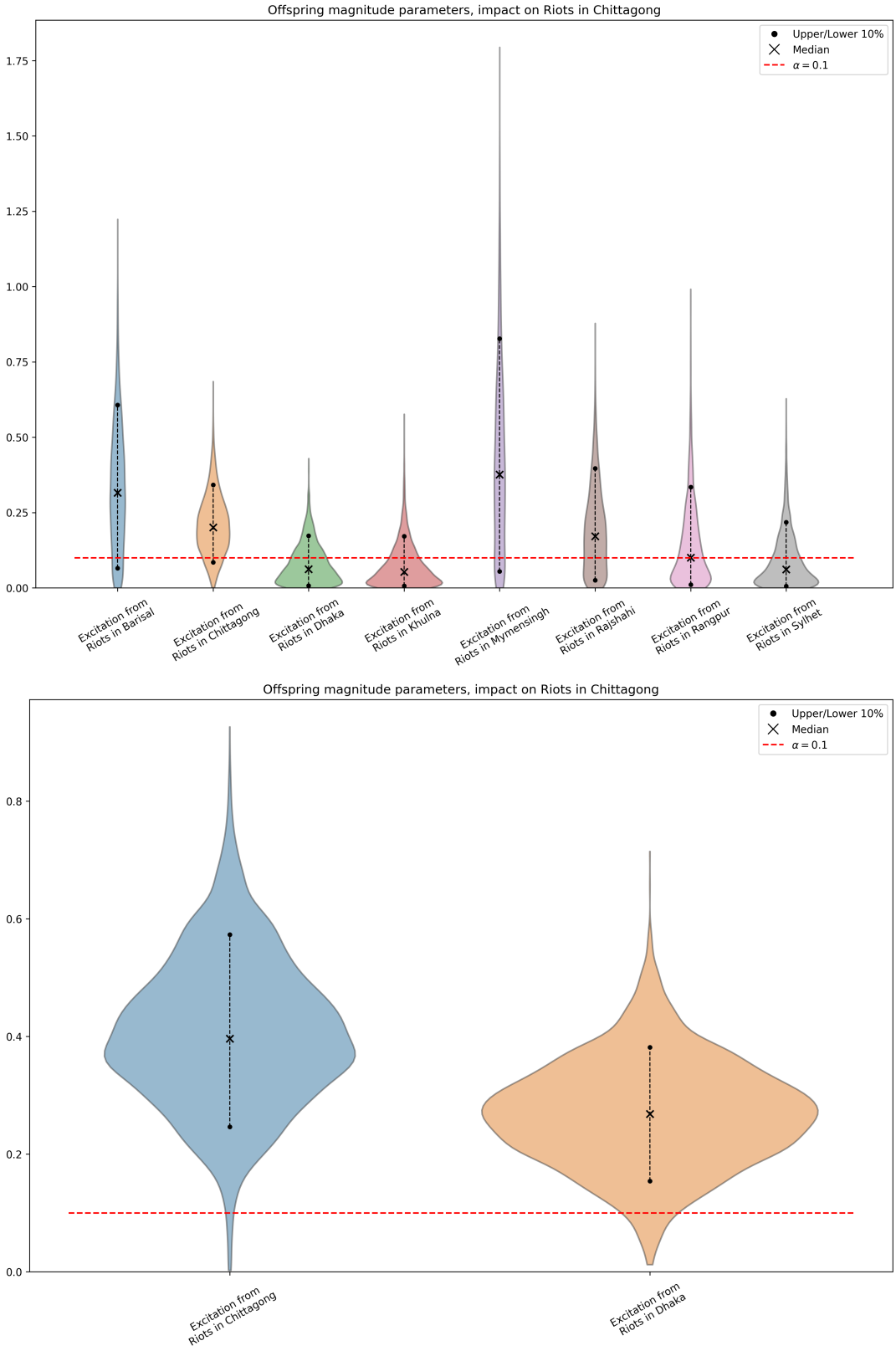


Figure 4.16: α parameters influencing Riots in Chittagong, for saturated and pre-trimmed model trained from 2016-2021. The impact of Dhaka appears higher in the pre-trimmed case; the saturated model has the posterior over the α associated with Dhaka more concentrated around zero.

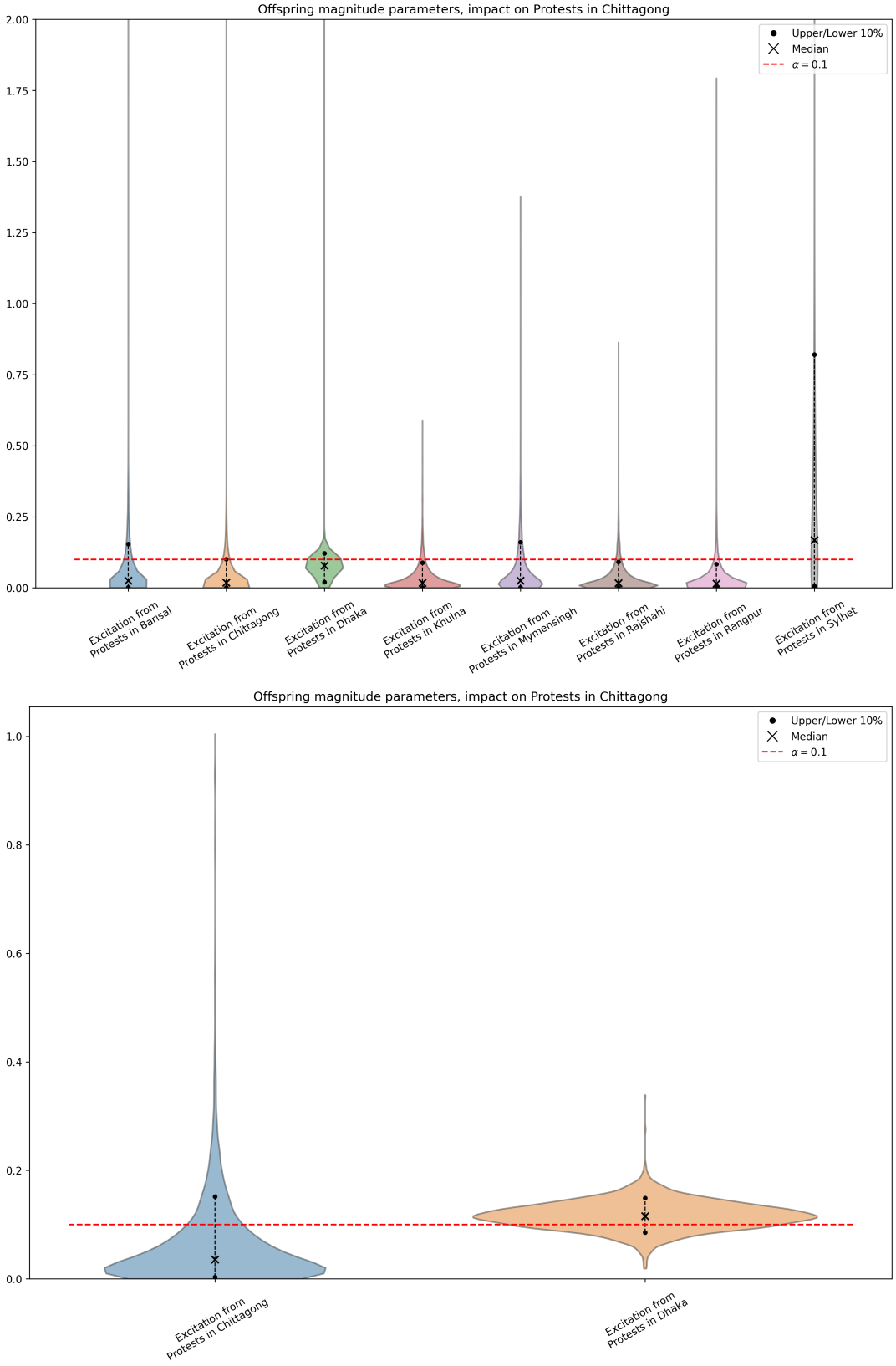


Figure 4.17: α parameters influencing Protests in Chittagong, for saturated and pre-trimmed model trained from 2016-2021. As in the Dhaka case, the interaction is most strong from Dhaka, which the model may learn more easily due to the higher amount of protests in that region.

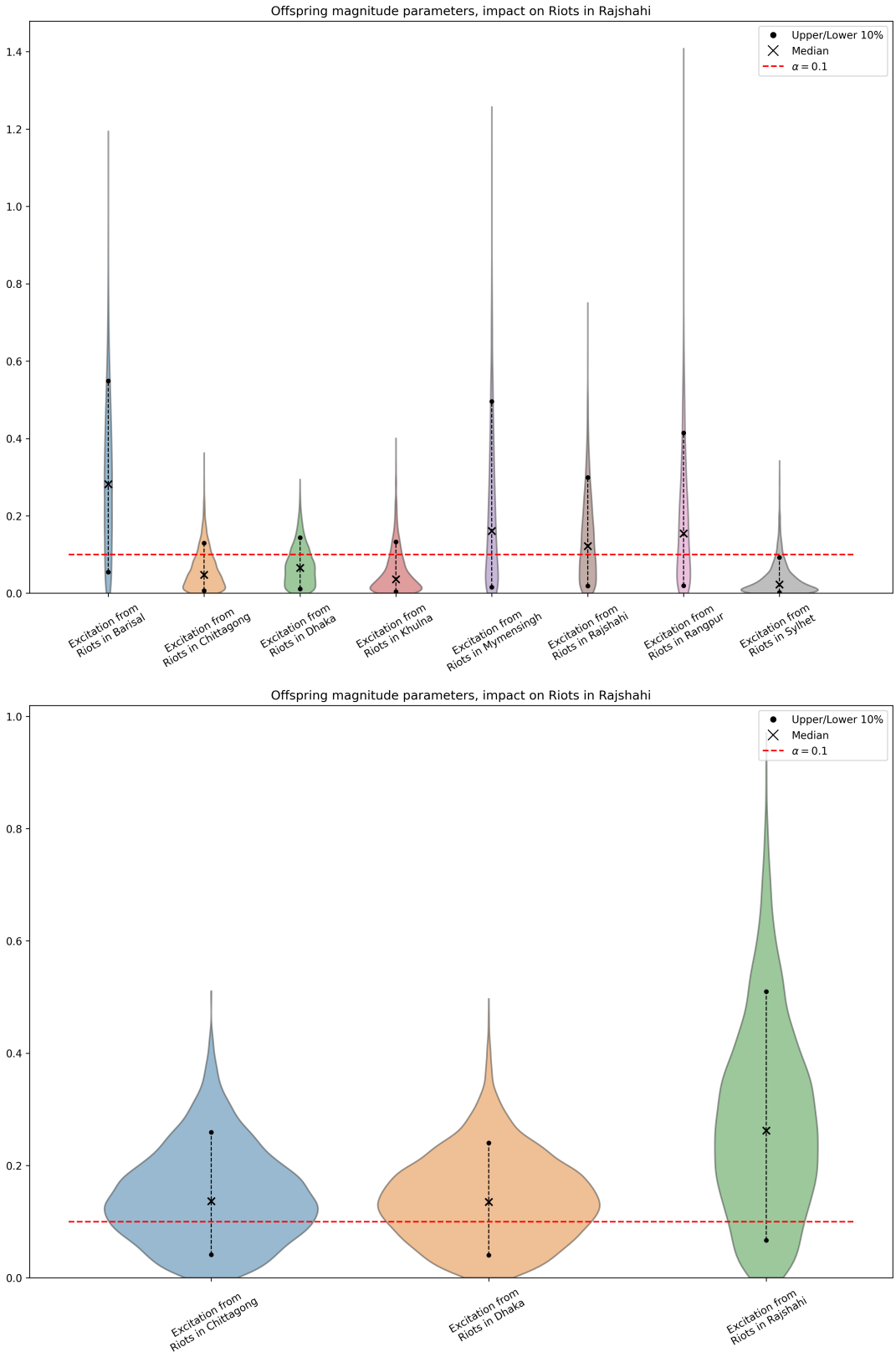


Figure 4.18: α parameters influencing Riots in Rajshahi, for saturated and pre-trimmed model trained from 2016-2021. The posteriors in the pre-trimmed model are better concentrated, by virtue of the reduced complexity of the parameter space.

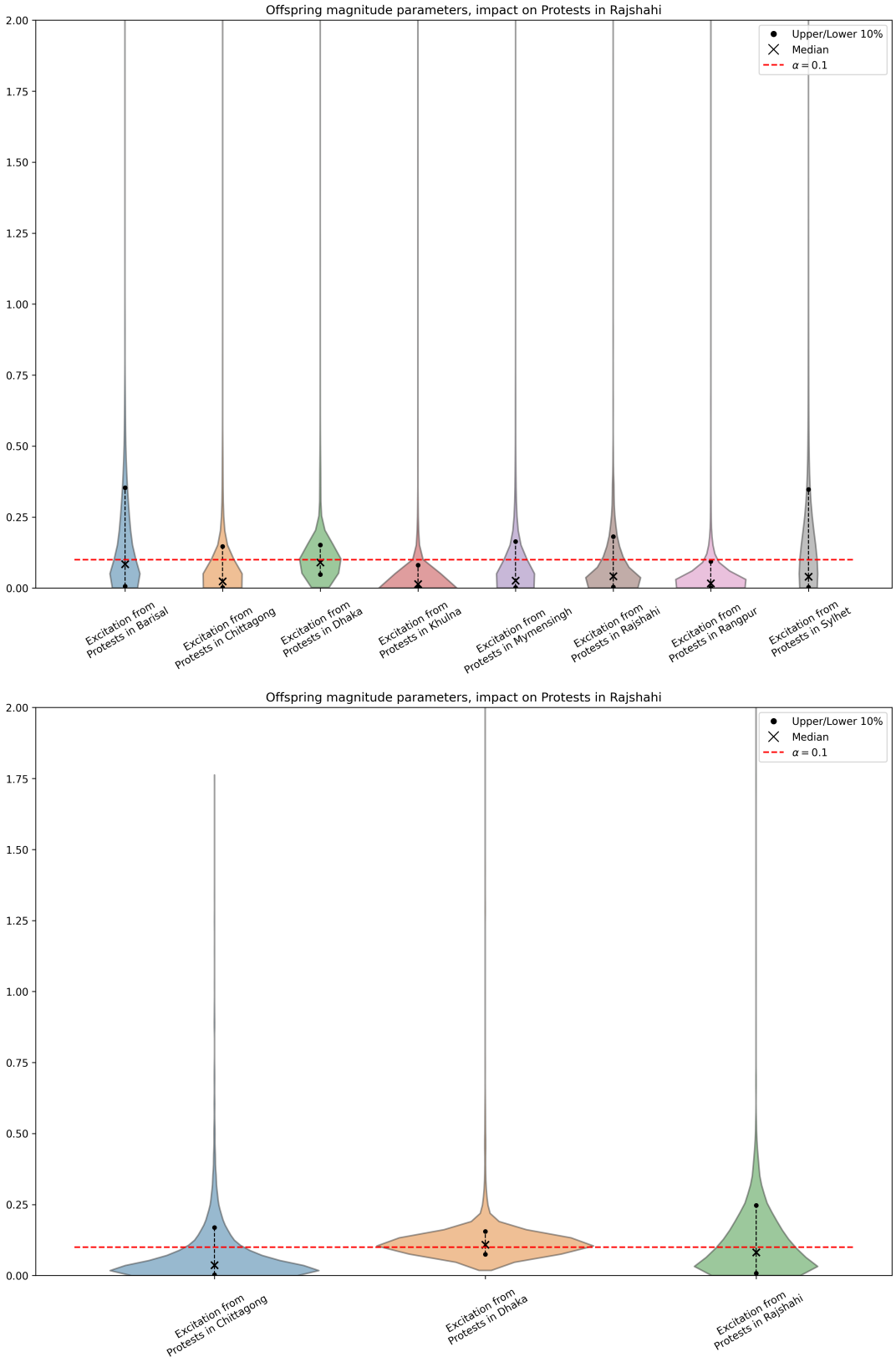


Figure 4.19: α parameters influencing Protests in Rajshahi, for saturated and pre-trimmed model trained from 2016-2021. As previously, the only meaningful interaction the model is able to learn is from Dhaka.

	Pre-trimmed Model		Saturated Model		Univariate Model	
Region	2010-2015	2016-2021	2010-2015	2016-2021	2010-2015	2016-2021
Barisal	1.000	1.000	1.000	1.000	1.000	0.887
Chittagong	0.962	0.849	0.943	0.925	0.962	0.906
Dhaka	0.849	0.774	0.868	0.792	0.830	0.755
Khulna	0.981	0.925	0.981	0.906	0.962	0.906
Mymensingh	1.000	1.000	1.000	0.981	1.000	0.962
Rajshahi	0.962	0.906	0.943	1.000	0.962	0.925
Rangpur	1.000	0.981	1.000	1.000	1.000	0.962
Sylhet	0.981	0.943	0.981	0.943	0.981	0.981

	Pre-trimmed Model		Saturated Model		Univariate Model	
Region	2010-2015	2016-2021	2010-2015	2016-2021	2010-2015	2016-2021
Barisal	2	17	3	44	1	4
Chittagong	3	5	2	5	2	6
Dhaka	5	11	5	11	6	10
Khulna	2	4	2	4	2	4
Mymensingh	1	6	1	3	1	3
Rajshahi	3	5	2	11	2	5
Rangpur	2	4	2	5	1	4
Sylhet	2	3	1	3	1	4

Table 4.3: Coverage (above) and median widths (below) of one-step-ahead 90% predictive bands for saturated, pre-trimmed, and univariate models, for **Protests** data. Predictive intervals are evaluated on the year following training.

	Pre-trimmed Model		Saturated Model		Univariate Model	
Region	2010-2015	2016-2021	2010-2015	2016-2021	2010-2015	2016-2021
Barisal	0.981	0.868	0.981	0.868	0.981	0.868
Chittagong	0.943	0.868	0.962	0.925	0.943	0.887
Dhaka	0.887	0.925	0.906	0.906	0.887	0.849
Khulna	0.943	0.906	0.962	0.943	0.962	0.925
Mymensingh	0.962	0.981	0.962	0.981	1.000	0.981
Rajshahi	0.962	0.925	0.962	0.925	0.962	0.943
Rangpur	1.000	0.981	1.000	0.981	1.000	0.943
Sylhet	0.981	0.943	0.981	0.981	0.981	0.925

	Pre-trimmed Model		Saturated Model		Univariate Model	
Region	2010-2015	2016-2021	2010-2015	2016-2021	2010-2015	2016-2021
Barisal	2	3	2	3	2	3
Chittagong	3	5	3	5	4	5
Dhaka	5	6	4	6	5	6
Khulna	3	3	3	3	3	4
Mymensingh	1	2	1	2	2	2
Rajshahi	3	4	3	4	3	4
Rangpur	2	2	2	2	2	2
Sylhet	3	2	3	2	3	2

Table 4.4: Coverage (above) and median widths (below) of one-step-ahead 90% predictive bands for saturated, pre-trimmed, and univariate models, for **Riots** data. Predictive intervals are evaluated on the year following training.



Figure 4.20: One-step ahead 90% posterior predictive bands, for modelling events in Dhaka (Top two are for Riots, Bottom two are for Protests). The unshaded region is out-of-sample and is the basis of the reported coverages.



Figure 4.21: One-step ahead 90% posterior predictive bands, for modelling events in Chittagong (Top two are for Riots, Bottom two are for Protests). The unshaded region is out-of-sample and is the basis of the reported coverages.



Figure 4.22: One-step ahead 90% posterior predictive bands, for modelling events in Rajshahi (Top two are for Riots, Bottom two are for Protests). The unshaded region is out-of-sample and is the basis of the reported coverages. We can see the predictive band is considerably wider in the saturated model for Protests.

4.5 Conclusion

By taking a multivariate Bayesian approach, we are able to learn complex temporal dependencies between different spatial regions in an interpretable, computationally tractable, and statistically principled way, while assuming very little about the nature or cause of the interaction, which may be very complex. Jointly, this approach allows us to model the impact on specific regions of interest without imposing any model on a broader set of impacting regions. Furthermore, the set of impacting regions can be adjusted at will, depending on the practitioner's prior knowledge about the likelihood of a meaningful spatial interaction.

We have demonstrated that an empirically driven prior choice for μ_j, alongside a one-sided Horseshoe-type prior for the α_{ij}, recovers well the strong spatial interactions while respecting a prior assumption that several spatial interactions may be negligible. The approach works particularly well in terms of predictive coverage; inference of individual parameters can provide meaningful results, although occasionally the tails of the posterior distribution may be more representative of the true structure of the data generating process. Alongside a careful choice of regions of possible impact, we believe this approach provides an excellent trade off between model complexity, and model expressiveness, allowing us to capture meaningful insights from complex data.

Our hope is that such analysis, especially when informed by regional experts, can assist in improving risk metrics for conflict, by providing a more finely grained understanding of how conflict is likely to evolve in space and time, while retaining interpretability. Our approach means that we are able to quantify uncertainty in these estimates and provide posterior predictive analytics which quantify uncertainty in our parameter estimates, and the downstream distribution of future conflict activity.

There are a number of possible extensions for further work. One direction would be to consider a blend of the model we present here and that of [16], such that a spatial kernel is adopted on a higher-resolution spatial grid, within a larger region that we model in a multivariate way. An alternative line of work, also suggested by [16], would seek to further understand the human impact of these conflict events, by modelling the ACLED data on fatalities jointly with (or conditional on) the events, possibly also introducing exogenous covariates at this step which relate to mortality.

Acknowledgements

The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the CDT StatML.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Bayesian Analysis of Conflict Risk with Multivariate Discrete-time Hawkes Processes
Publication Status	□Published □ Accepted for Publication □Submitted for Publication ☑Unpublished and unsubmitted work written in a manuscript style
Publication Details	Unpublished work written in a manuscript style. Produced in collaboration with Prof. Judith Rousseau (University of Oxford), Prof. Kerrie Mengersen (QUT), Dr. Hamish Patten (IFRC) and Dr Raiha Browning (University of Warwick).

Student Confirmation

Student Name:	Daniel Moss		
Contribution to the Paper	I personally contributed to the development of the methodology, wrote code for modeling, analysed model output, and wrote the manuscript. Collaboration was primarily limited to bi-weekly meetings to discuss project direction - most work was completed by myself, although I used existing code from Dr. Patten and Dr. Browning as a starting point. As such, collaborator's contributions were primarily limited to providing ideas to pursue, and general guidance on direction.		
Signature	D.P. Moss	Date	04/01/2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Judith Rousseau		
Supervisor comments			
Signature		Date	04/01/2024

This completed form should be included in the thesis, at the end of the relevant chapter.

5. Conclusion

In this thesis, we study two distinct topics in Bayesian statistics, which are thematically linked in their study of dependent data. In this section, we conclude with a brief summary of the contributions of each line of work; more detailed discussions of each contribution, together with directions for future work, are found in the conclusions of each individual chapter.

5.1 Semiparametric Hidden Markov Models

In Chapters 2 and 3, we focused on Bayesian estimation of Hidden Markov Models, from both a theoretical and practical perspective. While Chapter 2 heavily emphasized theoretical results, Chapter 3 focused on addressing a practical limitation of this method in a computationally tractable way, within the Bayesian framework.

Chapter 2 is the first theoretical work in the literature addressing semiparametric estimation of finite state space, nonparametric HMMs. Moreover, it builds on prior work of [102, 103] concerning posterior concentration, to show that posterior concentration (and indeed any frequentist rate of convergence) of marginal densities in L^1 translates to a corresponding L^1 rate on the emission densities. These emission densities, describing the latent structure of the data generating process, are often of more interest than the marginal when interpreting model estimates, and so it is vital that such estimates enjoy also the same theoretical guarantees.

By adapting the cut posterior approach, we have been able to identify a single joint distribution over the parameter space which effectively ties together the contributions concerning semiparametric estimation of Q, with the extension of prior L^1 concentration results to the emission densities $\mathbf{f}$. To our knowledge, this is the first work which studies cut posterior techniques in an asymptotic Bayesian context, and while our theory was directed towards Bayesian estimation of HMMs, the general framework we develop, extending the framework of [47, 48], is of broader interest. The cut posterior approach already enjoys use

in a wide range of applications, so studying its theoretical properties is well-motivated.

Although the semiparametric approach to estimating Q, as detailed in Chapter 2, is easy to implement in practice, one of its practical limitations is the need to know a priori the number R^* of hidden states. In Chapter 3, we analyse a Bayesian approach to estimating R^*, presenting theoretical justifications alongside empirical study. In this way, Chapter 3 addresses a practical shortcoming of Chapter 2, while simultaneously providing a Bayesian algorithm similar in spirit to the widely-used BIC. With the HMMs we study in Chapter 2 being so closely linked to the multidimensional mixture models considered in [45], our results were also demonstrated to hold in this case.

5.2 Bayesian Modelling of Conflict Data

In Chapter 4, we turned our attention to the Bayesian analysis of conflict risk, once again studying Bayesian models which Incorporated dependence between our observations.

The key outcome of this work was to develop an interpretable, predictive model, capable of uncovering spatial and temporal structure between conflict events, which has applications in modelling of the conflict data provided by ACLED. This was especially motivated by some of the limitations of the model considered in [16], especially in the appplication we consider of conflict in Bangladesh. We were able to demonstrate the efficacy of sparsity inducing priors, of the kind used by [31], as well as the benefit of expert prior knowledge, in understanding and predicting the spatiotemporal evolution of conflict events.

Bibliography

[1] Kweku Abraham, Ismael Castillo, and Elisabeth Gassiat. Multiple testing in non-parametric hidden markov models: An empirical bayes approach. *arXiv preprint arXiv:2101.03838*, 2021.

[2] Kweku Abraham, Zacharie Naulet, and Elisabeth Gassiat. Fundamental limits for learning hidden markov model parameters. *arXiv preprint arXiv:2106.12936*, 2021.

[3] Pierre Ailliot, Craig Thompson, and Peter Thomson. Space–time modelling of precipitation by using a hidden markov model and censored gaussian distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(3):405–426, 2009.

[4] Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. Nonparametric identification and maximum likelihood estimation for hidden markov models. *Biometrika*, 103(2):423–434, 2016.

[5] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6 A):3099–3132, 2009.

[6] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, 25, 2012.

[7] Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. mar 2012.

[8] Animashree Anandkumar, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. Technical report, 2014.

[9] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[10] James Bennett and Jon Wakefield. Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics*, 57(3):803–812, 2001.

[11] Peter J Bickel and Ya'acov Ritov. Inference in hidden markov models i: Local asymptotic normality in the stationary case. *Bernoulli*, 2(3):199–228, 1996.

[12] Peter J Bickel, Ya'acov Ritov, and Tobias Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

[13] Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. Inference of multivariate exponential hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4):91, 2023.

[14] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[15] Dan Braha. Global civil unrest: contagion, self-organization, and prediction. *PloS one*, 7(10):e48596, 2012.

[16] Raiha Browning, Judith Rousseau, Hamish Patten, and Kerrie Mengersen. *Bayesian Approaches For Modelling Discrete-Time Self-Exciting Processes and Their Applications*. PhD thesis, Queensland University of Technology, 2023. Chapter 7: Spatiotemporal modelling of conflict events in South Asia using discrete-time Hawkes processes, pages 148-206.

[17] Raiha Browning, Deborah Sulem, Kerrie Mengersen, Vincent Rivoirard, and Judith Rousseau. Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of covid-19. *PloS one*, 16(4):e0250015, 2021.

[18] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. In *Proceedings of EUSFLAT conference*, pages 14–16, 2009.

[19] Chris U Carmona and Geoff K Nicholls. Scalable semi-modular inference with variational meta-posteriors. *arXiv preprint arXiv:2204.00296*, 2022.

[20] Christian Carmona and Geoff Nicholls. Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. In *International Conference on Artificial Intelligence and Statistics*, pages 4226–4235. PMLR, 2020.

[21] Ismaël Castillo. Semiparametric bernstein–von mises theorem and bias, illustrated with gaussian process priors. *Sankhya A*, 74(2):194–221, 2012.

[22] Ismaël Castillo and Judith Rousseau. A bernstein–von mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, 2015.

[23] Antoine Chambaz and Judith Rousseau. Bounds for bayesian order identification with application to mixtures. *The Annals of Statistics*, 36(2):938–962, 2008.

[24] Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022.

[25] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51(1):58–87, 2007.

[26] Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden markov models. *The Journal of Machine Learning Research*, 17(1):3842–3884, 2016.

[27] Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models. *IEEE Transactions on Information Theory*, 63(8):4758–4777, 2017.

[28] M. C. M. de Gunst and O Shcherbakova. Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17(4):342–356, 2008.

[29] Stacy L. DeRuiter, Roland Langrock, Tomas Skirbutas, Jeremy A. Goldbogen, John Calambokidis, Ari S. Friedlaender, and Brandon L. Southall. A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11(1):362 – 392, 2017.

[30] Stacy L DeRuiter, Roland Langrock, Tomas Skirbutas, Jeremy A Goldbogen, John Calambokidis, Ari S Friedlaender, and Brandon L Southall. A multivariate mixed hidden markov model for blue whale behaviour and responses to sound exposure. 2017.

[31] Isabella Deutsch and Gordon J Ross. Bayesian estimation of multivariate hawkes processes with inhibition and sparsity. *arXiv preprint arXiv:2201.05009*, 2022.

[32] Randal Douc, Éric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, 32(5):2254–2304, oct 2004.

[33] Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear time series: Theory, methods and applications with R examples*. CRC press, 2014.

[34] Mathias Drton and Martyn Plummer. A bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):323–380, 2017.

[35] Paul Embrechts, Thomas Liniger, and Lu Lin. Multivariate hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378, 2011.

[36] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. The sticky hdp-hmm: Bayesian nonparametric hidden markov models with persistent states. *Arxiv preprint*, 2007.

[37] SEMP Franssen and AW van der Vaart. Bernstein-von mises theorem for the pitman-yor process of nonnegative type. *Electronic Journal of Statistics*, 16(2):5779–5811, 2022.

[38] David Freedman. Wald lecture: On the bernstein-von mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, 1999.

[39] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[40] Mark Gales, Steve Young, et al. The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2008.

[41] Elisabeth Gassiat. Mixtures of Nonparametric Components and Hidden Markov Models. In *Handbook of Mixture Analysis*, pages 343–360. 2019.

[42] Élisabeth Gassiat, Alice Cleynen, and Stephane Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

[43] Elisabeth Gassiat and Judith Rousseau. About the posterior distribution in hidden markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 2014.

[44] Elisabeth Gassiat and Judith Rousseau. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016.

[45] Elisabeth Gassiat, Judith Rousseau, and Elodie Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electronic Journal of Statistics*, 12(1):703–740, 2018.

[46] Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2):e1005390, 2017.

[47] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

[48] Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

[49] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[50] Richard D Gill and Boris Y Levit. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1(1-2):59–79, 1995.

[51] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

[52] Adrien Hairault, Christian P Robert, and Judith Rousseau. Evidence estimation in finite and infinite mixture models and applications. *arXiv preprint arXiv:2205.05416*, 2022.

[53] Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 33(3):438–443, 1971.

[54] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[55] Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

[56] Illia Horenko and Christof Schütte. Likelihood-based estimation of multidimensional langevin models and its application to biomolecular dynamics. *Multiscale Modeling & Simulation*, 7(2):731–773, 2008.

[57] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

[58] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[59] Qi Huang, Dwayne Cohen, Sandra Komarzynski, Xiao-Mei Li, Pasquale Innominato, Francis Lévi, and Bärbel Finkenstädt. Hidden markov models for monitoring circadian rhythmicity in telemetric activity data. *Journal of The Royal Society Interface*, 15(139):20170885, 2018.

[60] James P Hughes, Peter Guttorp, and Stephen P Charles. A non-homogeneous hidden markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30, 1999.

[61] Pierre E Jacob, Lawrence M Murray, Chris C Holmes, and Christian P Robert. Better together? statistical learning in models made of modules. *arXiv preprint arXiv:1708.08719*, 2017.

[62] Willem Kruijer, Judith Rousseau, and Aad Van Der Vaart. Adaptive bayesian density estimation with location-scale mixtures. *Electronic journal of statistics*, 4:1225–1257, 2010.

[63] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications*, 18(2):95–138, 1977.

[64] Roland Langrock, Thomas Kneib, Alexander Sohn, and Stacy L DeRuiter. Nonparametric inference in hidden markov models using p-splines. *Biometrics*, 71(2):520–528, 2015.

[65] Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

[66] Luc Lehéricy. Consistent order estimation for nonparametric hidden markov models. 2019.

[67] Rémi Lemonnier, Kevin Scaman, and Argyris Kalogeratos. Multivariate hawkes processes for large-scale inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[68] Brian G Leroux. Maximum-likelihood estimation for hidden markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

[69] Scott W Linderman and Ryan P Adams. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

[70] Thomas Josef Liniger. *Multivariate hawkes processes*. PhD thesis, ETH Zurich, 2009.

[71] Yang Liu and Robert JB Goudie. A general framework for cutting feedback within modularized bayesian inference. *arXiv preprint arXiv:2211.03274*, 2022.

[72] Yang Liu and Robert JB Goudie. Stochastic approximation cut algorithm for inference in modularized bayesian models. *Statistics and Computing*, 32(1):1–15, 2022.

[73] Thomas A Louis. Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, 1982.

[74] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[75] Rogemar S Mamon and Robert James Elliott. *Hidden Markov models in finance*, volume 4. Springer, 2007.

[76] Jean-Michel Marin and Christian Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Science & Business Media, 2007.

[77] Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724, 2013.

[78] Brad McNeney and Jon A Wellner. Application of convolution theorems in semiparametric models with non-iid data. *Journal of Statistical Planning and Inference*, 91(2):441–480, 2000.

[79] David RH Miller, Tim Leek, and Richard M Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221, 1999.

[80] George Mohler. Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics*, pages 1525–1539, 2013.

[81] George Mohler, Jeremy Carter, and Rajeev Raje. Improving social harm indices with a modulated hawkes process. *International Journal of Forecasting*, 34(3):431–439, 2018.

[82] Daniel Moss and Judith Rousseau. Efficient bayesian estimation and use of cut posterior in semiparametric hidden markov models. *arXiv preprint arXiv:2203.06081*, 2022.

[83] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

[84] Toby A Patterson, Marinelle Basson, Mark V Bravington, and John S Gunn. Classifying movement behaviour in relation to environmental conditions using hidden markov models. *Journal of Animal Ecology*, 78(6):1113–1123, 2009.

[85] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32, 2015.

[86] Ted Petrie. Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969.

[87] Martyn Plummer. Cuts in bayesian graphical models. *Statistics and Computing*, 25(1):37–43, 2015.

[88] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[89] Kolyan Ray and Aad van der Vaart. On the bernstein-von mises theorem for the dirichlet process. 2021.

[90] Lewis F Richardson. Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association*, 43(244):523–546, 1948.

[91] Vincent Rivoirard and Judith Rousseau. Bernstein–von mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523, 2012.

[92] Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5):689–710, 2011.

[93] Henry Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

[94] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

[95] Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Bayesian estimation of nonlinear hawkes process. *arXiv preprint arXiv:2103.17164*, 2021.

[96] Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Scalable variational bayes methods for hawkes processes. *arXiv preprint arXiv:2212.00293*, 2022.

[97] Ollivier Taramasco and Sebastian Bauer. *RHmm: Hidden Markov Models Simulations and Estimations*, 2018. R package version 2.1.9/r443.

[98] Stan Development Team. Stan modeling language users guide and reference manual, version 2.31. 2023.

[99] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[100] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[101] Aad Van Der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

[102] Elodie Vernet. Non parametric hidden markov models with finite state space: Posterior concentration rates. *arXiv preprint arXiv:1511.08624*, 2015.

[103] Elodie Vernet. Posterior consistency for nonparametric hidden markov models with finite state space. *Electronic Journal of Statistics*, 9(1):717–752, 2015.

[104] Gregory R. Warnes, Ben Bolker, and Thomas Lumley. *gtools: Various R Programming Tools*, 2020. R package version 3.8.2.

[105] Gentry White, Fabrizio Ruggeri, and Michael D Porter. Modelling the proliferation of terrorism via diffusion and contagion. *arXiv preprint arXiv:1612.02527*, 2016.

[106] Farid Yaghouby, Pradeep Modur, and Sridhar Sunderam. Naive scoring of human sleep based on a hidden markov model of the electroencephalogram. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5028–5031. IEEE, 2014.

[107] Christopher Yau, Omiros Papaspiliopoulos, Gareth O Roberts, and Christopher Holmes. Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57, 2011.

[108] Xuejun Yu, David J Nott, and Michael Stanley Smith. Variational inference for cutting feedback in misspecified models. *Statistical Science*, 38(3):490–509, 2023.