# A SAMPLING BASED APPROACH TO BAYESIAN INFERENCE IN HIDDEN MARKOV MODELS

by

Say Pham Hong

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

A SAMPLING BASED APPROACH TO BAYESIAN

INFERENCE IN HIDDEN MARKOV MODELS

BY

Say Pham Hong

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of

Manitoba in partial fulfillment of the requirement of the degree

MASTER OF SCIENCE

# Abstract

Hidden Markov models (HMM's) have been studied extensively both in the frequentist and Bayesian literature. Typically, the expectation maximization (EM) algorithm is used for likelihood inference, whereas Markov chain Monte Carlo (MCMC) has been applied in the Bayesian setting when the number of hidden states is known. When the number of hidden states is considered unknown, however, statistical computation and analysis of HMM's becomes extremely difficult due to the complexity of the model. In the frequentist perspective, penalized likelihood and penalized minimum distance methods have been used to estimate the number of hidden states, while the Bayesian approach typically relies on reversible jump MCMC to infer the number of hidden states.

The contribution of this thesis is to propose an alternative to the use of reversible jump MCMC for Bayesian inference in HMM's. Our methodology is based on a sampling procedure developed by Fu & Wang (2002). This method is based on the discretization of density functions with respect to the Lebesgue measure. One of the great features of this method is its mathematical simplicity which makes it easy to implement relative to most other sampling procedures (including reversible jump MCMC). In Chapter 5, several examples are used to show how this technique can be used to estimate the parameters as well as the number of hidden components of a HMM model.

# Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Alex Leblanc for his support, both academic and financial. I deeply appreciate his patience, stimulating suggestions and encouragements that gave me the possibility to complete this thesis.

I would like to thank my committee members Dr. XiKui Wang and Dr. Gary Wang for their valuable feedback.

I am grateful to Dr. Brian Macpherson for proofreading my thesis.

I would also like to thank my former supervisor Dr. Dean Slonowsky for his financial support.

Special thanks go to all my friends who gave me support and encouragement.

Finally, I am deeply indebted to my parents for their support and love throughout this journey.

Say Pham Hong

The University of Manitoba

August, 2007

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations and Notation

**AIC** — Akaike Information Criterion

**BIC** — Bayesian Information Criterion

**EM** — Expectation Maximization

**HMM** — Hidden Markov Model

**MCMC** — Markov Chain Monte Carlo

**MLE** — Maximum Likelihood Estimate

$T$=Total number of observations

$k$=Number of hidden states

$\mathbf{X} = (X_1, \ldots, X_T)$ — The sequence of hidden states

$\mathbf{Y} = (Y_1, \ldots, Y_T)$ — The sequence of observations

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & & a_{2k} \\ \vdots & & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$$ — Transition probability matrix of a hidden Markov chain

$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ — Initial probability vector (stationary distribution of the hidden chain)

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ — Parameters vector indexed by the hidden states

$\boldsymbol{\phi} = (\mathbf{A}, \boldsymbol{\theta})$ — The full set of parameters of a HMM

# Chapter 1

# Introduction

Consider a real world practical application in which a process of interest cannot be directly observed. Rather, it might be observed through another process, with either continuous or discrete state space. By analogy, we can think of a process (or signal) being "corrupted" by noise, with the noisy signal being the only observable quantity, the uncorrupted process being unobservable. When the unobservable process is a Markov chain, such a pair of processes is known as a Hidden Markov Model (HMM). In other words, a HMM can be thought of as a bivariate process, with one process being an underlying unobservable (hidden) discrete stochastic process which constitutes a Markov chain. The other process is observable given its hidden counterpart.

HMM's have been applied to a variety of fields for modeling weakly dependent observations, including genetics (Churchill, 1989), signal processing (Juang & Rabiner 1991), neurophysiology (Fredkin & Rice, 1992), biology (Leroux & Puterman, 1992), economics (Albert & Chib, 1993), and ecology (Guttorp, 1995). For an extensive list of applications, interested readers are referred to the monographs of MacDonald & Zucchini (1997) and Cappé *et al.* (2005).

Although the basic theory of HMM's was introduced in the late 1960's, Baum *et al.* (1970) were the first to develop an algorithm for obtaining the maximum likelihood estimates for a HMM. Note also that, HMM's were not extensively studied until the late 80's and early 90's, when the consistency and asymptotic normality of HMM

maximum likelihood estimators were proved. See the work of Leroux (1992), Bickel *et al.* (1998) and Douc & Matias (2001) for details. A good tutorial to HMM's can be found in Rabiner (1989).

In parallel, inference for HMM's from a Bayesian perspective was not considered until after the development of Markov chain Monte Carlo (MCMC) techniques. For that matter, Robert *et al.* (1993) were the first to apply MCMC techniques in the context of HMM's. Several important properties of the Markov chains thus introduced were proved, including geometric convergence, $\phi$ mixing and the central limit theorem. A method of simulating the hidden components of a HMM using data augmentation was also proposed. Chib (1996) proposed another method for simulating these hidden components using the so-called *forward-backward recursion* which will be introduced in Chapter 3. Robert & Titterington (1998) considered the use of non-informative priors based on a reparameterization of the model.

One of the important and difficult problems linked with HMM's is the estimation of the unknown number of states of the hidden component of a HMM. The likelihood ratio test has been considered by Rydén *et al.* (1998) and Giudici *et al.* (2000). The Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been used by Leroux & Puterman (1992) and MacDonald & Zucchini (1997). However, these methods have not been proved to lead to consistent estimators of the number of hidden states. Baras & Finesso (1992) developed a consistent estimator using the penalized likelihood method, whereas Mackay (2000) proposed a consistent estimator of the number of hidden states in the stationary hidden Markov model based on a penalized minimum distance method. In the Bayesian setting, inference for the number of hidden states has also been considered. Robert *et al.* (2000) applied

the reversible jump MCMC technique developed by Green (1995) in the context of HMM's.

In this thesis, we apply the sampling method developed by Fu & Wang (2002) to come up with an alternative to the use of reversible jump MCMC in performing Bayesian inference for HMM's. This method provides not only a way to estimate the parameters of the HMM, but also the number of components (or hidden states) of the underlying Markov chain when it is considered unknown. Specifically, Chapter 2 presents basic definitions and introduces HMM's. In this chapter, we also discuss the three important basic problems (as suggested by Rabiner, 1989) linked with HMM's. These are the evaluation of the HMM likelihood, the reconstruction (or estimation) of the sequence of hidden states, and parameter estimation. In Chapter 3, we review inference for HMM's using MCMC techniques by focusing on the simulation of the hidden chain, according to the methodologies suggested by Robert *et al.* (1993) and Chib (1996). In Chapter 4, we introduce the sampling method developed by Fu & Wang (2002) and present the methodology we developed for Bayesian inference in HMM's which relies on this technique. Chapter 5 presents various applications of our methodology, including cases of HMM's having an unknown number of components.

# Chapter 2

# Hidden Markov Models

As was previously mentioned, hidden Markov models can be thought of as bivariate stochastic processes. One process, being unobservable, constitutes a Markov chain, while the other process is observable given the hidden state of the Markov chain. In the first section of this chapter, we give a very brief introduction to Markov chains and discuss some related concepts that will be relevant later on. The rest of the chapter will focus on HMM's.

## 2.1 Markov Chains

A simple way to describe a discrete time stochastic process is as follows: it is a sequence of random variables $\mathbf{X} = \{X_t\}$, where $t \in \mathcal{T} = \{0, 1, 2, 3...\}$. Let the sample space of $\mathbf{X}$ be denoted by $\mathcal{S}$. Throughout this thesis, we consider only cases where $\mathcal{S}$ is a finite set. More specifically, we assume here that $\mathcal{S} = \{1, 2, \ldots, k\}$, for some integer $k$.

In this setting $t$ is called the time index, $X_t$ is called the *state of the process at time t*, and $\mathcal{S}$ is called the *state space* of the process. Now suppose that the stochastic process $\mathbf{X}$ has the property that the distribution of the next state $X_{t+1}$ given the current state $X_t$ and all the past states $X_0$, $X_1$,..., $X_{t-1}$ depends only on the current state. Mathematically, we then have

$$P(X_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1}, ..., X_0 = i_0) = P(X_{t+1} = i_{t+1} | X_t = i_t), \quad (2.1)$$

for all $i_0, i_1, ..., i_{t+1} \in \mathcal{S}$.

**Definition 2.1.1** *For a stochastic process* $\mathbf{X} = \{X_t\}$, *if* (2.1) *is satisfied for all* $t \geq 0$, *then the process* $\mathbf{X}$ *is called a Markov Chain. Furthermore, if*

$$P(X_{t+1} = j | X_t = i) = P(X_{k+1} = j | X_k = i)$$

*for all* $t, k \in \mathcal{T}$, *and all* $i, j \in \mathcal{S}$, *then the chain is said to be time-homogeneous.*

Note that the condition that the chain is time-homogeneous is equivalent to

$$P(X_{t+1} = j | X_t = i) = a_{ij},$$

for all $t \in \mathcal{T}$ and $i, j \in \mathcal{S}$, that is, the conditional distributions are the same as time evolves. In this case, $a_{ij}$ is referred to as a *one-step transition probability* and the square matrix $\mathbf{A} = (a_{ij})_{i,j \in \mathcal{S}}$ is called the *one-step transition probability matrix.* The $\mathbf{A}$ matrix is a *stochastic matrix*, that is, every row $\{A_i = (a_{ij})_{j \in \mathcal{S}} : i \in \mathcal{S}\}$ of $\mathbf{A}$ defines a distribution, since

$$0 \leq a_{ij} \leq 1,$$

and

$$\sum_{j \in \mathcal{S}} a_{ij} = 1, \qquad \text{for all } i \in \mathcal{S}.$$

Also, define the n-step transition probabilities as

$$a_{ij}^{(n)} = P(X_{t+n} = j | X_t = i),$$

and let $\mathbf{A}^{(n)}$ denote the n-step transition probability matrix. The following theorem can be obtained

**Theorem 2.1.2** *Let* $\mathbf{X} = \{X_t\}$ *be a time-homogeneous Markov chain. Then, the n-step transition probabilities satisfy*

1. *the Chapman-Kolmogorov equations:*

$$\mathbf{A}^{(t+n)} = \mathbf{A}^{(t)}\mathbf{A}^{(n)}, \quad \textit{for all } t,n \geq 0;$$

2. $\mathbf{A}^{(n)} = \mathbf{A}^n, \quad \textit{for all } n \geq 0.$

An important consequence of the previous result is that the transition probabilities of the Markov chain $\mathbf{X}$ are fully determined by the one-step transition probabilities, that is, by the $\mathbf{A}$ matrix.

**Definition 2.1.3** *The initial distribution of a discrete time Markov chain is the probability mass function (p.m.f.)* $\boldsymbol{\pi} = (\pi_i)_{i \in S}$, *where*

$$\pi_i = P(X_0 = i).$$

*In other words,* $\boldsymbol{\pi}$ *is the marginal distribution of* $X_0$.

**Proposition 2.1.4** *If* $\mathbf{X}$ *has an initial distribution* $\boldsymbol{\pi}$, *then* $(\mathbf{A}^n)^T\boldsymbol{\pi}$ *is the p.m.f. of* $X_n$, *that is*

$$\begin{aligned} P(X_n = j) &= \sum_{i \in \mathcal{S}} \pi_i a_{ij}^{(n)} \\ &= ((\mathbf{A}^n)^T \pi)_j \end{aligned}$$

*for all* $j \in \mathcal{S}$, *where* $T$ *denotes transposition of a matrix or vector.*

The proof of this result is straightforward, as conditioning on $X_0$ leads to

$$\begin{aligned} P(X_n = j) &= \sum_{i \in \mathcal{S}} P(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i \in \mathcal{S}} \pi_i \, a_{ij}^{(n)} \\ &= ((\mathbf{A}^{(n)})^T \boldsymbol{\pi})_j. \end{aligned}$$

The result then follows from Theorem 2.1.2.

**Definition 2.1.5** *Let* $\mathbf{X}$ *be a Markov chain with one-step transition probability matrix* $\mathbf{A}$. *Then, a distribution* $\{\pi_j : j \in \mathcal{S}\}$ *which satisfies*

$$\pi_j = \sum_{i \in \mathcal{S}} \pi_i \, a_{ij},$$

*for all* $j \in \mathcal{S}$ *is called a stationary distribution of* $\mathbf{X}$.

Note that the previous set of equations can be written in matrix form, as

$$\boldsymbol{\pi} = \mathbf{A}^T \boldsymbol{\pi},$$

implying that $\boldsymbol{\pi}$ is a properly normalized eigenvector of $\mathbf{A}^T$ associated with the eigenvalue 1. The terminology "stationary" in Definition 2.1.5 is due to the fact that if the initial distribution of the chain is taken to be $\boldsymbol{\pi}$, then the marginal distribution of $X_t$ is unchanged as time evolves, that is, for any $t \in \mathcal{T}$ and $j \in \mathcal{S}$,

$$P(X_t = j) = P(X_0 = j).$$

In other words, both the conditional and marginal distributions are unchanged. A formal proof of this is done by induction. We here consider only the case of $t = 1$.

In this case,

$$\begin{aligned} P(X_1 = j) &= \sum_{i \in \mathcal{S}} P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i \in \mathcal{S}} \pi_i \, a_{ij} \\ &= \pi_j \\ &= P(X_0 = j). \end{aligned}$$

The general case of $t > 1$ follows from using the same arguments.

The existence and uniqueness of the stationary distribution of $\mathbf{X}$ depends on the specific properties of the Markov chain. There exist many results that can be used to establish the existence/uniqueness of the stationary distribution of a Markov chain. We state one here that is going to be good enough for our purpose. First, note that a transition probability matrix $\mathbf{A}$ is said to be *regular* if there exists $n \geq 0$ such that

$$a_{ij}^{(n)} > 0 \qquad \forall i, j \in \mathcal{S}.$$

In other words, some power $n > 0$ of $\mathbf{A}$ leads to a matrix that is strictly positive in all its entries.

**Theorem 2.1.6** *If the one-step transition probability matrix* $\mathbf{A}$ *of a time-homogeneous Markov chain* $\mathbf{X}$ *is regular, then* $\mathbf{X}$ *admits a unique stationary distribution.*

We refer the reader to Taylor and Karlin (1998, section 4.1) for more details. However, as a note for readers that are experienced with Markov chain theory, we point out that a finite state Markov chain having a regular transition probability matrix is necessarily irreducible and ergodic. (The converse is also true.) The key point to be

made here is that a stationary Markov chain having a regular transition probability matrix is fully specified by its one-step transition probability matrix $\mathbf{A}$. Indeed, from Theorem 2.1.2, for any $n \geq 0$, the n-step transition probabilities are given by $\mathbf{A}^{(n)} = \mathbf{A}^n$. Also, because of Theorem 2.1.6, this implied the initial distribution of the chain is the unique distribution $\boldsymbol{\pi}$ which is a solution to

$$\boldsymbol{\pi} = \mathbf{A}^T \boldsymbol{\pi},$$

that is, the unique distribution $\boldsymbol{\pi}$ that is an eigenvector of $\mathbf{A}^T$ associated with the eigenvalue of 1. For more details on Markov chains, we refer the reader to the monographs of Karlin and Taylor (1998) and Ross (2007).

## 2.2 Hidden Markov Models: Notation, Definition and Assumptions

We now present the formal definition of a HMM (*cf.* MacKay, 2002).

**Definition 2.2.1** *A pair of stochastic processes $\{X_t, Y_t\}$ is said to constitute a Hidden Markov Model (HMM) if it satisfies the following two conditions:*

1. *$\{X_t\}$ is a time-homogeneous Markov chain with transition probability matrix $\mathbf{A} = \{a_{ij}\}$ and initial probability vector $\boldsymbol{\pi} = \{\pi_i\}$, where $i, j \in \mathcal{S} = \{1, 2, ...k\}$.*
2. *Conditionally on $X_t$, $Y_t$ is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_{t+1}, Y_{t+2}, \ldots, Y_T$ and $X_1, X_2, \ldots, X_{t-1}, X_{t+1}, X_{t+2}, \ldots, X_T$.*

Note that the condition of time-homogeneity is not necessarily required when defining HMM's. We, however include it here as it is going to be assumed by default

throughout the rest of this thesis. Note also that this definition implies a HMM is characterized by two probabilistic mechanisms: namely, an unobserved Markov chain $\{X_t\}$ with $k$ states, and a set of distribution functions for the observables $\{Y_t\}$ given each hidden state. Commonly, the distribution of $Y_t$ given the hidden state $X_t$ is assumed to follow a specified parametric family, that is,

$$Y_t|X_t = i \sim f(y|\theta_i),$$

where $f$ is a generic function that denotes either a density function (continuous case) or a probability mass function (discrete case) indexed by a parameter $\theta$. Here, $X_t$ can be thought of as a missing label that selects the parameter used to generate $Y_t$, but with the current label $X_t$ depending on the previous label $X_{t-1}$.

Marginalizing over the hidden state $X_t$ gives the following interesting form for the unconditional distribution of the observable $Y_t$ (*cf.* Chib, 1996)

$$f(y_t) = \begin{cases} \sum_{i=1}^{k} f(y_t|\theta_i)\pi_i & \text{if } t = 1, \\ \sum_{i=1}^{k} f(y_t|\theta_i)p(X_t = i) & \text{if } t \geq 2, \end{cases} \tag{2.2}$$

which is the distribution of a finite mixture. Note that, since the hidden states constitute a Markov chain, the observations $\{Y_t\}$ generated through a HMM are dependent and possibly not identically distributed. However, if the initial distribution is taken to be a stationary distribution of the hidden Markov chain, then the distribution of $Y_t$ becomes

$$f(y_t) = \sum_{i=1}^{k} f(y_t|\theta_i)\pi_i \qquad \forall\ t \geq 1, \tag{2.3}$$

thus leading to identically distributed, but dependent observations. Note that the correlation between observations approaches zero as the time difference between these

observations becomes large (*cf.* Albert, 1991). Hence, HMM's can be thought of as an extension of i.i.d. mixture models which are suitable for modeling weakly dependent observations. By opposition, the most common mixture setups call for independent observations arising from the same mixture distribution. In this thesis, we will focus on the identically distributed dependent mixture models of the type (2.3). That is, we will concentrate on stationary HMM's.

According to Rabiner (1989), there are three basic problems associated with HMM's. They are:

1. How do we efficiently evaluate the likelihood function of the HMM?

2. How do we uncover the hidden states of the model parameters?

3. How do we estimate the model parameters?

The first problem is the problem of evaluating the likelihood function for specific values of the parameters and is usually referred to as the scoring problem. The second is the problem of "reconstructing" or finding the "most likely" hidden state sequence given the observation sequence. This is at the core of areas such as speech recognition, and is referred to as the decoding problem. Note that when the number of components $k$ (that is, the number of hidden states) is known, the third problem is simply one of estimating the parameters of the model, whereas when $k$ is unknown, it also becomes a problem of model selection, that is, finding a model that best fits the given observations among a class of HMM's with different complexity.

## 2.3 The Likelihood Function of a HMM

Let us assume for the moment that given $X_t$, the distribution of $Y_t$ is indexed by a single parameter. For example, a Poisson distribution with mean $\theta$, or a normal distribution with mean zero and unknown variance $\sigma^2$. In this setup, for a given number of hidden states $k$, we have $k^2$ parameters to be estimated from the hidden Markov chain (the elements of $\mathbf{A}$), and $k$ parameters to be estimated from the distribution of the observables $\{Y_t\}$. Remember that we assume the initial distribution $\boldsymbol{\pi}$ is a stationary distribution of $\mathbf{X}$, so that $\mathbf{X}$ is here fully specified by its one-step transition probability matrix. Let $\boldsymbol{\phi} = (a_{11}, a_{12}, ..., a_{kk}, \theta_1, \theta_2, ..., \theta_k)$. Note that $\boldsymbol{\phi}$ contains all the parameters to be estimated.

### 2.3.1 The Form of the Likelihood Function

For a fixed number $k$ of components and given $\{X_t : t = 1, 2, ..., T\}$, the joint distribution (density or mass function) of $\boldsymbol{Y} = (Y_1, ..., Y_T)$ is, by the assumption of conditional independence,

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}) &= P(y_1, ..., y_T|x_1, ..., x_T, \boldsymbol{\phi}) \\ &= \prod_{t=1}^{T} P(y_t|x_t, \boldsymbol{\phi}) \\ &= \prod_{t=1}^{T} f(y_t|\theta_{x_t}). \end{aligned} \tag{2.4}$$

On the other hand, the joint probability of the hidden state sequence $\mathbf{X} = (X_1, ..., X_T)$ is

$$P(\mathbf{x}|\boldsymbol{\phi}) = P(x_1, ..., x_T|\boldsymbol{\phi}) = \pi_{x_1} \prod_{t=2}^{T} a_{x_{t-1}, x_t}, \tag{2.5}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ is the initial distribution of $\mathbf{X}$ (and thus, by assumption, is also the unique stationary distribution of the chain). Multiplying (2.4) and (2.5) produces the joint density of $\mathbf{x}$ and $\mathbf{y}$ as

$$\begin{aligned} P(\mathbf{x}, \mathbf{y}|\boldsymbol{\phi}) &= P(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi})\, P(\mathbf{x}|\boldsymbol{\phi}) \\ &= \left(\prod_{t=1}^{T} f(y_t|\theta_{x_t})\right)\left(\pi_{x_1}\prod_{t=2}^{T} a_{x_{t-1},x_t}\right). \end{aligned} \tag{2.6}$$

Thus, the likelihood of the observed sequence, obtained by summing over all the possible hidden state sequences, is

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = P(y_1, ..., y_T|\boldsymbol{\phi}) = \sum_{x_1=1}^{k} \cdots \sum_{x_T=1}^{k} \pi_{x_1} f(y_1|\theta_{x_1}) \prod_{t=2}^{T} a_{x_{t-1},x_t} f(y_t|\theta_{x_t}). \tag{2.7}$$

Note that this form of the likelihood involves a sum of $k^T$ terms, each of which is itself a product of $T^2$ terms. Obviously then, it quickly becomes infeasible to evaluate this likelihood except for very small values of $k$ and $T$. However, with a slight modification, (2.7) can be expressed in the following matrix form (*cf.* MacDonald & Zucchini, 1997, p.78)

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = \boldsymbol{\pi}^* \left[\prod_{t=2}^{T} (\mathbf{D}(y_t)\mathbf{A})\right] \mathbf{1}, \tag{2.8}$$

with

$$\mathbf{D}(y_t) = Diag\{f(y_t|\theta_1), ..., f(y_t|\theta_k)\},$$

$$\boldsymbol{\pi}^* = \boldsymbol{\pi} \cdot Diag\{f(y_1|\theta_1), ..., f(y_1|\theta_k)\},$$

$$\mathbf{1} = k \times 1 \text{ vector of ones},$$

where Diag $\{d_1, \ldots, d_n\}$ stands for a $n \times n$ diagonal matrix with diagonal entries $d_1, \ldots, d_n$. This form of the likelihood is actually a lot easier to compute than that given in (2.7). However, evaluating the likelihood is still not a simple task. For this, we rely on the so-called *forward-backward algorithm* developed by Baum *et*

*al.* (1970). Note, Leroux & Putterman (1992) observed that this algorithm can be unstable since it may converge to zero or diverge to infinity. All of these issues will be addressed in the next two subsections.

### 2.3.2 The Forward-Backward Algorithm

It should be clear that the evaluation of the likelihood function is of particular importance as most statistical procedures typically depend on it (for example, likelihood ratio testing and model selection). As was mentioned above, the likelihood can be computed using the forward-backward algorithm. We present here a simple version of the algorithm that was suggested by Baum *et al.* (1970), which is essentially a "purely forward" version of their algorithm. For this, let

$$\alpha_t(i) = P(y_1, ..., y_t, X_t = i|\boldsymbol{\phi}),$$

for $t = 1, \ldots, T$ and $i = 1, \ldots, k$. These are referred to as the *forward variables* and can be solved for recursively. Indeed, for $t = 1$, we have

$$\begin{aligned} \alpha_1(i) &= P(y_1, X_1 = i|\boldsymbol{\phi}) \\ &= P(X_1 = i|\theta) f(y_1|X_1 = i, \boldsymbol{\phi}) \\ &= \pi_i f(y_1|\theta_i), \end{aligned} \tag{2.9}$$

for $i = 1, \ldots, k$. Also, for $t = 2, 3, \ldots, T$, we can write

$$\begin{aligned}
\alpha_t(j) &= P(y_1, y_2, ..., y_t, X_t = j|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} P(y_1, y_2, ..., y_t, X_{t-1} = i, X_t = j|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} P(X_t = j, y_t|X_{t-1} = i, y_1, ..., y_{t-1}, \boldsymbol{\phi}) P(X_{t-1} = i, y_1, .., y_{t-1}|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} P(X_t = j, y_t|X_{t-1} = i, \boldsymbol{\phi}) P(X_{t-1} = i, y_1, ..., y_{t-1}|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} P(y_t|X_t = j, X_{t-1} = i, \boldsymbol{\phi}) P(X_t = j|X_{t-1} = i, \boldsymbol{\phi}) P(X_{t-1} = i, y_1, ..., y_{t-1}|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} f(y_t|X_t = j, \boldsymbol{\phi}) a_{ij} \alpha_{t-1}(i) \\
&= f(y_t|\theta_j) \sum_{i=1}^{k} \alpha_{t-1}(i) a_{ij}. \qquad (2.10)
\end{aligned}$$

Equations (2.9) and (2.10) together allow a recursion to be set up that will lead to obtaining $\alpha_t(i)$ for $i = 1, \ldots, k$ and $t = 1, \ldots, T$. First, the values of $\alpha_1(1), \ldots, \alpha_1(k)$ are calculated from (2.9). Then, $\alpha_t(1), \ldots, \alpha_t(k)$ are obtained from $\alpha_{t-1}(1), \ldots, \alpha_{t-1}(k)$ through equations (2.10) recursively for $t = 2, \ldots, T$. Note that equations (2.9) and (2.10) are known as the *forward equations*. The usefulness of the forward equations comes from the fact that, in terms of the forward variables, the likelihood can be written as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) &= f(y_1, y_2, ..., y_T|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} P(y_1, y_2, ..., y_T, X_T = i|\boldsymbol{\phi}) \\
&= \sum_{i=1}^{k} \alpha_T(i).
\end{aligned}$$

In principle, this last expression allows efficient computation of the likelihood. However, as was mentioned earlier, Leroux & Putterman (1992) observed that this method of evaluating the likelihood can be unstable because $\alpha_t(i)$ may be too small to be distinguishable from zero. One remedy to this instability is to rescale the $\alpha_t(i)$'s throughout the recursion by dividing them by $\sum_{i=1}^{k} \alpha_t(i)$. Other scaling techniques are available, see for instance, MacDonald & Zucchini (1997, p.79). We outline one such method in the next subsection.

### 2.3.3 Rescaling the Forward Equations

Evaluating the likelihood using the forward equations may suffer from the so-called underflow/overflow problem, that is, the likelihood may converge to zero or diverge to infinity as $t$ increases. To see why this is the case, note that the forward variable can also be expressed in the following vector form

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_1 \prod_{s=1}^{t} (\mathbf{D}(y_s)\mathbf{A}) \tag{2.11}$$

where

$$\boldsymbol{\alpha}_t = \{\alpha_t(1), \alpha_t(2), ..., \alpha_t(k)\},$$

$$\mathbf{D}(y_s) = Diag\{f(y_s|\theta_1), ..., f(y_s|\theta_k)\},$$

It can be seen from (2.11) that the forward variable contains a summation involving terms of the form

$$\left(\prod_{s=1}^{t-1} a_{x_s,x_{s+1}}\right)\left(\prod_{s=1}^{t} f(y_s|\theta_{x_s})\right).$$

Hence, in the case where $f(\cdot|\theta_{x_s})$ is a probability mass function, each term of the previous product is less than 1, and so, as $t$ gets large, $\boldsymbol{\alpha}_t$ converges to zero. On the other hand, if $f(\cdot|\theta_{x_s})$ is a highly concentrated density function, then each of the

$f(\cdot|\theta_{x_s})$ can be extremely large, possibly leading to $\boldsymbol{\alpha}_t$ diverging to infinity. One way of getting around this difficulty is by way of rescaling the forward variables as we go through the iterative process. To do this, define the first scaling coefficient as

$$c_1 = \left[\sum_{j=1}^{k} \alpha_1(j)\right]^{-1}.$$

Hence, at $t = 1$, the forward variables are rescaled by multiplying by $c_1$, that is,

$$\hat{\alpha}_1(i) = c_1\alpha_1(i), \qquad \text{for } i = 1, \ldots, k,$$

where $\hat{\alpha}$ is used to denote the rescaled coefficients. Then, applying (2.10) directly to the rescaled forward variables, we define at $t = 2$,

$$\begin{aligned}\alpha_2^*(j) &= f(y_2|\theta_j)\sum_{i=1}^{k}\hat{\alpha}_1(i)a_{ij} \\ &= c_1 f(y_2|\theta_j)\sum_{i=1}^{k}\alpha_1(i)a_{ij} \\ &= c_1\alpha_2(j).\end{aligned}$$

Now, let

$$\hat{\alpha}_2(i) = c_2\alpha_2^*(i),$$

where the second scaling coefficient is defined as

$$c_2 = \left[\sum_{j=1}^{k} \alpha_2^*(j)\right]^{-1}.$$

Obviously,

$$\hat{\alpha}_2(i) = c_1c_2\alpha_2(i),$$

for $i = 1, \ldots, k$. For general $t > 1$, using the same trick leads to

$$\begin{aligned}\alpha_t^*(j) &= f(y_t|\theta_j) \sum_{i=1}^{k} \hat{\alpha}_{t-1}(i) a_{ij} \\ &= \left( \prod_{s=1}^{t-1} c_s \right) f(y_t|\theta_j) \sum_{i=1}^{k} \alpha_{t-1}(i) a_{ij} \\ &= \left( \prod_{s=1}^{t-1} c_s \right) \alpha_t(j),\end{aligned}$$

and allows us to further define

$$\hat{\alpha}_t(j) = \alpha_t^*(j) \left[ \sum_i \alpha_t^*(i) \right]^{-1} = c_t \alpha_t^*(j) = \left( \prod_{s=1}^{t} c_s \right) \alpha_t(j). \tag{2.12}$$

As this last result is valid for $t = T$, it is possible to compute the likelihood from

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = \sum_{i=1}^{k} \alpha_T(i) = \left( \prod_{s=1}^{T} c_s \right)^{-1} \sum_{i=1}^{k} \hat{\alpha}_T(i) = \left( \prod_{s=1}^{T} c_s \right)^{-1},$$

since, conveniently,

$$\sum_{i=1}^{k} \hat{\alpha}_T(i) = \frac{\sum_{i=1}^{k} \alpha_T^*(i)}{\sum_{j=1}^{k} \alpha_T^*(j)} = 1,$$

which should be clear upon looking at (2.12) once again. Obviously this implies that the log-likelihood can be computed as

$$\ell(\boldsymbol{\phi}|\mathbf{y}) = -\sum_{t=1}^{T} \log\, c_t,$$

and thus, involves only the scaling constants! Since the log-likelihood is computed as a sum of the log scaling factors, this will avoid the underflow/overflow problems. We point out that the rescaling method presented here was originally suggested by Rabiner (1989).

## 2.4 Reconstructing the Hidden States

In many practical applications, estimating the hidden states is the central question of interest. For some applications, finding an individually most likely state is sufficient, while other applications may require that a most likely sequence of states be estimated. For example, in gene sequencing, $X_t$ could represent a DNA base element whereas $\mathbf{X} = (X_1, ..., X_T)$ would then represent a gene. The implementation of the first approach is done by calculating $P(X_t = i|y_1, ..., y_T, \boldsymbol{\phi})$ for each $t$ and $i$, and determining which state $i$ gives the highest probability for each $t$. The second approach requires finding a sequence of states $\hat{\mathbf{X}} = (\hat{X}_1, \hat{X}_2, ..., \hat{X}_T)$ maximizing the joint probability $P(x_1, x_2, ..., x_T|y_1, y_2, ..., y_T, \boldsymbol{\phi})$. We now briefly outline how these problems can be approached.

### 2.4.1 Estimating Individually Most Likely States

In order to find individually most likely states, we define for $i = 1, \ldots, k$

$$\beta_t(i) = f(y_{t+1}, ..., y_T|X_t = i, \boldsymbol{\phi}),$$

for $t = 1, \ldots, T-1$, and, by convention, $\beta_T(i) = 1$.

Note that

$$\begin{aligned}
\beta_t(i) &= f(y_{t+1}, ..., y_T | X_t = i; \boldsymbol{\phi}) \\
&= \frac{1}{P(X_t = i|\boldsymbol{\phi})} P(y_{t+1}, ..., y_T, X_t = i|\boldsymbol{\phi}) \\
&= \sum_{j=1}^{k} \frac{1}{P(X_t = i|\boldsymbol{\phi})} P(y_{t+1}, ..., y_T, X_t = i, X_{t+1} = j|\boldsymbol{\phi}) \\
&= \sum_{j=1}^{k} \frac{1}{P(X_t = i|\boldsymbol{\phi})} f(y_{t+1}, ..., y_T | X_t = i, X_{t+1} = j, \boldsymbol{\phi}) P(X_t = i, X_{t+1} = j|\boldsymbol{\phi}) \\
&= \sum_{j=1}^{k} f(y_{t+1}, ..., y_T | X_{t+1} = j, \boldsymbol{\phi}) P(X_{t+1} = j | X_i = i, \boldsymbol{\phi}) \\
&= \sum_{j=1}^{k} f(y_{t+1} | X_{t+1} = j, \boldsymbol{\phi}) f(y_{t+2}, ..., y_T | X_{t+1} = j, \boldsymbol{\phi}) a_{ij} \\
&= \sum_{j=1}^{k} a_{ij} f(y_{t+1} | \theta_j) \beta_{t+1}(j),
\end{aligned}$$

for $t = 1, \ldots, T-1$. Notice, however, that the $\beta$'s are computed recursively going backwards, that is from $T$ down to $t$, and thus are referred to as the *backward variables.*

Now, finding the individually most likely state is to find a single state $X_t$ that maximizes $P(X_t = i|y_i, ..., y_T, \boldsymbol{\phi})$, that is,

$$\begin{aligned}
\hat{X}_t &= \arg\max_i P(X_t = i|y_1, ..., y_T, \boldsymbol{\phi}) \\
&= \arg\max_i P(y_1, \ldots, y_T, X_t = i|\boldsymbol{\phi}), \qquad (2.13)
\end{aligned}$$

this last equality holding because $f(y_1, ..., y_T|\boldsymbol{\phi}) = \mathcal{L}(\boldsymbol{\phi}|\mathbf{y})$ is a constant with respect

to $X_t$. However, note that

$$\begin{aligned} P(y_1, ..., y_T, X_t = i|\boldsymbol{\phi}) &= f(y_1, ..., y_T|X_t = i, \boldsymbol{\phi})P(X_t = i|\boldsymbol{\phi}) \\ &= f(y_1, ..., y_t|X_t = i, \boldsymbol{\phi})f(y_{t+1}, ..., y_T|X_t = i, \boldsymbol{\phi})P(X_t = i|\boldsymbol{\phi}) \\ &= P(y_1, ..., y_t, X_t = i|\boldsymbol{\phi})f(y_{t+1}, ..., y_T|X_t = i, \boldsymbol{\phi}) \\ &= \alpha_t(i)\beta_t(i). \end{aligned} \tag{2.14}$$

Thus the estimator of the individually most likely state can be found to be

$$\begin{aligned} \hat{X}_t &= \arg\max_i f(y_1, ..., y_T, X_t = i|\boldsymbol{\phi}) \\ &= \arg\max_i \alpha_t(i)\beta_t(i), \end{aligned}$$

which can be easily identified after having gone through the full forward-backward recursion. Note that the caluculation of the backward variables can also lead to underflow/overflow problems. However, remedial measures do exist. (For instance, see Devijver, 1985.)

### 2.4.2 Estimating the Most Likely Sequence of States

In light of the argument leading to (2.13), estimating the most likely sequence of states, that is, finding $\hat{\mathbf{X}} = (\hat{X}_1, ..., \hat{X}_T)$ that maximizes $P(x_1, ..., x_T|y_1, ..., y_T; \boldsymbol{\phi})$ is equivalent to maximizing the joint distribution $P(x_1, ..., x_T, y_1, ..., y_T|\boldsymbol{\phi})$. Thus, the problem of finding the most likely sequence of states becomes the problem of maximizing the complete data likelihood (*i.e.* without averaging on the hidden states). Even though the notation could be somewhat confusing, it should be clear to the reader that the most likely sequence of states $\hat{\mathbf{X}}$ is not made from the individually most likely states.

A technique based on a recursive procedure, called the Viterbi algorithm (*cf.* Viterbi, 1967) is designed to solve this problem. To carry out the Viterbi algorithm, the following two quantities are needed,

$$\delta_t(j) = \max_{x_1,...,x_{t-1}} P(x_1, ..., x_{t-1}, X_t = j, y_1, ..., y_t|\boldsymbol{\phi}),$$

and

$$\psi_t(j) = \arg\max_i[\delta_{t-1}(i)a_{ij}].$$

These definitions imply that $\delta_t(j)$ is the highest density along a single path leading to $X_t = j$, and $\psi_t(j)$ stores each maximal stage as the algorithm progresses. It can be shown that $\delta_t(j)$ satisfies the following recursion,

$$\delta_t(j) = f(y_t|\theta_j) \max_i[\delta_{t-1}(i)a_{ij}] \tag{2.15}$$

for $t = 2, ..., T$. To solve for $\hat{\mathbf{X}} = (\hat{X}_1, ..., \hat{X}_T)$, with initialization $\delta_1(i) = \pi_i f(y_1|\theta_i)$, run $\delta_t(j)$ recursively, storing the argument maximizing $\delta_t(j)$ in $\psi_t(j)$. Then, choose

$$\begin{aligned}\hat{X}_t &= \psi_{t+1}(\hat{X}_{t+1}) \\ &= \arg\max_i[\delta_t(i)a_{i\hat{X}_{t+1}}],\end{aligned}$$

for $t = T-1, T-2, ..., 1$. Note that, the computation of the Viterbi algorithm is similar to the forward procedure, except for the maximization of equation (2.15) replacing the summation of equation (2.10). In additon, the Viterbi algorithm can also suffer from underflow/overflow problems. For remedies to this problem, readers are referred to Rabiner (1989) and Scott (2002).

## 2.5 Parameter Estimation using the EM algorithm

Traditionally, parameter estimation in HMM's is done through maximum likelihood. In practice, the Maximum Likelihood Estimator (MLE) is often solved for using the Expectation-Maximization (EM) algorithm (*cf.* Dempster *et al.*, 1977), which is known to handle problems where missing data have occurred. The adaptation of the EM algorithm to the context of HMM's is credited to Baum *et al.* (1970). The basic principle of the EM algorithm is to iterate between the E-step and the M-step. In the E-step, the conditional expectation of the unobserved states is computed given the parameters. This is done by using the forward-backward algorithm. Then, in the M-step, the likelihood function is maximized given the data and the expected states. The implementation of the EM algorithm always leads to explicit formulae for the transition probabilities at the M-step if the parametric family $f(y|\theta)$ under consideration belongs to an exponential family. In specific cases, like in the case of the Poisson family with unknown parameter $\theta$, the M-step is also explicit in $\theta$.

For a time-homogeneous HMM with unknown initial probabilities, the EM algorithm can be carried out as follows. The complete-data likelihood function (that is, if the hidden states were also observed) is

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = \pi_{x_1} f(y_1|\theta_{x_1}) \prod_{t=2}^{T} a_{x_{t-1},x_t} f(y_t|\theta_{x_t}),$$

so that the complete-data log-likelihood, denoted $\ell(\boldsymbol{\phi})$, is

$$\begin{aligned}\ell(\boldsymbol{\phi}|\mathbf{y}) &= \log \pi_{x_1} + \sum_{t=1}^{T} \log f(y_t|\theta_{x_t}) + \sum_{t=2}^{T} \log a_{x_{t-1},x_t} \\ &= \sum_{i=1}^{k} u_i(1) \log \pi_i + \sum_{i=1}^{k} \sum_{t=1}^{T} u_i(t) f(y_t|\theta_i) + \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{t=2}^{T} \nu_{ij}(t) \log a_{ij},\end{aligned}$$

where $u_i(t)$ and $\nu_{ij}(t)$ are indicator functions defined as

$$u_i(t) = \begin{cases} 1 & \text{if } X_t = i, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\nu_{ij}(t) = \begin{cases} 1 & \text{if } X_{t-1} = i \text{ and } X_t = j, \\ 0 & \text{otherwise.} \end{cases}$$

The E-step replaces $u_i(t)$ and $\nu_{ij}(t)$ by their conditional expectations given by

$$\hat{u}_i(t) = E(u_i(t)|y_1, ..., y_T, \boldsymbol{\phi}) = P(X_t = i|y_1, ..., y_T, \boldsymbol{\phi}),$$

and

$$\hat{\nu}_{ij}(t) = E(\nu_{ij}(t)|y_1, ..., y_T, \boldsymbol{\phi}) = P(X_{t-1} = i, X_t = j|y_1, ..., y_T, \boldsymbol{\phi}).$$

These quantities can be computed using the forward-backward recursions. Indeed, we have that

$$\begin{aligned} \hat{u}_i(t) &= P(X_t = i|y_1, ..., y_T, \boldsymbol{\phi}) \\ &= \frac{P(y_1, .., y_T, X_t = i|\boldsymbol{\phi})}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})}, \end{aligned}$$

from equation (2.14), and that

$$\begin{aligned}
\hat{\nu}_{ij}(t) &= P(X_{t-1}=i, X_t=j|y_1,...,y_T,\boldsymbol{\phi}) \\
&= \frac{P(X_{t-1}=i, X_t=j, y_1,...,y_T|\boldsymbol{\phi})}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\
&= \frac{f(y_1,...,y_T|X_{t-1}=i, X_t=j,\boldsymbol{\phi})P(X_{t-1}=i, X_t=j|\boldsymbol{\phi})}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\
&= \frac{f(y_1,...,y_{t-1}|X_{t-1}=i,\boldsymbol{\phi})f(y_t,...,y_T|X_t=j,\boldsymbol{\phi})P(X_{t-1}=i|\boldsymbol{\phi})a_{ij}}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\
&= \frac{P(y_1,...,y_{t-1}, X_{t-1}=i|\boldsymbol{\phi})f(y_t,...,y_T|X_t=j,\boldsymbol{\phi})a_{ij}}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\
&= \frac{\alpha_{t-1}(j)f(y_t|X_t=i,\boldsymbol{\phi})f(y_{t+1},...,y_T|X_t=j,\boldsymbol{\phi})a_{ij}}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})} \\
&= \frac{\alpha_{t-1}(i)\beta_t(j)f(y_t|\theta_i)a_{ij}}{\mathcal{L}(\boldsymbol{\phi}|\mathbf{y})}.
\end{aligned}$$

The M-step maximizes the expectation of the log-likelihood. For this, the maximizing values for $\pi_i$, and $a_{ij}$ at iteration $n+1$ are taken to be

$$\pi_i^{(n+1)} = \hat{u}_i^{(n)}(1),$$

and

$$a_{ij}^{(n+1)} = \frac{\sum_{t=2}^{T}\hat{\nu}_{ij}^{(n)}(t)}{\sum_{t=2}^{T}\sum_{j}^{N}\hat{\nu}_{ij}^{(n)}(t)}.$$

In the case where the hidden Markov chain is assumed to be stationary, $\pi_i^{(n+1)}$ should instead be obtained by finding the unique distribution $\boldsymbol{\pi}$ satisfying

$$\boldsymbol{\pi} = (\mathbf{A}^{(n+1)})^T\boldsymbol{\pi}.$$

(See section 2.1.)

In addition, if the distributional form of $f(y_t|\theta_{x_t})$ is known, then $\theta_{x_t}$ can also be solved at the $M$ step. This iterative process is repeated until convergence is reached.

Practically, when changes in parameter estimates become negligible, the estimation procedure is considered to have converged.

Note that there are some well known shortcomings of the EM algorithm. Firstly, convergence is often slow. Secondly, due to the large number of parameters, the likelihood surface usually contains many local maxima, so that the algorithm does not necessarily locate the global maximum. In such a case, the algorithm may very well converge to a local maximum or even a saddle point of the log-likelihood function. In particular, starting values are very important for using this algorithm in the HMM setup. Leroux & Puterman (1992) suggested a method to select reasonable starting values.

Alternatives to the use of the EM algorithm do exist. For instance, the direct numerical maximization suggested by MacDonald & Zucchini (1997, Chapter 2) uses a derivative free method known as the downhill simplex algorithm to locate the maximum likelihood estimates.

Different approaches to the estimation of HMM parameters are also possible. In particular, Bayesian methodologies have been considered. Working from that perspective, the hidden states can be treated as unknown parameters and simulated along side with the other parameters of the model by Gibbs sampling methods. This idea is known as data augmentation and is presented in the next chapter. Finally, we present in Chapter 4 an alternative approach to Gibbs sampling that allows Bayesian inference for HMM's without using data augmentation.

# Chapter 3

# Bayesian Inference for Hidden Markov Models

## 3.1 The Idea of Bayesian Inference

The major difference between the so-called frequentist and Bayesian approaches to inference is that from the frequentist perspective, the parameters are considered as fixed constants, to be estimated from the observations whereas, from the Bayesian perspective, the parameters and observations are put on the same conceptual level. That is, the parameters are treated as random quantities, where the uncertainty on the parameters can be modeled through a probability distribution.

Now, let $x$ be the observed data and $\theta$ be the vector of parameters. The sampling distribution, or likelihood, is denoted $f(x|\theta)$. The marginal distribution of the parameters is denoted as $\pi(\theta)$, and is referred to as the *prior distribution.* This prior distribution is used to model what is known about the problem at hand (including the absence of any information, be it the case) prior to the gathering of data. (See Robert 2001, Chapter 3 for more details.) Based on $f(x|\theta)$ and $\pi(\theta)$, we can derive the joint distribution of $(X, \theta)$, given by

$$h(x, \theta) = f(x|\theta)\pi(\theta),$$

and the marginal density of $X$ as

$$\begin{aligned} m(x) &= \int_{\Theta} h(x, \theta) d\theta \\ &= \int_{\Theta} f(x|\theta)\pi(\theta) d\theta, \end{aligned}$$

where $\Theta$ denotes the parameter space.

By applying Bayes' theorem, we can obtain the conditional distribution of the parameter vector $\theta$ given the observed data $x$ as

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

All Bayesian inference is driven by the above conditional distribution, called the *posterior distribution* of $\theta$. The posterior distribution can be thought of as a mechanism that takes, as an input, the information that one has on the parameter $\theta$ prior to the experiment being conducted, and combines it with the information contained in the observed data $x$ as an input to produce a valuable output. Generally speaking, prior distributions can be classified into two categories: *informative* and *noninformative*. When a reasonable amount of information on the parameters is available, an informative prior can be used. When information on the parameters is too vague or unavailable, then one can turn to a noninformative prior, usually a uniform, very flat and/or heavy tailed distribution. In using such a noninformative prior, it is usually hoped that the prior distribution will have minimal effect on the posterior distribution and the resulting inference.

In particular, Bayesian inference is often based on the posterior expectation of some function $f(\theta)$ given by

$$E[f(\theta)|x] = \frac{\int f(\theta)f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta}, \tag{3.1}$$

and/or the maximization of the posterior distribution. In many practical applications, computing the above posterior expectation and maximizing the posterior distribution may be very difficult because no analytical expression of this posterior

distribution is available. Furthermore, when an explicit expression is available for the posterior distribution, posterior expectations like (3.1) might still not be available in closed form. Finally, there are cases where everything can be obtained in closed form but where computation may still be intractable due to a large sample size. This is the so-called "information paradox". (See Robert 2001, Chapter 6, p.319 for more details.) This is the case, for instance, with mixtures where the posterior distribution takes into account all the possible partitions of the sample. To address these issues, many techniques have been developed for approximating (3.1), one of which is known as the Markov Chain Monte Carlo method (*cf.* Tanner, 1993 and Evans & Schwantz, 2000). For a complete treatment of the Bayesian approach to inference, including the role, impact and selection of prior distributions, we refer the reader to the books of Berger (1985), Robert (2001), Carlin and Louis (2000), and for a manuscript more geared towards application, Gelman *el al.* (2003).

## 3.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a method that allows for (approximately) sampling from posterior distributions, and thus, can be used for approximating posterior expectations as in (3.1). The idea of MCMC is to construct a Markov chain on the parameter space $\Theta$, which is irreducible and aperiodic, and whose stationary distribution is the posterior $\pi(\theta|x)$. Then, this Markov chain is run for a long time to generate a sequence of identically distributed (but not independent) random variables or vectors. In practice, some large value $M$ is selected and the random variables $\theta^{(M+1)}, \theta^{(M+2)}, \ldots, \theta^{(M+N)}$ are used to mimic random sampling from the

posterior $\pi(\theta|x)$. These values can be used to approximate the expected value of some function $f$ of the parameter vector with respect to the posterior distribution. Specifically, we consider

$$E[f(\theta)|x] \approx \frac{1}{N}\sum_{k=1}^{N} f(\theta^{(M+k)}).$$

In this context $M$ is referred to as the *burn-in period*, that is, the first $M$ randomly generated values are discarded. This is done to reduce the influence of the starting value, and to ensure that the randomly generated values come from a distribution that is "close" to the stationary distribution. The convergence of the above empirical average to the proper expectation is ensured by the Ergodic Theorem. Thus, if one can construct a Markov chain having $\pi(\theta|x)$ as its stationary distribution, then inference based on $\pi(\theta|x)$ can be done by simulating realizations from this Markov chain. A widely applicable method of constructing such a Markov chain, due to Gelfand and Smith (1990), is known as Gibbs sampling.

The main idea behind Gibbs sampling is to sequentially generate random samples from many low-dimensional conditional distributions in order to approximate sampling from a high-dimensional (and possibly more complex) joint distribution. To illustrate how this is done, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ be a random vector, such that the conditional distribution of $\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_p$ is available for all $i = 1, 2, \ldots, p$. Note that $\theta_1, \theta_2, \ldots, \theta_p$ can themselves be vectors, although the simplest implementation of the method usually considers univariate components. Then, given an arbitrary starting point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$, the iteration scheme of the Gibbs sampler can be summarized in the following manner.

For $m = 1, 2, ...,$ draw

$$\begin{aligned}
\theta_1^{(m)} &\sim f(\theta_1|\theta_2^{(m-1)}, \theta_3^{(m-1)}, \ldots, \theta_p^{(m-1)}) \\
\theta_2^{(m)} &\sim f(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)}, \ldots, \theta_p^{(m-1)}) \\
\theta_3^{(m)} &\sim f(\theta_3|\theta_1^{(m)}, \theta_2^{(m)}, \theta_4^{(m-1)} \ldots, \theta_p^{(m-1)}) \\
&\vdots \\
\theta_p^{(m)} &\sim f(\theta_p|\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_{p-1}^{(m)}).
\end{aligned}$$

The above steps are to be repeated many times to ensure the Gibbs sampler has converged (or is close enough) to the distribution of interest. Typically, a reasonably long burn-in period would be considered and only the generated vectors $\theta^{(M+1)}, \theta^{(M+2)}, \ldots$ (for some large $M$) are used at the next stage of the analysis. The vectors $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots$ can be shown to form a Markov chain and, under mild conditions, the joint distribution of $(\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_p^{(m)})$ can be shown to converge geometrically to the joint distribution of $(\theta_1, \theta_2, \ldots, \theta_p)$ as $m \to \infty$. Typically, this allows the user to generate random vectors $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ by effectively sampling from univariate conditional distributions, making the use of the technique fairly straightforward. For more on advantages and potential pitfalls of Gibbs sampling, see Gelfand & Smith (1990), Casella & George (1992) and Fu & Wang (2002). We also refer the readers to the monograph of Robert & Casella (2004) for an in-depth discussion of MCMC and Gibbs sampling.

## 3.3 Bayesian inference for HMM's with a Known Number of Hidden States

In this section, we discuss Bayesian inference for HMM's using Gibbs sampling. In the HMM setup, recall that the likelihood is

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = P(y_1, ..., y_T|\boldsymbol{\phi}) = \sum_{x_1=1}^{k} \cdots \sum_{x_T=1}^{k} \pi_{x_1} f(y_1|\theta_{x_1}) \prod_{t=2}^{T} a_{x_{t-1},x_t} f(y_t|\theta_{x_t}),$$

when the number of hidden states is known to be $k$. Given a prior distribution $\pi$ on the parameter vector $\boldsymbol{\phi} = (\boldsymbol{\theta}, \mathbf{A})$, the posterior distribution of $\boldsymbol{\phi}$ given the observed data $\mathbf{y}$ involves a sum of $k^T$ terms. This makes it very difficult to sample directly from the posterior distribution of $\boldsymbol{\phi}$, even in the case of a moderate sample size. (Note, however, that this is the approach we will be taking later, using the sampling method of Fu & Wang, 2002). The usual way around this difficulty is to sample using the idea of data augmentation. The principle behind data augmentation, in the current setup, is that by augmenting the observed data by generating the hidden states (thus treating them as missing data), the complete-data likelihood $P(\mathbf{x}, \mathbf{y}|\boldsymbol{\phi})$ has a simple form, which in turn produces a simple posterior distribution $\pi(\boldsymbol{\phi}|\mathbf{x}, \mathbf{y})$, so that, using a proper prior, all the parameters can be simulated from $\pi(\boldsymbol{\phi}|\mathbf{x}, \mathbf{y})$. For this to work, the hidden states are sampled from the conditional density $P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$.

In the HMM setup, the implementation of the Gibbs sampling thus alternately simulates $\mathbf{x}$ and $\boldsymbol{\phi}$ by iterating between the following two steps:

1. simulate

$$\mathbf{x}^{(m+1)} \sim P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}^{(m)}),$$

2. simulate

$$\boldsymbol{\phi}^{(m+1)} \sim \pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{x}^{(m+1)}).$$

Iterating between the above two steps produces a random sequence $\{(\boldsymbol{\phi}, \mathbf{x})^{(m)} : m = 1, 2, ...\}$ that forms a Markov chain, which under some general conditions will admit the posterior of interest $\pi(\boldsymbol{\phi}, \mathbf{x}|\mathbf{y})$ as its stationary distribution. In addition, the sequence $(\mathbf{x}^{(\mathbf{m})})$ also forms a Markov chain with transition kernel density

$$P(\mathbf{x}, \mathbf{x}') = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{x}^{(\mathbf{m})} = \mathbf{x}) \pi(\mathbf{x}^{(\mathbf{m+1})} = \mathbf{x}'|\boldsymbol{\phi}, \mathbf{y}) \mathbf{d}\boldsymbol{\theta}.$$

Under the assumption that the state-space of the hidden chain is finite (keep in mind we have assumed $\mathcal{S}$=$\{1, 2, \cdots, k\}$), many convergence results can be easily established, and tranferred to the sequence $\boldsymbol{\phi}^{(m)}$ by the so-called duality principle (see Robert *et al.*, 1993 and Diebolt & Robert, 1994). These results include geometric convergence, $\phi$ mixing, and a central limit theorem. We now take a closer look at each of the two steps used in the iterative scheme outlined earlier.

### 3.3.1 Simulating from $\pi(\boldsymbol{\phi}|\mathbf{y}, \mathbf{x})$

Let $\pi(\boldsymbol{\theta})$ and $\pi(\mathbf{A})$ be respectively the prior distribution of the distributional parameters and the prior distribution of the transition probability matrix. We here assume as is commonly done in the literature, that $\boldsymbol{\theta}$ and $\mathbf{A}$ are *a priori* independent. Then, when the observed $\mathbf{y}$ is augmented by the vector of hidden states $\mathbf{x}$, the HMM model becomes hierarchical with the following structure

$$\mathbf{y}|\mathbf{X} = \mathbf{x} \sim f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}),$$

$$\mathbf{x} \sim P(\mathbf{x}|\mathbf{A}),$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}),$$

$$\mathbf{A} \sim \pi(\mathbf{A}),$$

and, the joint posterior distribution of $\boldsymbol{\phi} = (\boldsymbol{\theta}, \mathbf{A})$ given $\mathbf{y}$ and $\mathbf{x}$ satisfies

$$\begin{aligned}\pi(\boldsymbol{\phi}|\mathbf{x}, \mathbf{y}) &\propto f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) P(\mathbf{x}|\mathbf{A}) \pi(\boldsymbol{\theta}) \pi(\mathbf{A}) \\ &\propto \pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \pi(\mathbf{A}|\mathbf{x}). \end{aligned} \tag{3.2}$$

So that $\boldsymbol{\theta}$ and $\mathbf{A}$ are independent *a posteriori*. Note that $\mathbf{A}$ is also *a posteriori* independent of the observed data $\mathbf{y}$. These independence properties make generating from $\pi(\boldsymbol{\phi}|\mathbf{x}, \mathbf{y})$ very simple in many cases. Indeed, we can write

$$\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto \left( \prod_{t=1}^{T} f(y_t|\theta_{x_t}) \right) \pi(\boldsymbol{\theta}),$$

so that *a priori* independence of $\theta_1, \ldots, \theta_k$ and the use of conjugate priors lead to $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ having a form that is easy to simulate from. Similarly, we have that

$$\pi(\mathbf{A}|\mathbf{x}) \propto \left( \prod_{t=1}^{T} a_{x_{t-1}, a_{x_t}} \right) \pi(\mathbf{A}),$$

when assuming a start in a fixed state $x_0$ (as done by Robert *et al.*, 1993), which can also be written as

$$\pi(\mathbf{A}|\mathbf{x}) \propto \left( \prod_{i=1}^{k} \prod_{j=1}^{k} a_{ij}^{n_{ij}} \right) \pi(\mathbf{A}), \tag{3.3}$$

where $n_{ij} = \sum_{t=1}^{T} I(x_{t-1} = i, x_t = j)$ is the number of transitions from state $i$ to state $j$. Now let $\mathbf{A_i}$ denote the $i^{th}$ row of $\mathbf{A}$ and assume prior independence between $\mathbf{A_1}, \ldots, \mathbf{A_k}$. Given the multinomial form of the first term on the right hand side of (3.3), a family of conjugate prior distributions for $\mathbf{A_i}$ is the family of Dirichlet distributions defined on the $k$-dimensional simplex, with density

$$\pi(\mathbf{A_i}) \propto \prod_{j=1}^{k} a_{ij}^{\alpha_{ij}-1} \mathbf{I}_{\{\sum_j a_{ij}=1\}}, \tag{3.4}$$

which we denote $\mathbf{A_i} \sim D_k(\alpha_{i1}, \ldots, \alpha_{ik})$, with $\alpha_{ij} > 0$. From (3.3) and (3.4), we get posterior independence and the following posterior distributions

$$\mathbf{A_i}|\mathbf{x} \sim \mathcal{D}_k(\alpha_{i1} + n_{i1}, ..., \alpha_{ik} + n_{ik}) \tag{3.5}$$

for each row $i = 1, \ldots, k$ of $\mathbf{A}$. This makes simulating from $\pi(\mathbf{A}|\mathbf{x})$ very easy when working with independent Dirichlet priors for the rows of $\mathbf{A}$.

### 3.3.2 Simulation of the Hidden States

When simulating $\mathbf{x}$ from

$$P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) \propto \pi_{x_1} f(y_1|\theta_{x_1}) \prod_{t=2}^{T} a_{x_{t-1},x_t} f(y_t|\theta_{x_t}), \tag{3.6}$$

the dependence structure between hidden states must be taken into account. There are two methods available for simulating $\mathbf{x}$. One possibility is working from a sequence of univariate conditional distributions (*cf.* Robert *et al.* 1993) to simulate each of the $T$ components of $\mathbf{x}$ individually. This is done through a sequential univariate update of $P(x_t|\mathbf{x_{\backslash t}}, \mathbf{y}, \boldsymbol{\phi})$, for $t = 1, 2, ..., T$, where $\mathbf{x_{\backslash t}} = (x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T)$ corresponds to the sequence of hidden states with component $t$ omitted. It turns out that, for $t = 1, 2, \cdots, T-1$,

$$\begin{aligned} P(x_t|\mathbf{x_{\backslash t}}, \mathbf{y}, \boldsymbol{\phi}) &= P(x_t|x_{t-1}, x_{t+1}, \mathbf{y}, \boldsymbol{\phi}) \\ &= \frac{a_{x_{t-1},x_t} f(y_t|\theta_{x_t}) a_{x_t,x_{x+1}}}{\sum_{i=1}^{k} a_{x_{t-1},i} f(y_t|\theta_i) a_{i,x_{t+1}}}, \end{aligned} \tag{3.7}$$

and that

$$\begin{aligned} P(x_T|\mathbf{x_{\backslash T}}, \mathbf{y}, \boldsymbol{\phi}) &= P(x_T|x_{T-1}, \mathbf{y}, \boldsymbol{\phi}) \\ &= \frac{a_{x_{T-1},x_T} f(y_T|\theta_{x_T})}{\sum_{i=1}^{k} a_{x_{t-1},i} f(y_T|\theta_i)}. \end{aligned} \tag{3.8}$$

The realization of the hidden states can then be simulated, $x_t$ being obtained from (3.7), for $t < T$, and from (3.8) when $t = T$.

An alternative simulation method for generating a realization of the hidden states is to simulate $\mathbf{x}$ as a full sequence each time, that is, to sample $\mathbf{x}$ directly from the joint distribution $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$. Note that $P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$ can be decomposed in the following way

$$\begin{aligned} P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) &= P(x_T|\mathbf{y}, \boldsymbol{\phi})P(x_{T-1}|x_T, \mathbf{y}, \boldsymbol{\phi})\ P(x_{T-2}|x_{T-1}, \mathbf{y}, \boldsymbol{\phi}) \cdots P(x_1|x_2, \boldsymbol{\phi}) \\ &= P(x_T|\mathbf{y}, \boldsymbol{\phi}) \prod_{t=1}^{T-1} P(x_t|\mathbf{y}, x_{t+1}, \boldsymbol{\phi}), \end{aligned}$$

by simply making use of the general multiplication rule for conditional probability. Hence, $\mathbf{x}$ is generated by first drawing $x_T$ from $P(x_T|\mathbf{y}, \boldsymbol{\phi})$, and then by sequentially simulating $x_t$ going backwards making use of $P(x_t|x_{t+1}, \boldsymbol{\phi})$. For full details, readers are referred to Chib (1996). Note, however, that generating the $\mathbf{x}$ vector directly from its joint distribution can speed up the convergence of the Gibbs sampler (*cf.* Chib, 1996 and Scott, 2002).

## 3.4 Bayesian Inference for HMM's with an Unknown Number of Hidden States

So far, we have considered the MCMC algorithm that is only suitable for a fixed number of hidden states $k$, and thus only valid for a fixed number of parameters. However, in many practical applications, inference for a fixed $k$ is too restrictive since $k$ is itself unknown and a quantity of greatest interest. Assuming that $k$ varies between $1 \leq k \leq K$, and letting $\mathcal{L}_k(\boldsymbol{\phi}^k|\mathbf{y})$ denote the likelihood conditional on there

being exactly $k$ hidden states, the full likelihood becomes

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = \sum_{l=1}^{K} \mathbf{I}(k=l)\mathcal{L}_l(\boldsymbol{\phi}^l|\mathbf{y}) = \mathcal{L}_k(\boldsymbol{\phi}^k|\mathbf{y}), \tag{3.9}$$

where $\boldsymbol{\phi} = (k, \boldsymbol{\phi}^1, \boldsymbol{\phi}^2, \ldots, \boldsymbol{\phi}^K)$, and includes the parameters involved with each possible number of hidden states.

### 3.4.1 Models with a Random Number of Hidden States

The Bayesian way to tackle the difficulties linked with having an unknown number $k$ of hidden states consists of assuming $k$ is random, that is, to treat $k$ just like any other parameter. Then, given the full likelihood (3.9), and a prior distribution on $k$, the posterior distribution of the complete parameter vector is of the form

$$\pi(\boldsymbol{\phi}|\mathbf{y}) = \pi(\boldsymbol{\phi}^k|\mathbf{y})P(k).$$

Inference on $k$ is based on its marginal posterior distribution, which is easily approximated by

$$\begin{aligned} P(k=i|\mathbf{y}) &= E[\mathbf{I}(k=i)|\mathbf{y}] \\ &\approx \frac{1}{N}\sum_{j=1}^{N} \mathbf{I}(k^{(M+j)}=i), \end{aligned} \tag{3.10}$$

for $i = 1, 2, \ldots, K$ and where $k^{(M+j)}$ are values obtained through Gibbs sampling or some other form of MCMC simulation method and using a burn-in period of $M$ steps. Note that the dimension of $\boldsymbol{\phi}^k = (\boldsymbol{\theta}^k, A^k)$ varies with $k$, so that the usual MCMC described in the previous section is not applicable to HMM's with an unknown number of hidden states. More sophisticated methods are required. One such method, called reversible jump Markov chain Monte Carlo is briefly discussed next.

### 3.4.2 Reversible Jump MCMC

Reversible jump MCMC was first introduced by Green (1995) and applied to multiple change point problems in one and two dimensions. It is a sampling algorithm that allows the dimension of the parameter vector to vary within the sampling process. Richardson and Green (1997) applied this methodology to univariate normal mixture models with an unknown number of components. More recently, Robert *et al.* (2000) considered this methodology and applied it in the context of HMM's. In particular, their version of reversible jump MCMC augmented the Gibbs sampler with the following two steps:

1. splitting a component (or hidden state) into two, or combining two into one,

2. birth or death of an empty component.

We refer the reader to Robert & Casella (2004, Chapter 11) for more details.

It should be noted that although reversible jump MCMC provides a way of generating samples with a varying number of components, its algebraic complexity makes it very difficult to implement and its use is mainly restricted to a limited number of experts. We will see in the next chapter that another sampling algorithm developed by Fu & Wang (2002) can be used to generate samples with a varying number of components $k$. This method only requires the knowledge of the joint distribution function up to a multiplicative constant and is easy to implement relative to the reversible jump MCMC algorithm.

# Chapter 4

# Discretization-Based Sampling in the Bayesian HMM Setup

In this chapter we describe the sampling method introduced by Fu & Wang (2002) and discuss its use in the context of Bayesian analysis of HMM's (see also Wang & Fu, 2007 and Xue *et al.*, 2005). The basic idea behind their method is to discretize the density function with respect to Lebesgue measure. Compared with the MCMC methods mentioned in Chapter 3, this method is dimension-free and non-iterative. The approach has many advantages over the Markov chain Monte Carlo (MCMC) or reversible jump MCMC. Firstly, it is easy to implement. Secondly, knowledge of the density function is required only up to a normalizing constant, and the full set of univariate conditional distributions related to the complete joint posterior distribution is not required. Thirdly, we can simultaneously approximate the posterior expectation, posterior mode (also known as generalized MLE, or sometimes, penalized MLE), and maximum likelihood estimate at no extra cost in terms of computational effort.

## 4.1 Sampling Algorithm

Suppose that we have a $d$-variate density function $f(x)$, either known completely or up to a multiplicative constant with support $S(f) \subseteq R^d$. Now, assume our objective is to generate a random sample of size $m$ from the significant region of

this density function. (Note that, the significant region is defined to be the region where $f(x) > 0$, thus the region with $f(x) \approx 0$ is regarded as a negligible region). Following the method of Fu & Wang (2002), this can be accomplished in the following five steps.

1. **Determination of the initial compact cover:**

   Lower and upper limits $0 < a_i^{(0)} < b_i^{(0)} < \infty$ for $i = 1, \ldots, d$ are determined that constitute an initial compact set

   $$S_0(f) = [a_1^{(0)}, b_1^{(0)}] \times [a_2^{(0)}, b_2^{(0)}] \times \cdots \times [a_d^{(0)}, b_d^{(0)}],$$

   which is large enough to cover the significant region of the density function $f(x)$. In the case where $f(x)$ has bounded support $S(f)$, the initial region is chosen to be $S_0(f) = S(f)$.

2. **Discretization:**

   With the initial compact cover determined in step 1, a set of independent random uniform points $D_n(f) = \{x_j \in S_0(f), j = 1, 2, \ldots, n\}$ is generated, where $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jd}\}$. For large $n$, the set of points $D_n(f)$ approximates the initial support $S_0(f)$.

3. **Contourization:**

   The points generated in step 2 are first rearranged such that $f(x_i) \geq f(x_j)$, if $i < j$. Then, for a given integer $N \in \mathbb{N}$, we partition $D_n(f)$ into $N$ contours,

   $$E_i = \{x_j : (i-1)l < j \leq il\}, \quad i = 1, 2, \ldots, N$$

   where $l = n/N$ is the number of points within each contour. Note that these

contours form a partition of the sample space $D_n(f)$, that is,

$$\cup_{i=1}^{N} E_i = D_n(f),$$

and

$$E_i \cap E_j = \emptyset,$$

for $i \neq j$.

4. **Sampling:**

   To sample $m$ points from $f(x)$, first sample $m$ contours with replacement from the set of contours $\{E_i,\ i = 1, 2, \ldots, N\}$ according to the contour probabilities $\{P_N(i),\ i = 1, 2, \ldots, N\}$ defined by

$$P_k(i) = \frac{\bar{f}_i}{\sum_{j=1}^{k} \bar{f}_j} \quad i = 1, 2, \ldots, k \tag{4.1}$$

   where

$$\bar{f}_i = \frac{1}{l} \sum_{x \in E_i} f(x),$$

   that is, $\bar{f}_i$ represents the average value of $f(x)$ over all points included in the $i^{th}$ contour. Let $m_i$ denote the number of occurrences of $E_i$ in the $m$ draws. It should be clear that $\sum_{i=1}^{N} m_i = m$. Then, randomly sample $m_i$ (again with replacement) points with equal probability from $E_i$. Denote $O_i$ as the set of points sampled within contour $E_i$, for each $i = 1, 2, \ldots, N$. Then $\cup_{i=1}^{N} O_i$ gives the desired sample of size $m$.

5. **Visualization:**

   To visualize the significant region and the negligible region of the sample space $S_0(f)$, plot histograms over all dimensions using the sample obtained in step 4

above. Alongside, calculate the minimum and maximum values of the sample for each coordinate. Denote

$$S_1(f) = [a_1^{(1)}, b_1^{(1)}] \times [a_2^{(1)}, b_2^{(1)}] \times \ldots \times [a_d^{(1)}, b_d^{(1)}],$$

where $a_i^{(1)} = \min_j(x_{ji})$ and $b_i^{(1)} = \max_j(x_{ji})$, for $i = 1, 2, \ldots, d$. If $S_1(f) = S_0(f)$, the sample generated from step 4 is accepted. Otherwise, replace $S_0(f)$ with $S_1(f)$ and repeat steps 2 to 5 until the significant region is formed.

Unlike Gibbs sampling, where samples are generated using marginal conditional distributions, the above procedure samples directly from the joint distribution $f(x)$, so that even complicated dependence structures are taken into account. Also, the procedure does not become more complex when $d$ gets large. Finally, the contourization step of the procedure provides information on the shape and location of the distribution $f(x)$, and the visualization step enables us to view the significant region.

## 4.2 Sampling from the Posterior in the HMM Setup

In this section, we adapt the sampling method of Fu & Wang (2002) described in Section 4.1 to the Bayesian HMM setup. The methodology we introduce here requires knowledge of the joint posterior distribution up to a multiplicative constant only, and allows us to omit using data augmentation, so that full knowledge of $P(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$ is not required. In addition, cases where the number of hidden states $k$ is unknown can be handled in a convenient way.

### 4.2.1 The Log-Posterior Density

Recall the likelihood of the HMM model is

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{y}) = \sum_{x_1=1}^{k} \cdots \sum_{x_T=1}^{k} \pi_{x_1} f(y_1|\theta_{x_1}) \prod_{t=2}^{T} a_{x_{t-1},x_t} f(y_t|\theta_{x_t}),$$

when the number of hidden states is known to be $k$ and where $\boldsymbol{\phi} = (\mathbf{A}, \boldsymbol{\theta})$. Thus, for a fixed number of hidden states $k$, assuming prior independence for $\mathbf{A}$ and $\boldsymbol{\theta}$, and given priors $\pi(\mathbf{A})$ on $\mathbf{A}$ and $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, the full posterior distribution is

$$\pi(\boldsymbol{\phi}|\mathbf{y}) \propto \mathcal{L}(\boldsymbol{\phi}|\mathbf{y})\pi(\mathbf{A})\pi(\boldsymbol{\theta}).$$

However, as mentioned in Chapter 2, computation of the likelihood using the forward-backward recursion can lead to underflow/overflow problems as the sample size $T$ increases. Hence, instead of working directly from the posterior distribution, all calculations are based on the log-posterior distribution

$$\log\left(\pi(\boldsymbol{\phi}|\mathbf{y})\right) = \ell(\boldsymbol{\phi}|\mathbf{y}) + \log \pi(\mathbf{A}) + \log \pi(\boldsymbol{\theta}) + C(\mathbf{y}), \tag{4.2}$$

for some constant $C(\mathbf{y})$ related to the normalizing constant of $\pi(\boldsymbol{\phi}|\mathbf{y})$ (actually we have that the normalizing constant is $e^{C(\mathbf{y})}$), and where $\ell(\boldsymbol{\phi}|\mathbf{y})$ is the log-likelihood to be computed using the scaling method of Section 2.3.3.

Now suppose that the number of hidden states $k$ is unknown, but it is known that $k \in \{1, \ldots, K\}$. In other words, we have $K$ models to choose from and our objective is no longer only to estimate the parameters; we are also interested in finding the best model that fits the given data. As was discussed in Section 3.4, the Bayesian approach to this problem treats the number of components $k$ as a random parameter to be estimated along with the other parameters. It should be clear that

the parameters $\mathbf{A}$ and $\boldsymbol{\theta}$ both depend on $k$, so that the full posterior distribution takes the following hierarchical form

$$\pi(\boldsymbol{\phi}|\mathbf{y}) \propto \mathcal{L}_k(\boldsymbol{\phi}^k|\mathbf{y})\pi(\mathbf{A}|k)\pi(\boldsymbol{\theta}|k)P(k).$$

(*cf.* Section 3.4 and, in particular equation (3.12)). As before, all calculations have to be based on the log-posterior

$$\log\left(\pi(\boldsymbol{\phi}|\mathbf{y})\right) = \ell_k(\boldsymbol{\phi}^k|\mathbf{y}) + \log\pi(\mathbf{A}|k) + \log\pi(\boldsymbol{\theta}|k) + \log P(k) + C(\mathbf{y}), \qquad (4.3)$$

in order to avoid difficulties linked with underflow/overflow problems, and where $C(\mathbf{y})$ again denotes some constant related to the normalizing constant of $\pi(\boldsymbol{\phi}|\mathbf{y})$. We next outline how contour probabilities can be computed in the current setup, and show that the exact value of $C(\mathbf{y})$ plays no role in the sampling process.

### 4.2.2 Defining the Contour Probabilities

In the current setup, the contour probabilities are calculated from the posterior distribution. We outline how this is done from the log-posterior. First, assume that we have generated $n$ independent random uniform points denoted $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ according to steps 1 and 2 of the Fu & Wang algorithm, and that $\log\pi_i$ has been calculated for each of the $n$ points, where

$$\pi_i = \mathcal{L}_{k_i}(\boldsymbol{\phi}_i^{k_i}|\mathbf{y})\pi(\mathbf{A_i}|k_i)\pi(\boldsymbol{\theta}_i|k_i)P(k_i).$$

Note that equation (4.3) and this definition allow us to write

$$\log\pi(\boldsymbol{\phi}_i|\mathbf{y}) = \log\pi_i + C(\mathbf{y}),$$

or

$$\pi(\boldsymbol{\phi}_i|\mathbf{y}) = \pi_i e^{C(\mathbf{y})}.$$

Now, define $M = \max_i(\log \pi_i)$ and, for $i = 1, \ldots, n$,

$$\pi_i^* = \exp(\log \pi_i - M).$$

It turns out that the contour probabilities can be obtained as

$$P_N(i) = \frac{\bar{\pi}_i^*}{\sum_{j=1}^N \bar{\pi}_j^*} \qquad i = 1, 2, \ldots, N, \tag{4.4}$$

where

$$\bar{\pi}_i^* = \frac{1}{l} \sum_{\boldsymbol{\phi}_j \in E_i} \pi_j^*,$$

and $l = n/N$ corresponds to the number of points falling into each contour. Note that the contour probabilities obtained through (4.4) are exactly the same as the contour probabilities that would be obtained working directly from the posterior distribution provided the normalizing constant were known. Indeed, we have that

$$\begin{aligned}
\bar{\pi}_i^* &= \frac{1}{l} \sum_{\boldsymbol{\phi}_j \in E_i} \pi_j^* \\
&= \frac{1}{l} \sum_{\boldsymbol{\phi}_j \in E_i} \exp(\log \pi_j - M) \\
&= \frac{\exp\{-(M + C(\mathbf{y}))\}}{l} \sum_{\boldsymbol{\phi}_j \in E_i} \exp\{\log \pi_j + C(\mathbf{y})\} \\
&= \frac{\exp\{-(M + C(\mathbf{y}))\}}{l} \sum_{\boldsymbol{\phi}_j \in E_i} \pi(\boldsymbol{\phi}_j | \mathbf{y}),
\end{aligned}$$

so that the extra constants naturally cancel out when computing the contour probabilities using equation (4.4). The important advantage of working from the log-posterior and calculating the contour probabilities from the previous scheme using the $\pi_i^*$, is the computational stability of the resulting procedure. We now move on to apply this methodology to a series of examples (most of which have appeared in

the literature) to see how it can perform in practice. Some of the practical issues that arise with the use of this technique will also be discussed.

# Chapter 5

# Applications

In this chapter, we apply the method outlined in Chapter 4 to the Bayesian analysis of various data sets for which stationary and time-homogeneous HMM's are considered adequate. We also compare our results with those obtained by other authors using maximum likelihood estimation and other Bayesian approaches based on reversible jump Markov chain Monte Carlo. The numerical analyses and simulations presented in this chapter were done on *R* (version 2.2.1) on a Windows platform.

## 5.1 A Two-state Poisson HMM

We first look at the so-called *epileptic seizure count data*. The data records the number of myoclonic seizures a patient suffered on each day for 225 consecutive days. Le *et al.* (1992) fit a two-state hidden Markov model to these data. As noted in MacDonald & Zucchini (1997, Chapter 4, p. 147), this is not a correct version of the data, and observations 92-112 inclusive should be discarded. However, for the purpose of comparison with the result of Le *et al.* (1992), we will use the full set of 225 observations. These data are replicated in Figure 5.1.

For these data, we assume a Poisson distribution for the seizure counts $Y_t$ given the true state of the process $X_t$, linked with the patient's epileptic activity level

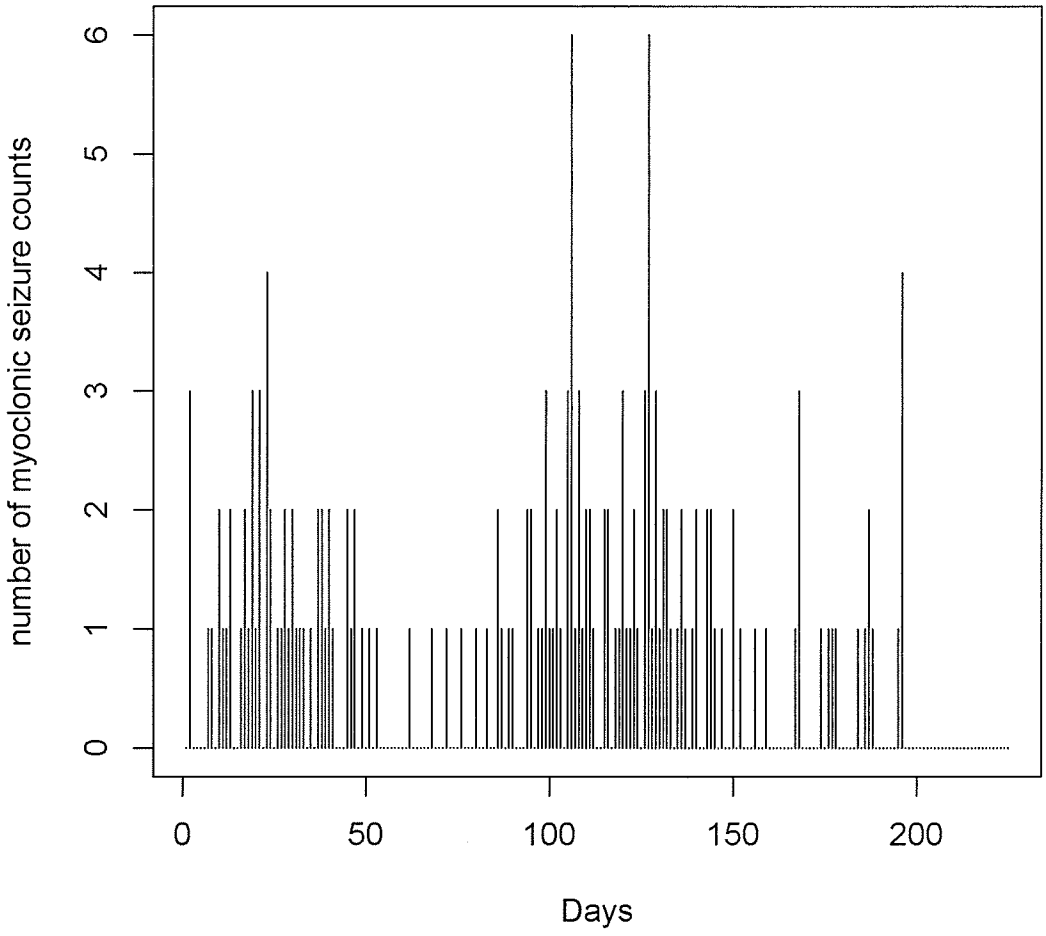


Figure 5.1: Myoclonic seizure series

according to

$$X_t = \begin{cases} 1 & \text{if the patient is in a period of low seizure activity at time } t, \\ 2 & \text{if the patient is in a period of high seizure activity at time } t. \end{cases}$$

It shoud be clear that the true activity level $X_t$ of the patient's epilepsy is not observable and that the conditional independence implied by the use of a HMM is reasonable here. In other words, we use a HMM where

$$Y_t | X_t = i \sim Poisson(\theta_i),$$

for $i = 1, 2$ and $\{X_t\}$ is a two-state Markov chain.

For the Poisson parameters, we use independent Gamma priors

$$\theta_i \sim gamma(a, b)$$

for $i = 1$ and 2, but with the extra restriction that $\theta_1 < \theta_2$ so that the model is identifiable. The parameters for the prior distributions are specified to be $a = 1$, and $b = 0.0001$ so as to make the prior "nearly noninformative". Such a prior (with a very large variance) is sometimes referred to as a *diffuse* or *vague* prior (*cf.* Robert, 2001). Each row $\mathbf{A}_i = (a_{i1}, a_{i2})$ of the transition probability matrix $\mathbf{A}$ is assigned an independent Dirichlet prior,

$$\mathbf{A}_i \sim D(\alpha_1, \alpha_2)$$

with $(\alpha_1, \alpha_2) = (1, 1)$, implying $\mathbf{A}_i$ is uniformly distributed over the two-dimensional simplex. This is usually considered as a noninformative prior specification and is a standard procedure in the Bayesian literature.

With the above specifications, we have essentially four unknown paramters, namely $a_{11}, a_{21}, \theta_1$, and $\theta_2$, the other two being fixed as $a_{i2} = 1 - a_{i1}$ $(i = 1, 2)$. The posterior simulation was run according to the steps outlined in Section 4.1 and using the rescaled forward-backward recursion. The reader might want to refer to Section 4.2.2 for the notation.

1. The initial compact intervals were chosen to be

$$a_{11},\ a_{21} \in (0, 1),$$

$$\theta_1,\ \theta_2 \in (0, 6),$$

   with the requirement that $\theta_1 < \theta_2$.

2. We generated $10^6$ uniform points from the initial compact interval, and evaluated $\log \pi_i$ for $i = 1, \ldots, 10^6$. These points were then partitioned into $10^5$ contours, so that each contour contained 10 points.

3. The contour probabilities $P_N(j)$ for $j = 1, \ldots, 10^5$ were then calculated using (4.4).

4. A random sample of $m = 2000$ was drawn by first sampling contours (with replacement) from the probability mass function obtained in step 3, and then by drawing points from within each contour with equal probability.

5. Marginal histograms were plotted using the sample obtained in step (4).

After the first iteration, the initial compact intervals were reduced to

$$a_{11} \in (0.7, 1), \quad a_{21} \in (0, 0.26),$$

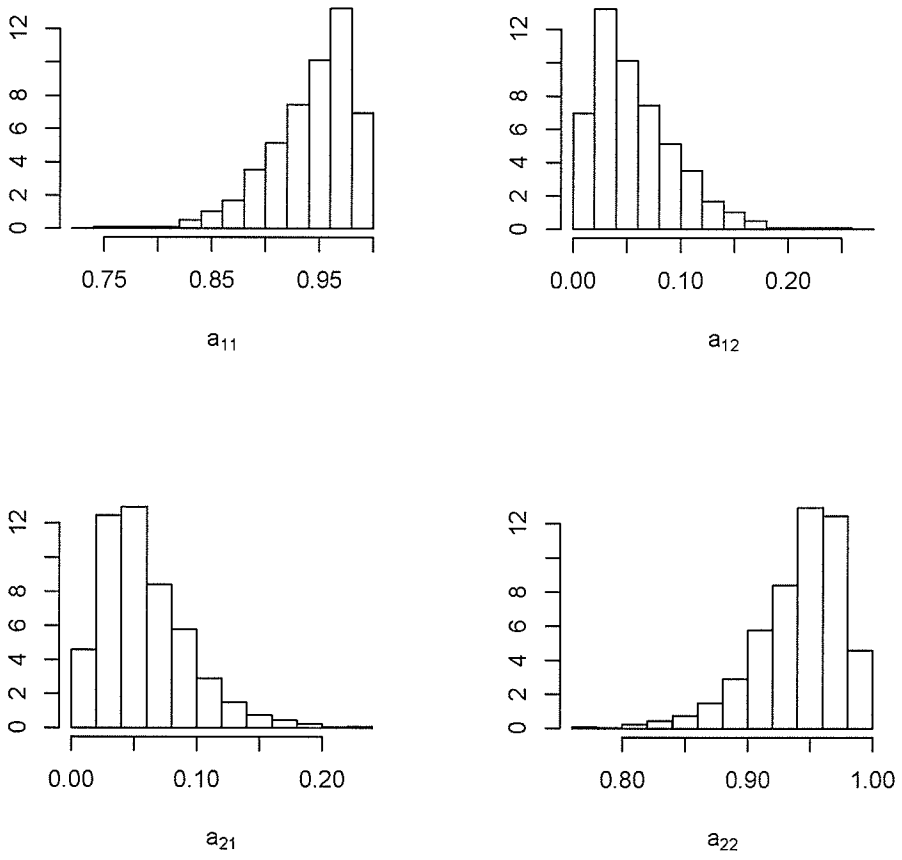


Figure 5.2: Approximated posterior distributions of $a_{ij}$ $(i, j = 1, 2)$, epileptic seizure count data.

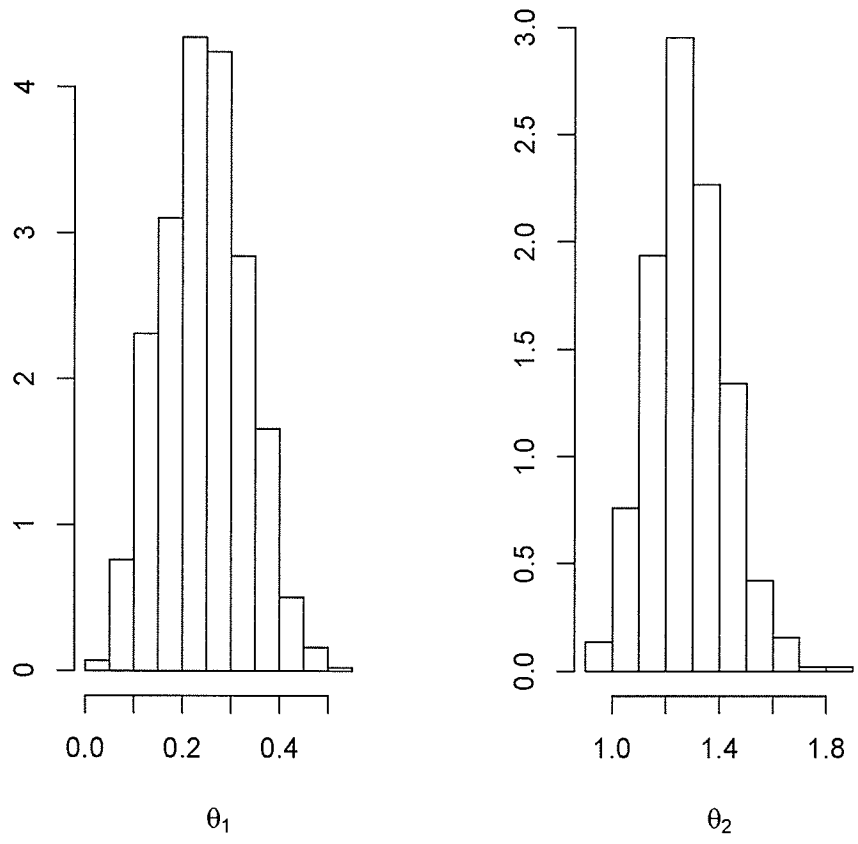


Figure 5.3: Approximated posterior distributions of $\theta_i$ $(i = 1, 2)$, epileptic seizure count data.

These results are compared to the results given in Le *et al.* (1992). The MLE given in their paper is

$$\hat{\theta}_{1,MLE} = 0.287, \quad \hat{\theta}_{2,MLE} = 1.255,$$

and

$$\hat{\mathbf{A}}_{MLE} = \begin{pmatrix} 0.986 & 0.014 \\ 0.024 & 0.976 \end{pmatrix}.$$

These results are very similar to our approximate posterior mode.

## 5.2 A Two-state Normal HMM

The *S & P 500 stock index* data consists of 1700 observations of daily returns during the 1950's. These data were previously analyzed by Rydén *et al.* (1998) using maximum likelihood estimation and in a Bayesian framework by Robert *et al.* (2000) using mixtures of zero-mean normal distributions. The data were preprocessed such that each observation falling outside the range $\bar{y} \pm 4s$ was replaced by the limit of the interval, where $\bar{y}$ is the sample mean and $s$ is the sample standard deviation. For more information, readers are referred to Rydén *et al.* (1998). For computational issues (since the data set contains 1700 observations), we will analyze this set of data with a fixed number of components $k = 2$ (note that, Rydén *et al.* (1998) and Robert *et al.* (2000) both concluded $k = 2$). A histogram of this data set is shown in Figure 5.4.

To analyze these data, we use the following Bayesian HMM model. First, like Rydén *et al.* (1998) and Robert *et al.* (2000), we assume the daily returns $Y_t$ follow a normal distribution with mean zero and unknown variance linked to the unobservable

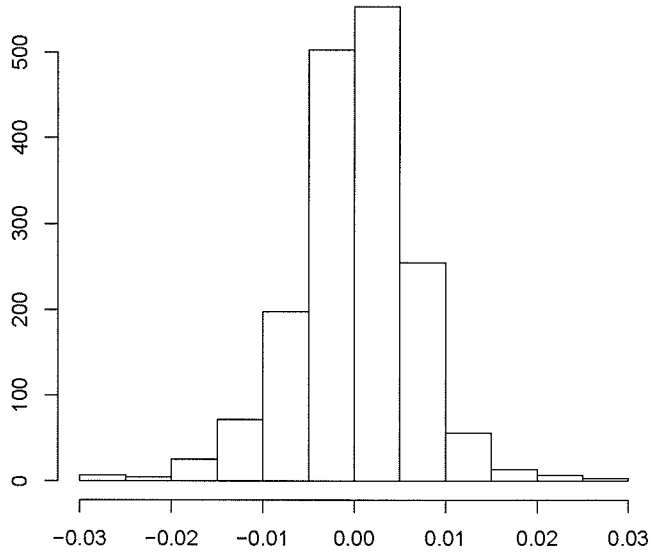


Figure 5.4: Histogram of S&P 500 Stock index data

true state $X_t$ of the U.S. economy at time $t$ according to

$$Y_t|X_t = i \sim N(0, \sigma_i^2),$$

where $X_t$ is assumed to be a two-state Markov chain.

For the prior distributions, we use an inverse-gamma distribution for $\sigma_i^2$, that is

$$\sigma_i^2 \sim IG(2, (R_y/6)^2)$$

with again, for identifiability, $\sigma_1^2$ and $\sigma_2^2$ being sorted in an ascending order, and where $R_y$ is the range of the data (see Wang & Fu (2007) for this prior specification). Again, as before, the priors on each row $\mathbf{A}_i = (a_{i1}, a_{i2})$ of the transition probability matrix

are taken to be independent Dirichlet

$$\mathbf{A}_i \sim D(1,1).$$

Again, these prior distributions, that are uniform over the two-dimensional simplex, were selected because they are often considered to be noninformative.

When sampling from the posterior distribution, the initial compact intervals were chosen to be

$$a_{11}, a_{21} \in (0,1),$$

$$\sigma_1^2,\, \sigma_2^2 \in (0, 9.0 \times 10^{-4}),$$

with $\sigma_1^2 < \sigma_2^2$. Following the steps outlined in Section 5.1, we generated $10^6$ uniform points to discretize the support. These points were then partitioned into $10^5$ contours, so that each contour contained 10 points. Then, a random sample of size 2000 was obtained, and marginal histograms were plotted to visualize the significant regions. After three iterations, the significant regions were found to be

$$a_{11} \in (0.9, 1), \quad a_{21} \in (0, 0.2),$$

$$\sigma_1^2 \in (1.6 \times 10^{-5}, 2.8 \times 10^{-5}), \quad \sigma_2^2 \in (6.0 \times 10^{-5}, 1.2 \times 10^{-4}).$$

From these intervals, a final sample of size 2000 was obtained, leading to the approximated posterior distributions displayed in Figure 5.5 and 5.6.

The posterior mean and standard deviation of $\sigma_1^2$, $\sigma_2^2$ are found to be

$$\mathbb{E}\left((\sigma_1^2, \sigma_2^2)|\mathbf{y}\right) \simeq (2.1724 \times 10^{-5}, 8.6983 \times 10^{-5}),$$

$$S(\sigma_1^2|\mathbf{y}) = 1.607 \times 10^{-6},$$

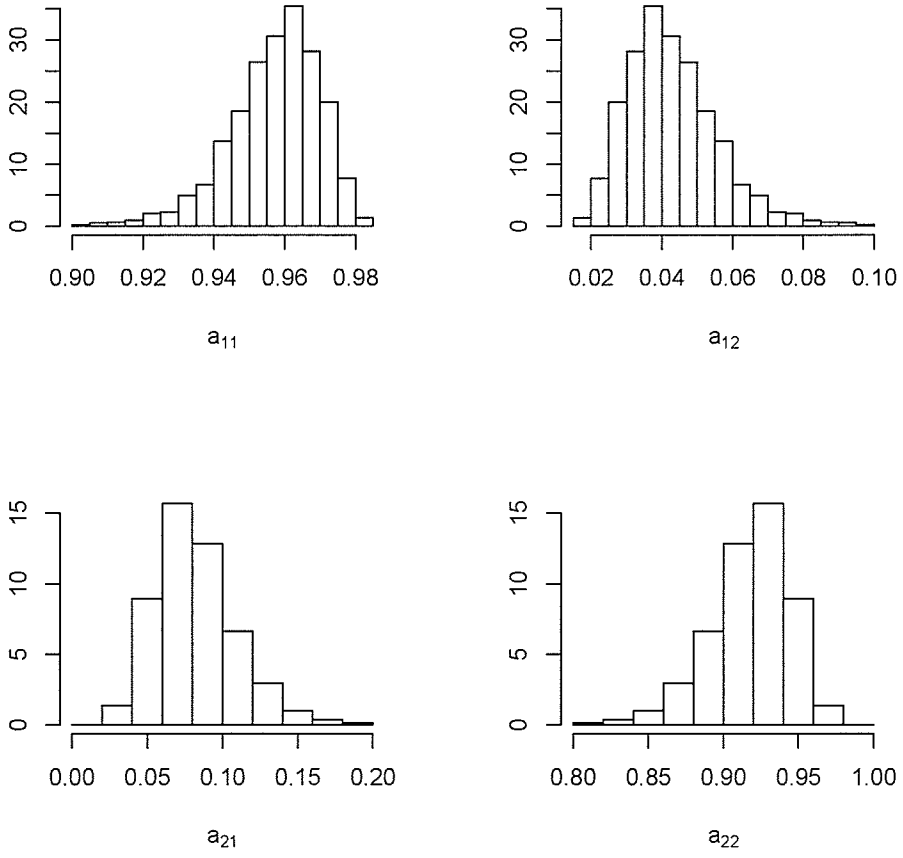


Figure 5.5: Approximated posterior distributions of $a_{ij}$ $(i, j = 1, 2)$, S & P 500 data.

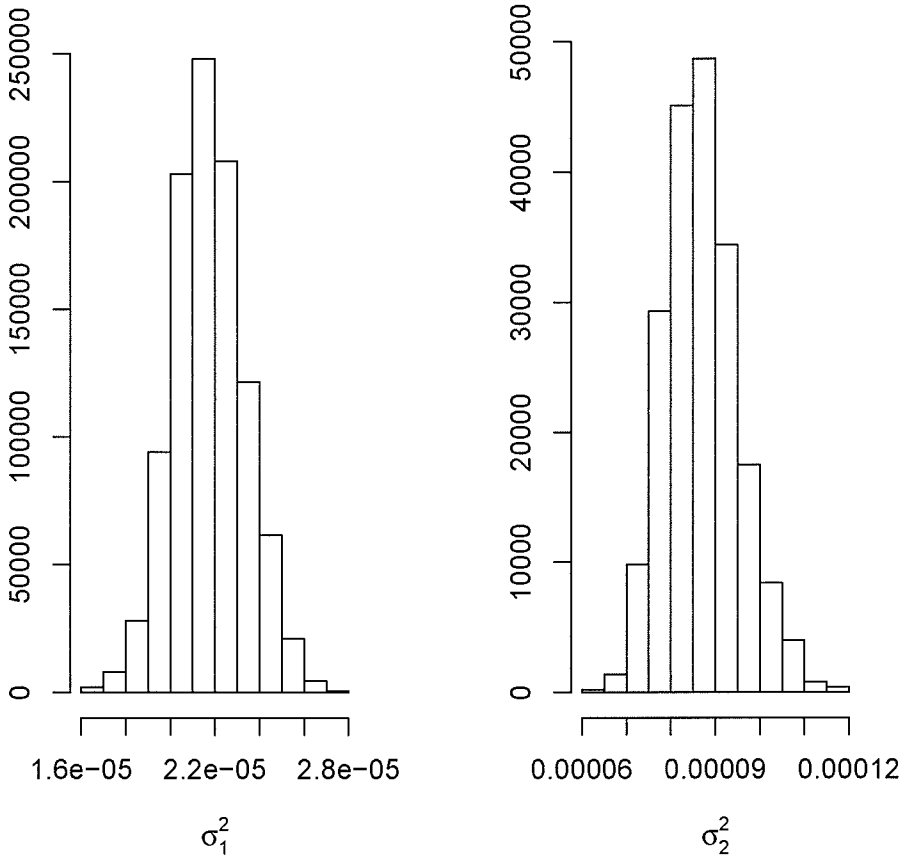


Figure 5.6: Approximated posterior distributions of $\sigma_i^2$ $(i = 1, 2)$, S & P 500 data.

and

$$S(\sigma_2^2|\mathbf{y}) = 8.1680 \times 10^{-6},$$

whereas, the posterior means and standard deviations of the elements of $\mathbf{A}$ are found to be, respectively,

$$\mathbb{E}(\mathbf{A}|\mathbf{y}) \simeq \begin{pmatrix} 0.9571 & 0.0429 \\ 0.0820 & 0.9180 \end{pmatrix},$$
$$S(a_{11}|\mathbf{y}) = S(a_{12}|\mathbf{y}) \simeq 0.0127,$$

and

$$S(a_{21}|\mathbf{y}) = S(a_{22}|\mathbf{y}) \simeq 0.0268.$$

We found the AMLE and APM also gave the same values for this data. These values are

$$\hat{\sigma}^2_{1,AMLE} = \hat{\sigma}^2_{1,APM} = 2.1882 \times 10^{-5},$$

$$\hat{\sigma}^2_{2,AMLE} = \hat{\sigma}^2_{2,APM} = 8.5683 \times 10^{-5},$$

and

$$\hat{\mathbf{A}}_{AMLE} = \hat{\mathbf{A}}_{APM} = \begin{pmatrix} 0.9630 & 0.0370 \\ 0.0686 & 0.9314 \end{pmatrix}.$$

These values are very close to those given by Rydén *et al.* (1998) and by Robert *et al.* (2000). For the purpose of comparison, we report their results here. The MLEs for $\sigma_1^2$, $\sigma_2^2$ and $\mathbf{A}$ reported by in Rydén *et al.* (1998) are

$$\hat{\sigma}^2_{1,MLE} = 2.116 \times 10^{-5}, \quad \hat{\sigma}^2_{2,MLE} = 8.464 \times 10^{-5},$$

and

$$\hat{\mathbf{A}}_{MLE} = \begin{pmatrix} 0.963 & 0.037 \\ 0.069 & 0.931 \end{pmatrix},$$

while the posterior mean of $\sigma_1^2$, $\sigma_2^2$ and $\mathbf{A}$ found in Robert *et al.* (2000) are

$$\bar{\sigma}_1^2 = 2.116 \times 10^{-5}, \quad \bar{\sigma}_2^2 = 8.464 \times 10^{-5},$$

and

$$\bar{\mathbf{A}} = \begin{pmatrix} 0.956 & 0.044 \\ 0.083 & 0.917 \end{pmatrix}.$$

respectively.

## 5.3 A First Example where $k$ is Unknown

We generated a random sample of size $N = 750$ observations from a normal HMM with $k = 3$ hidden states, under the specifications $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$, with common variance $\sigma^2 = 0.25$ and the following transition probability matrix

$$\mathbf{A} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.25 & 0.5 & 0.25 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

A histogram of the simulated data set is shown in Figure 5.7. We here investigate the performance of our sampling approach to Bayesian inference for HMM's when the number of components $k$ is unknown. In particular, we wish to see if the Bayesian methodology will correctly identify the number of components $k$.

In accordance with the methodology presented in Section 3.4, we analyzed this data set by considering $k$ as an unknown parameter to be estimated along with all other parameters. In this example, given the number of hidden states $k$ and $X_t = i$ (with $1 \leq i \leq k$), we assumed $Y_t$ satisfies

$$Y_t | X_t = i, k, \boldsymbol{\mu}_k, \sigma_k \sim N(\mu_{ki}, \sigma_k^2),$$

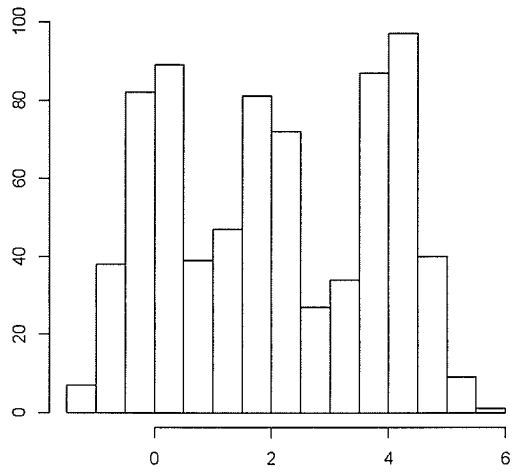


Figure 5.7: Histogram of simulated data set

where $X_t$ has no physical sense but is used as an index generating $Y_t$. Note that this implies the true state of the process $X_t$ only affects the mean of the normal distribution.

The prior distribution for $k$ is chosen to be uniform over the set $\{1, 2, \ldots, K\}$, where the maximum number of hidden state is set to $K = 4$. Given $k$, the rows of the transition probability matrix $\mathbf{A}_k$ are again assumed independent, with a prior distribution that is taken to be $\mathbf{A}_{k,i} \sim D(\alpha_{k,i1}, \alpha_{k,i2}, \ldots, \alpha_{k,ik})$. Thus, for general values of the hyperparameters $\alpha_{k,ij}$, the joint prior for the elements of $\mathbf{A}_k$ is

$$\pi(\mathbf{A}_k) = \prod_{i=1}^{k} \left( \frac{\Gamma(\sum_{j=1}^{k} \alpha_{k,ij})}{\prod_{j=1}^{k} \alpha_{k,ij}} \prod_{j=1}^{k} a_{k,ij}^{\alpha_{k,ij}-1} \right).$$

For objectivity reasons, we again set $\alpha_{k,ij} = 1$ $(i, j = 1, \ldots, k)$ so that each row of $\mathbf{A}_k$ is uniformly distributed over the $k$-dimensional simplex.

Given $k$, we choose ordered normal priors for the vector means $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kk})$, with the restriction $\mu_{k1} < \mu_{k2} < \cdots < \mu_{kk}$ for identifiability purposes. The joint prior distribution of the location parameters is thus

$$\pi(\boldsymbol{\mu}_k|k) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\mu_{kj} - \mu)^2}{2\sigma^2}\right].$$

For the hyperparameters $\mu$ and $\sigma^2$, we follow Richardson and Green (1997), and Wang & Fu (2007) by using $\mu = M_y$ and $\sigma^2 = R_y^2$, where $M_y$ and $R_y$ are the mid-range and range of the data respectively. For our simulated data set, $M_y = 2.1303$ and $R_y = 6.8462$.

Finally, the inverse gamma is used as a prior distribution for $\sigma_k^2$, with density

$$\pi(\sigma_k^2|k) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_k^2)^{-\alpha-1}e^{-\beta/\sigma_k^2}.$$

Here, we set $\alpha = 1$ and $\beta = 2$ as this corresponds to a vague prior (none of its moments exist).

With the above specifications, the full parameter vector becomes

$$\boldsymbol{\phi} = (k, \boldsymbol{\phi}^1, \ldots, \boldsymbol{\phi}^4),$$

where we have that $\boldsymbol{\phi}^1 = \boldsymbol{\theta}_1$, and

$$\boldsymbol{\phi}^k = (\boldsymbol{\theta}_k, \mathbf{A}_k),$$

for $k \geq 2$, with

$$\boldsymbol{\theta}_k = (\mu_{k1}, \ldots, \mu_{kk}, \sigma_k^2),$$

and

$$\mathbf{A}_k = (a_{k,11}, \ldots, a_{k,kk}).$$

Thus here, the parameter vector $\boldsymbol{\phi}$ has

$$1 + \sum_{k=1}^{4}(k+1) + \sum_{k=2}^{4} k^2 = 44$$

components since for all values of $k$, $\boldsymbol{\theta}_k$ has $k+1$ components, and $\mathbf{A}_k$ has $k^2$ components for $k \geq 2$.

For this data set, we sampled from the posterior distribution of $\boldsymbol{\phi}$ using the following steps:

1. The initial compact cover is shown in Table 5.1.

2. We discretized the sample space by generating $7 \times 10^6$ uniform base points from the initial compact intervals (we explain below what we mean by base points) and partitioned these values into $n = 3.5 \times 10^6$ contours, so that each contour contains only 2 points.

3. The contour probabilities were then calculated using the method outlined in section 4.2.2.

4. A random sample of $m = 2000$ was drawn by first sampling contours (with replacement) from the probability mass function obtained in step 3, and then by drawing points from within each contour with equal probability.

5. Finally, we visualized the significant region by constructing histograms using the sample from step 4 above.

This process was repeated 4 times and the significant regions located. These regions are shown in Table 5.2. Note that $k = 1$ and $k = 2$ were excluded from the significant region. From these intervals we repeated the above steps 2-5 to draw a

| $k$ | $\mathbf{A}_k$ | $\boldsymbol{\mu}_k$ | $\sigma_k^2$ |
|---|---|---|---|
| 1 | | $\mu_{11} \in (0,5)$ | $\sigma_1^2 \in (0,2)$ |
| 2 | $a_{ij} \in (0,1)$ $i,j = 1,2$ | $\mu_{21} \in (-1,3)$ $\mu_{22} \in (1,5)$ | $\sigma_2^2 \in (0,2)$ |
| 3 | $a_{ij} \in (0,1)$ $i,j = 1,2,3$ | $\mu_{31} \in (-2,2)$ $\mu_{32} \in (0,4)$ $\mu_{33} \in (2,6)$ | $\sigma_3^2 \in (0,2)$ |
| 4 | $a_{ij} \in (0,1)$ $i,j = 1,2,3,4$ | $\mu_{41} \in (-2,2)$ $\mu_{42} \in (-1,3)$ $\mu_{43} \in (1,5)$ $\mu_{44} \in (2,6)$ | $\sigma_4^2 \in (0,2)$ |

Table 5.1: Initial compact intervals, simulated data set

sample of size $10^4$. In this example, we generated approximately $77 \times 10^6$ uniform points in order to keep $7 \times 10^6$ base points. The following rule was used to determine our base points. Let $P_{\max}$ be the maximum log posterior value of the previous iteration and $P_i$ be all the log posterior values of the current iteration. Finally, let

$$d_i = P_i - P_{\max}.$$

Then, if $e^{d_i} \approx 0$, the point is dropped, otherwise the point is kept and is included in the base points. Thus, base points are "good points" in the sense that points that are kept do not have a zero probabibility of later being sampled. Unfortunately, this is very often the case in such high-dimensional setups as the likelihood function can become mostly flat with spikes that are difficult to detect. This explains why, for example, $77 \times 10^6$ uniform points were generated to retain only $7 \times 10^6$ "reasonably" useful points.

The prior and approximate posterior distributions for $k$ are given in Table 5.3. The highest posterior probability (and by far) is $\mathbb{P}(k = 3|\mathbf{y}) \simeq 0.838$ suggesting

| $k$ | $\mathbf{A}_k$ | | $\boldsymbol{\mu}_k$ | $\sigma_k^2$ |
|---|---|---|---|---|
| 3 | $a_{11} \in (0.6, 1)$ | $a_{12} \in (0, 0.4)$ | $\mu_1 \in (-0.16, 0.1)$ | $\sigma_3^2 \in (0.21, 0.34)$ |
| | $a_{22} \in (0.35, 0.85)$ | $a_{23} \in (0.15, 0.65)$ | $\mu_2 \in (1.74, 2.06)$ | |
| | $a_{32} \in (0.18, 0.5)$ | $a_{33} \in (0.5, 0.82)$ | $\mu_3 \in (3.94, 4.18)$ | |
| 4 | $a_{11} \in (0.6, 1)$ | $a_{12} \in (0, 0.4)$ | $\mu_1 \in (-0.11, 0.1)$ | $\sigma_4^2 \in (0.22, 0.33)$ |
| | $a_{13} \in (0, 0.4)$ | | $\mu_2 \in (1.5, 2.2)$ | |
| | $a_{ij} \in (0, 1)$ | $i = 2, 3, 4$ | $\mu_3 \in (1.6, 4.2)$ | |
| | | $j = 1, 2, 3, 4$ | $\mu_4 \in (3.8, 4.24)$ | |

Table 5.2: Significant regions after 4 iterations, simulated data set.

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Prior | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| Posterior | 0.000 | 0.000 | 0.838 | 0.162 |

Table 5.3: Prior and approximate posterior distribution of $k$, simulated data set.

$k = 3$ is the best candidate for $k$. The posterior means, standard deviations and approximate MLEs for $k = 3$ are given in Table 5.4 along with the true values.

It is interesting to note that, in the current setup, our proposed methodology recovers the number of components $k = 3$ and leads to conditional estimates that are quite close to the true parameter values. The approximate posterior distributions of $\mathbf{A}_3$, $\boldsymbol{\mu}_3$ and $\sigma_3^2$ are shown in Figures 5.8, 5.9 and 5.10.

| | $\mathbf{A}_3$ | | | $\boldsymbol{\mu}_3$ | | | $\sigma_3^2$ |
|---|---|---|---|---|---|---|---|
| True | 0.8000 | 0.1000 | 0.1000 | 0.0000 | 2.0000 | 4.0000 | 0.2500 |
| | 0.2500 | 0.5000 | 0.2500 | | | | |
| | 0.1000 | 0.3000 | 0.6000 | | | | |
| Mean | 0.7364 | 0.1505 | 0.1131 | −0.0261 | 1.8978 | 4.0598 | 0.2677 |
| | 0.2207 | 0.4882 | 0.2911 | | | | |
| | 0.0569 | 0.2992 | 0.6439 | | | | |
| Std. Dev. | 0.0294 | 0.0243 | 0.0209 | 0.0356 | 0.0402 | 0.0330 | 0.0160 |
| | 0.0291 | 0.0349 | 0.0313 | | | | |
| | 0.0149 | 0.0291 | 0.0305 | | | | |
| AMLE | 0.7357 | 0.1454 | 0.1189 | −0.0255 | 1.9015 | 4.0653 | 0.2680 |
| | 0.2232 | 0.4675 | 0.3092 | | | | |
| | 0.0635 | 0.3362 | 0.6003 | | | | |

Table 5.4: True values, posterior means, standard deviations and approximate MLE of $\mathbf{A}_3$, $\boldsymbol{\theta}_3$ and $\sigma_3^2$, simulated data set.

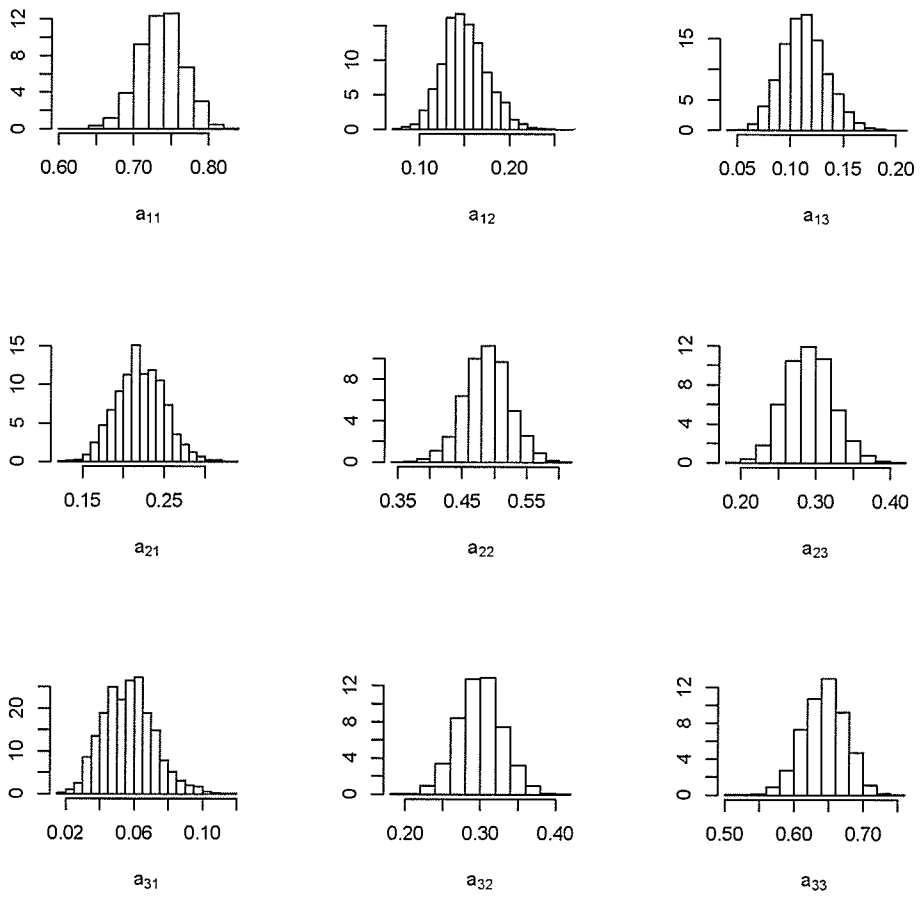


Figure 5.8: Approximate posterior distributions of $a_{3,ij}$, $(i, j = 1, 2, 3)$, simulated data set.

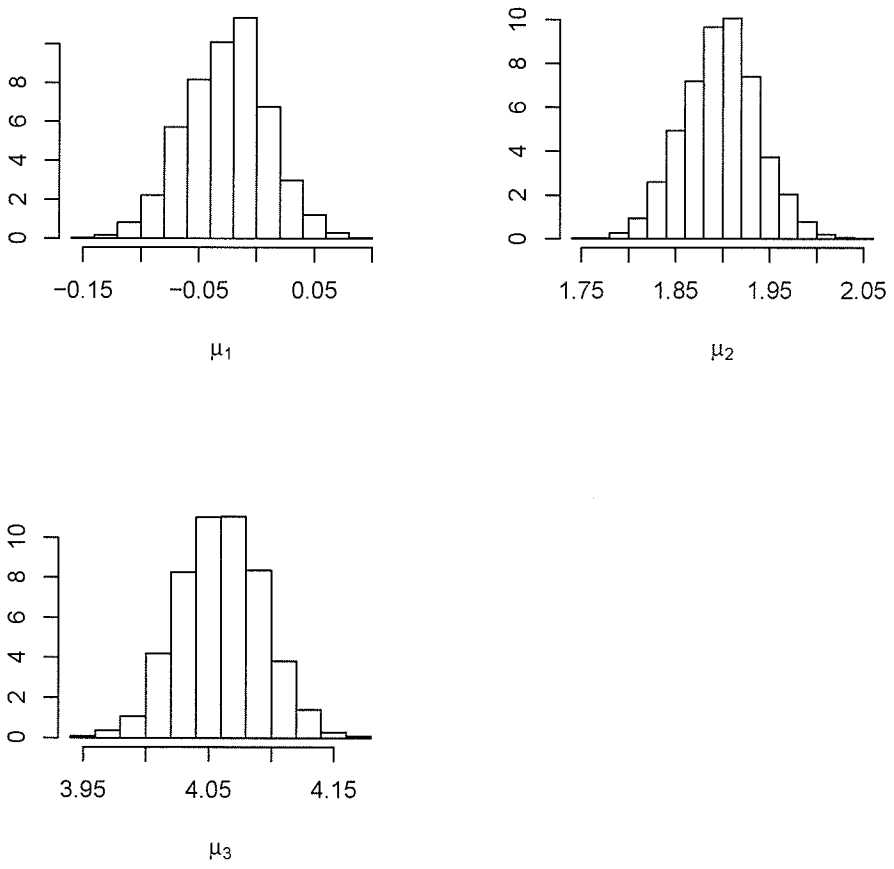


Figure 5.9: Approximate posterior distributions of $\mu_{3i}$, $(i = 1, 2, 3)$, simulated data set.

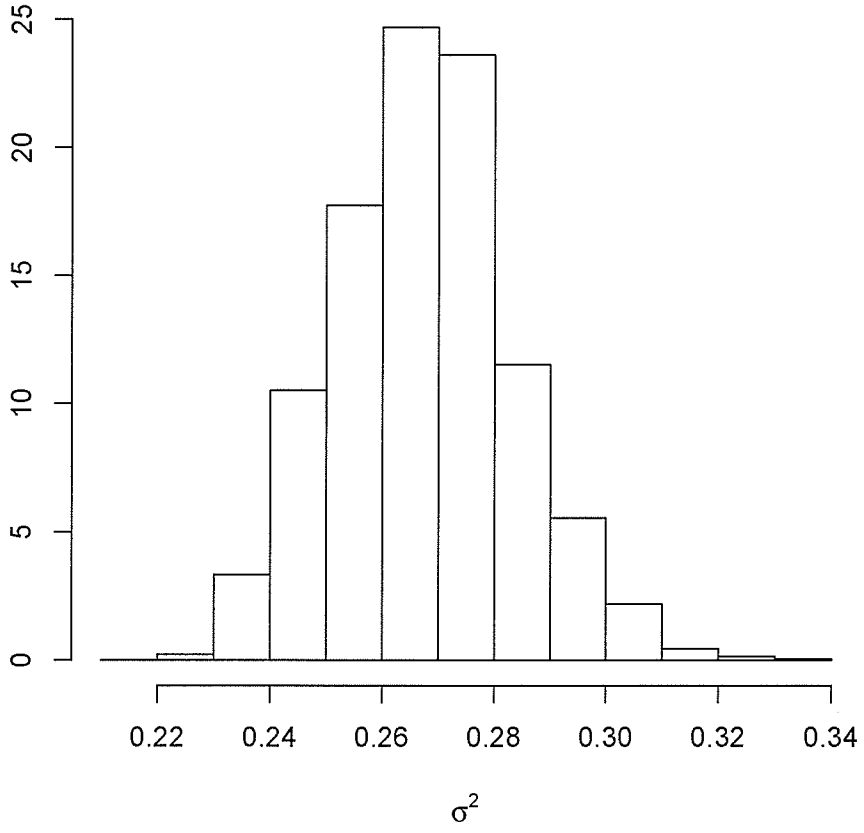


Figure 5.10: Approximate posterior distribution of $\sigma_3^2$, simulated data set.

## 5.4 A Poisson HMM where $k$ is Unknown

We now consider the *fetal lamb movement data* which were studied by Leroux & Puterman (1992) and Scott (2002). These data consist of the number of movements produced by a fetal lamb over 240 consecutive 5 second intervals. These data are shown in Figure 5.11. In their paper, Leroux and Puterman (1992) concluded that the best model included $k = 2$ hidden states using the BIC method, whereas they favored $k = 3$ with the AIC method. Here, we treat the number of hidden components as random rather than a fixed constant, and apply our sampling method to do Bayesian inference on $k$ as well as other unknown parameters.

For this data set, we assume that the movement count $Y_t$ during the $t^{th}$ interval satisfies

$$Y_t|X_t = i, k, \boldsymbol{\theta}_k \sim Poisson(\theta_{ki}),$$

where $X_t \in \{1, \ldots, k\}$ stands for the level of fetal activity during the $t^{th}$ interval (see Scott, 2002 for details).

The prior distribution for $k$ is taken as

$$P(k = i) = 1/K,$$

for $i = 1, 2, \ldots, K$, and where $K$ is again set to 4. As before, the prior for each row $\mathbf{A}_{k,i}$ of the transition probability matrix $\mathbf{A}_k$ is taken to be a Dirichlet prior with $\alpha_{k,ij} = 1$ $(i, j = 1, \ldots, k)$. For the unknown Poisson means, we use independent Gamma priors, but sorted in increasing order so that the model is identifiable, that is

$$\pi(\boldsymbol{\theta}_k|k) = k! \prod_{j=1}^{k} \theta_{kj}^{\alpha-1} \frac{\beta^\alpha e^{-\beta\theta_{kj}}}{\Gamma(\alpha)},$$

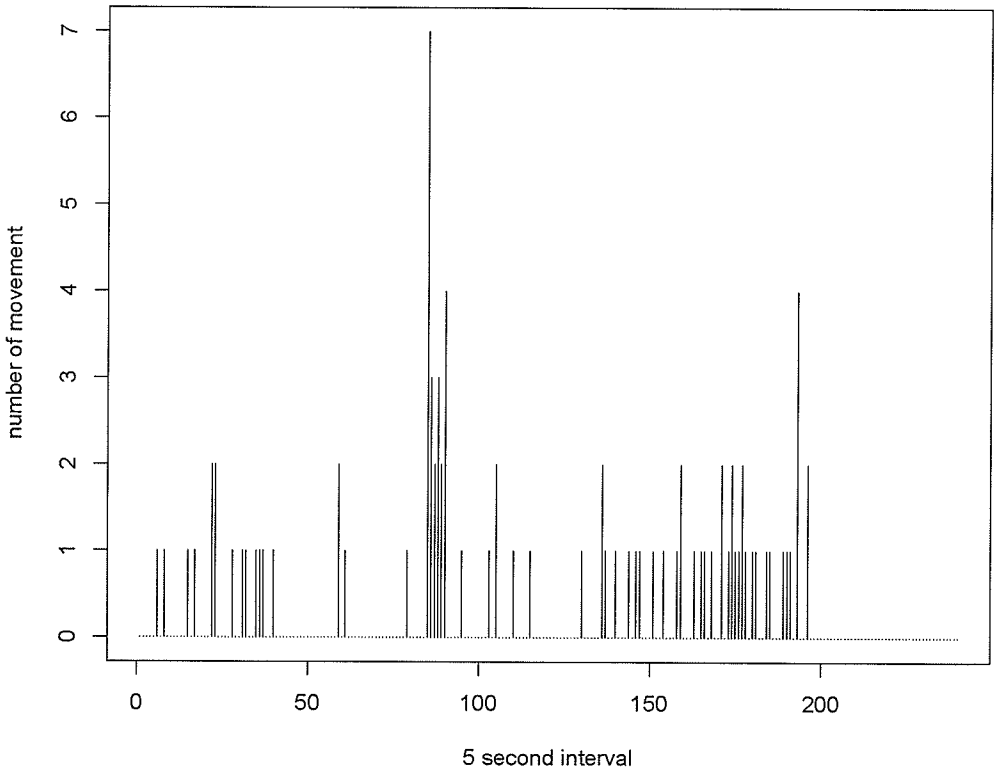


Figure 5.11: The fetal lamb movement series

for $\theta_{k1} < \theta_{k2} < \cdots < \theta_{kk}$, and where the hyperparameters were set to $\alpha = 1$ and $\beta = 0.1$, leading to a vague prior (*cf.* Section 5.1). Under the above specifications, we see that the full parameter vector is

$$\boldsymbol{\phi} = (k, \boldsymbol{\phi}^1, \ldots, \boldsymbol{\phi}^4),$$

where $\boldsymbol{\phi}^1 = \boldsymbol{\theta}_1$ and for $k \geq 2$,

$$\boldsymbol{\phi}^k = (\boldsymbol{\theta}_k, \mathbf{A}_k),$$

where

$$\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kk}),$$

and

$$\mathbf{A}_k = (a_{k,11}, \ldots, a_{k,kk}),$$

Note that $\boldsymbol{\phi}$ here has

$$1 + \sum_{k=1}^{4} k + \sum_{k=2}^{4} k^2 = 40$$

components.

For our analysis, the initial compact intervals are displayed in Table 5.5. We followed the exact same steps as those outlined in Section 5.3. After 5 iterations, the significant regions were reduced to the intervals given in Table 5.6. Note that $k = 1$ was excluded from the significant region. In this example, the total number of uniform random points that we generated to keep $7 \times 10^6$ base points in the final run is approximately $15 \times 10^7$. Here, we used the same rule as in Section 5.3 to determine the base points. The prior and resulting approximated posterior distributions of $k$ are shown in Table 5.7, with $k = 3$ having the highest posterior probability.

| $k$ | $\mathbf{A}_k$ | | $\boldsymbol{\theta}_k$ | |
|---|---|---|---|---|
| 1 | | | $\theta_1 \in (0,2)$ | |
| 2 | $a_{ij} \in (0,1)$ | $i,j = 1,2$ | $\theta_1 \in (0,1)$ | $\theta_2 \in (0,7)$ |
| 3 | $a_{ij} \in (0,1)$ | $i,j = 1,2,3$ | $\theta_1 \in (0,1)$<br>$\theta_3 \in (0,7)$ | $\theta_2 \in (0,4)$ |
| 4 | $a_{ij} \in (0,1)$ | $i,j = 1,2,3,4$ | $\theta_1 \in (0,1)$<br>$\theta_3 \in (0,4)$ | $\theta_2 \in (0,3)$<br>$\theta_4 \in (0,7)$ |

Table 5.5: Initial compact interval, fetal lamb movement data.

| $k$ | $\mathbf{A}_k$ | | $\boldsymbol{\theta}_k$ | |
|---|---|---|---|---|
| 2 | $a_{11} \in (0.86,1)$ | $a_{21} \in (0.01,0.8)$ | $\theta_1 \in (0.07,0.38)$ | $\theta_2 \in (0.87,5.6)$ |
| 3 | $a_{11} \in (0.6,1)$ | $a_{12} \in (0,0.4)$ | $\theta_1 \in (0,0.31)$ | $\theta_2 \in (0.19,2.63)$ |
| | $a_{ij} \in (0,1)$ | $i = 2,3\ j = 1,2,3$ | $\theta_3 \in (1.3,5.5)$ | |
| 4 | | | $\theta_1 \in (0,0.32)$ | $\theta_2 \in (0,2.14)$ |
| | $a_{ij} \in (0,1)$ | $i,j = 1,2,3,4$ | $\theta_3 \in (0.17,3.22)$ | $\theta_4 \in (0.98,6.1)$ |

Table 5.6: Significant region after 5 iterations, fetal lamb movement data

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Prior | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| Posterior | 0.0000 | 0.2744 | 0.4730 | 0.2526 |

Table 5.7: Prior and posterior distribution of $k$, fetal lamb movement.

| | $\mathbf{A}_2$ | | $\boldsymbol{\theta}_2$ | |
|---|---|---|---|---|
| Mean | 0.9759 | 0.0241 | 0.2376 | 2.7143 |
| | 0.3505 | 0.6495 | | |
| Std. Dev. | 0.0165 | 0.0165 | 0.0433 | 0.8451 |
| | 0.1410 | 0.1410 | | |

Table 5.8: Posterior means and standard deviations of $\mathbf{A}_2$ and $\boldsymbol{\theta}_2$, fetal lamb movement data.

We here report the posterior means and standard deviations along with all approximate posterior distributions for both $k = 2$ and $k = 3$. The posterior means and standard deviations are given in Table 5.8 ($k = 2$) and Table 5.9 ($k = 3$), whereas the approximate posterior distributions of the elements of $\boldsymbol{\theta}_2$ and $\mathbf{A}_2$ are displayed in Figures 5.12, 5.13 and those of $\boldsymbol{\theta}_3$ and $\mathbf{A}_3$, in Figures 5.14, and 5.15, respectively. We also report the approximate MLE of $\mathbf{A}_k$ and $\boldsymbol{\theta}_k$ for $k = 2$, 3 in Table 5.10.

For comparison, the results of Leroux & Puterman (1992) are given in Table 5.11. Note that the posterior means of $\mathbf{A}_2$ and $\boldsymbol{\theta}_2$ are similar to Leroux & Putterman's MLE, whereas the posterior means of $\mathbf{A}_3$ and $\boldsymbol{\theta}_3$ are quite different from their results. Note, however, that the AMLE of $\mathbf{A}_k$ and $\boldsymbol{\theta}_k$ we obtained for $k = 2$ and $k = 3$ are very similar to their MLE, except for the estimated value of $\theta_{33}$.

## 5.5 A Normal HMM where $k$ is Unknown

The last example we consider consists of the *wind velocity data*, a series of the first 500 hourly wind velocity differences measured at Athens in January 1990. This data set was first considered by Francq & Roussignol (1995), and later by Robert

| | $\mathbf{A}_3$ | | | $\boldsymbol{\theta}_3$ | | |
|---|---|---|---|---|---|---|
| Mean | 0.9167 | 0.0611 | 0.0222 | 0.0861 | 0.5893 | 3.0426 |
| | 0.1177 | 0.8346 | 0.0477 | | | |
| | 0.2417 | 0.2001 | 0.5583 | | | |
| Std. Dev. | 0.0579 | 0.0564 | 0.0182 | 0.0650 | 0.3370 | 0.7814 |
| | 0.1536 | 0.2091 | 0.1014 | | | |
| | 0.1628 | 0.1640 | 0.1843 | | | |

Table 5.9: Posterior means and standard deviations of $\mathbf{A}_3$ and $\boldsymbol{\theta}_3$, fetal lamb movement data.

| $k$ | 2 | | 3 | | |
|---|---|---|---|---|---|
| A | 0.9886 | 0.0114 | 0.9503 | 0.0291 | 0.0206 |
| | | | 0.0415 | 0.9505 | 0.0080 |
| | 0.3112 | 0.6888 | 0.1928 | 0.0272 | 0.7800 |
| $\boldsymbol{\theta}$ | 0.2582 | 3.1280 | 0.0328 | 0.4639 | 4.3720 |

Table 5.10: Approximate MLE of $\mathbf{A}_k$ and $\boldsymbol{\theta}_k$ $(k = 2, 3)$, fetal lamb movement data.

| $k$ | 2 | | 3 | | |
|---|---|---|---|---|---|
| A | 0.9884 | 0.0116 | 0.9468 | 0.0433 | 0.0099 |
| | | | 0.0424 | 0.9576 | 0 |
| | 0.3083 | 0.6917 | 0.1838 | 0 | 0.8162 |
| $\boldsymbol{\theta}$ | 0.2560 | 3.1006 | 0.0447 | 0.5090 | 3.4138 |

Table 5.11: MLE of $\mathbf{A}_k$ and $\boldsymbol{\theta}_k$ $(k = 2, 3)$ obtained by Leroux & Puterman (1992), fetal lamb movement data.

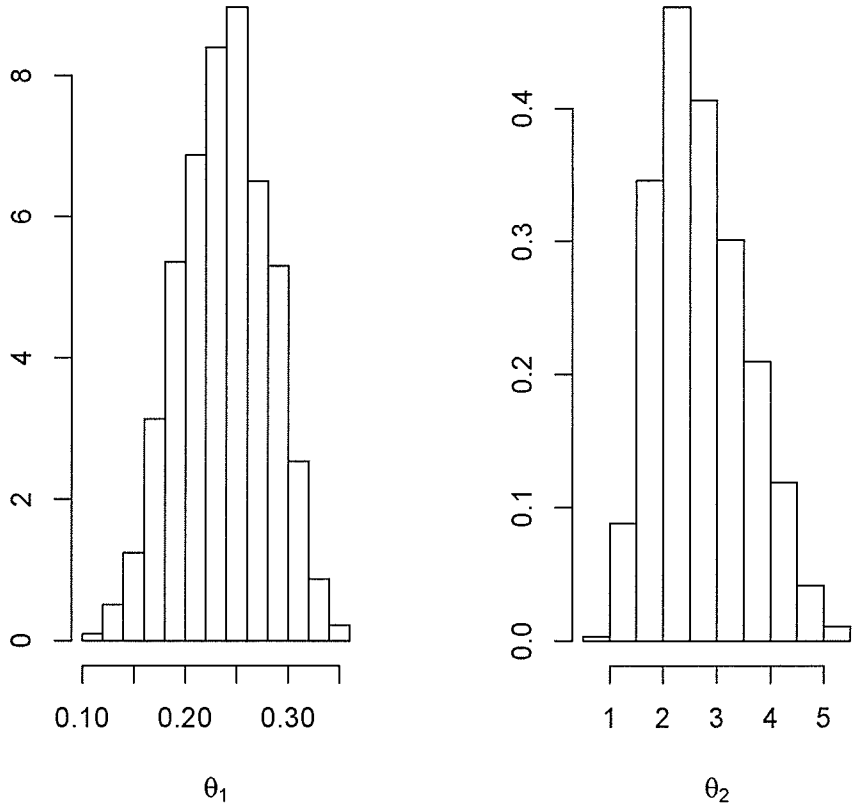


Figure 5.12: Approximate posterior distributions of $\theta_{2i}$ $(i = 1, 2)$, fetal lamb movement data.

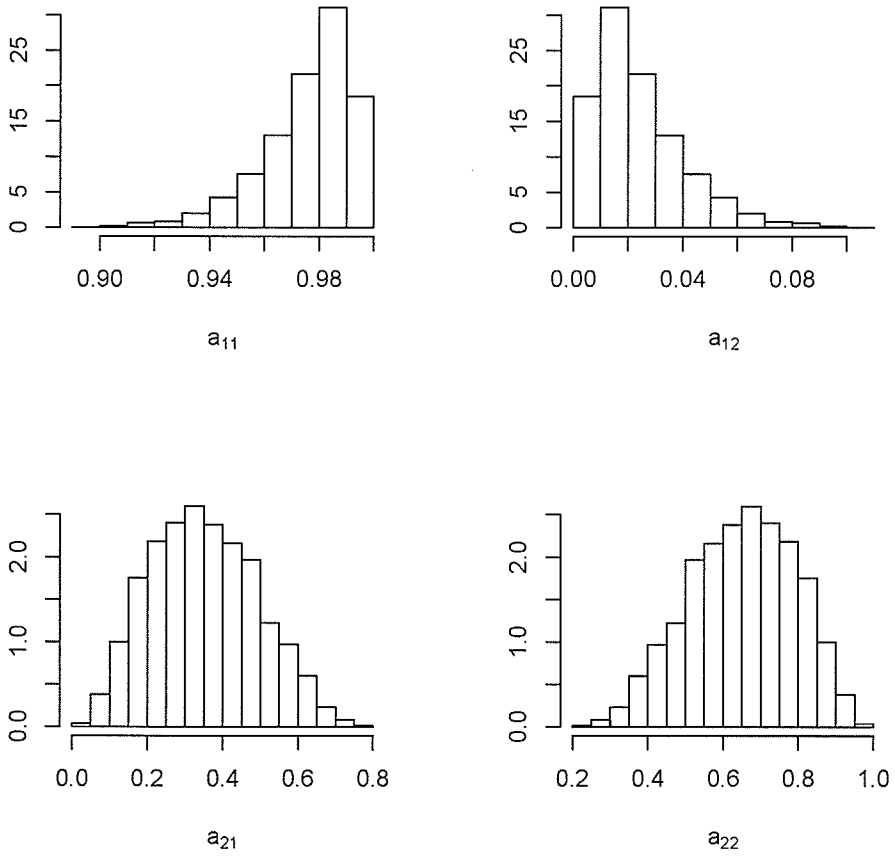


Figure 5.13: Approximate posterior distributions of $a_{2,ij}$ $(i, j = 1, 2)$, fetal lamb movement data.

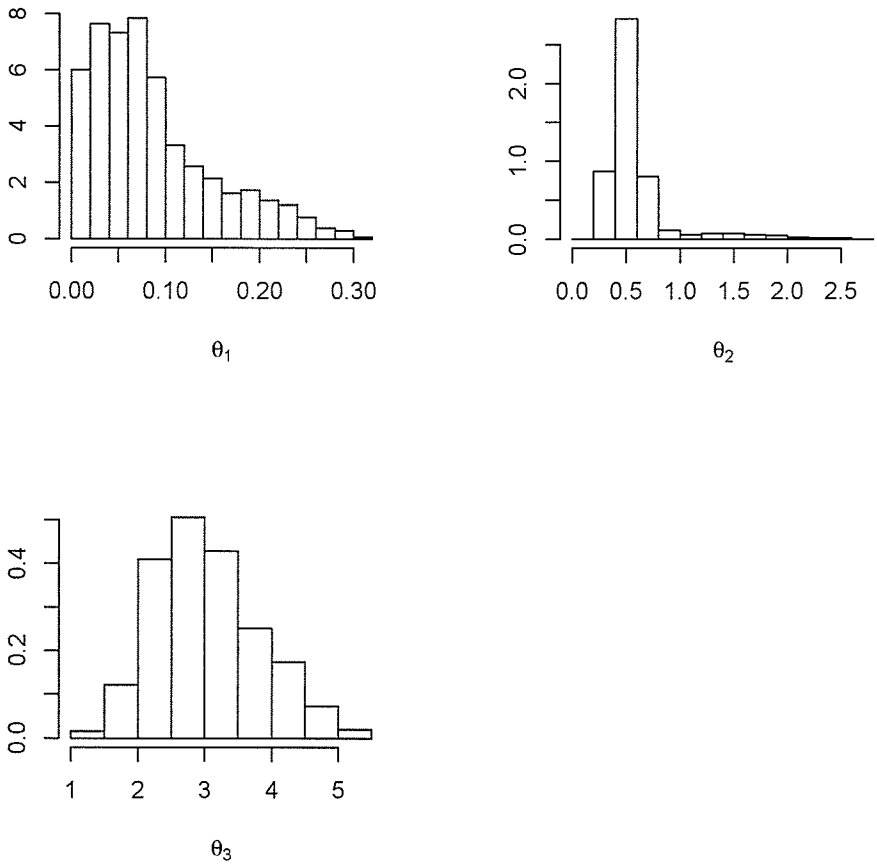


Figure 5.14: Approximate posterior distributions of $\theta_{3i}$ $(i = 1, 2, 3)$, fetal lamb movement data.

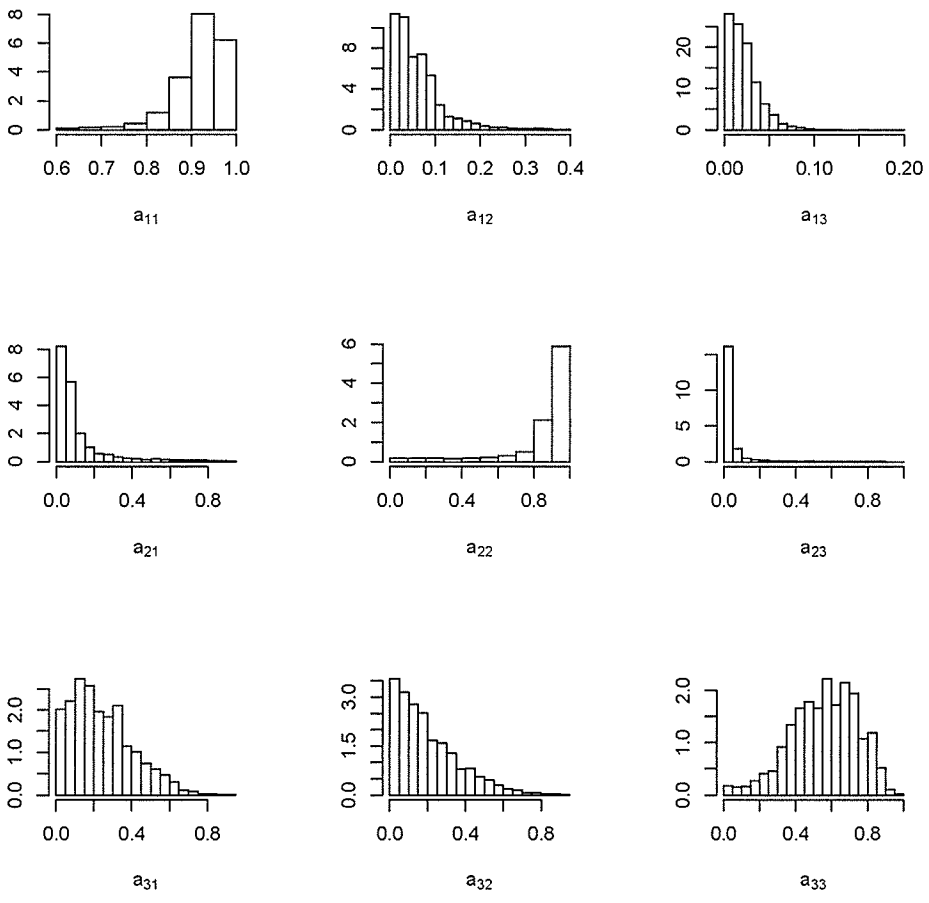


Figure 5.15: Approximate posterior distributions of $a_{3,ij}$ $(i, j = 1, 2, 3)$, fetal lamb movement data.

*et al.* (2000). The model we consider in what follows is identical to the one used by Robert *et al.* (2000), although they relied on reversible jump MCMC for all computations. Note this model differs considerably from the one originally used by Francq & Roussignol (1995).

The HMM that was used here is based on zero-mean normal distributions with unknown variances. Specifically, let $V_t$ denote the wind velocity measured at time $t$, and

$$Y_t = V_t - V_{t-1} \qquad \text{for } t = 1, \ldots, 500,$$

denote the wind velocity differences. Then, the observable differences $Y_t$ were assumed to satisfy

$$Y_t | X_t = i, k, \boldsymbol{\sigma}_k^2 \sim N(0, \sigma_{ki}^2).$$

where $X_t \in \{1, \ldots, k\}$ stands for the planetary geomagnetic activity index. In their paper, Robert *et al.* considered $k$ to vary between 1 and $K = 7$. They concluded that the number of components is $k = 3$ with the approximate posterior probability for $k \geq 5$ being shown to be less than 1%. For the purpose of comparison with their results, we analyzed these data by also treating $k$ as an unknown parameter and estimating it alongside with other unknown parameters. However, for computational issues, we considered the maximum number of hidden states to be $K = 4$. The histogram of the data set is reproduced in Figure 5.16.

As with previous examples, the prior distribution for $k$ was chosen to be uniform, that is $P(k = i) = 1/4$ for $i = 1, \ldots, 4$, whereas each row of the transition probability matrix $\mathbf{A}_k$ was assigned a uniform prior on the $k$-dimensional simplex. For the variance parameters, the prior distribution is taken to be the joint distribution of a

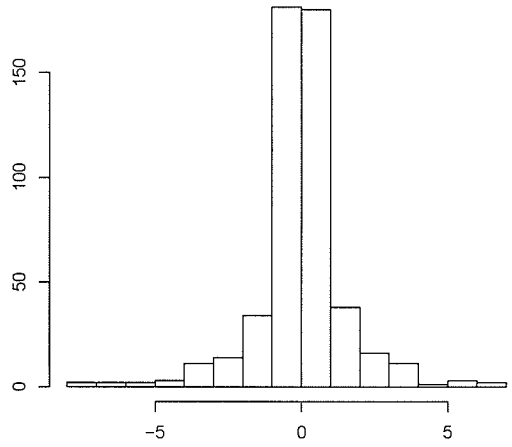


Figure 5.16: Histogram of hourly wind velocity differences.

set of $k$ ordered inverse gamma random variables, that is

$$\pi(\boldsymbol{\sigma}_k^2|k) = k! \prod_{j=1}^{k} \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_{kj}^2)^{-\alpha-1} e^{-\beta/\sigma_{kj}^2}, \quad \sigma_{k1}^2 < \sigma_{k2}^2 < \cdots < \sigma_{kk}^2,$$

for $k = 1, \ldots, 4$. The hyperparameters were set to $\alpha = 2$ and $\beta = 1$. The complete parameter vector for this problem is

$$\boldsymbol{\phi} = (k, \phi^1, \ldots, \phi^4)$$

with $\boldsymbol{\phi}^1 = \boldsymbol{\sigma}_1^2$ and, for $k \geq 2$,

$$\boldsymbol{\phi}^k = (\boldsymbol{\sigma}_k^2, \mathbf{A}_k),$$

where

$$\boldsymbol{\sigma}_k^2 = (\sigma_{k1}^2, \ldots, \sigma_{kk}^2),$$

and

$$\mathbf{A}_k = (a_{k,11}, \ldots, a_{k,kk}),$$

and so, has

$$1 + \sum_{k=1}^{4} k + \sum_{k=2}^{4} k^2 = 40$$

elements.

In order to simulate from the posterior distribution $\pi(\boldsymbol{\phi}|\mathbf{y})$ using our methodology, the initial compact regions we used are shown in Table 5.12. Again, we followed the steps outlined in Section 5.3, generating $7 \times 10^6$ uniform base points from the initial intervals, which we partitioned into $3.5 \times 10^6$ contours. We then sampled $m = 10000$ points from these contours and studied the associated marginal histograms. After 5 iterations, the significant regions were obtained for all the coordinates. These are given in Table 5.13. In this case, we generated approximately $71 \times 10^6$ uniform random points to keep $7 \times 10^6$ base points. The prior and resulting approximated posterior distributions of $k$ are given in Table 5.14. These show that $k = 3$ has by far the highest posterior probability, which agrees with the results obtained by Robert *et al.* (2000). The posterior means and standard deviations along with the approximate MLE are shown in Table 5.15. The approximate posterior distributions of the elements of $\mathbf{A}_3$ and $\boldsymbol{\sigma}_3^2$ are displayed in Figure 5.17 and 5.18 respectively.

## 5.6 Concluding Remarks

Through the examples given above, we see that our approach to sampling from the posterior distribution, which is based essentially on the sampling method of Fu & Wang (2002) and the use of the forward/backward recursion, seems to work quite well in HMM setups with both a fixed and unknown number of hidden states. The results

| $k$ | $\mathbf{A}_k$ | | $\boldsymbol{\sigma}_k^2$ | |
|---|---|---|---|---|
| 1 | | | $\sigma_1^2 \in (0,4)$ | |
| 2 | $a_{ij} \in (0,1)$ | $i,j = 1,2$ | $\sigma_1^2 \in (0,1)$ | $\sigma_2^2 \in (0,12)$ |
| 3 | $a_{ij} \in (0,1)$ | $i,j = 1,2,3$ | $\sigma_1^2 \in (0,1)$<br>$\sigma_3^2 \in (2,18)$ | $\sigma_2^2 \in (0,6)$ |
| 4 | $a_{ij} \in (0,1)$ | $i,j = 1,2,3,4$ | $\sigma_1^2 \in (0,1)$<br>$\sigma_3^2 \in (0,12)$ | $\sigma_2^2 \in (0,6)$<br>$\sigma_4^2 \in (2,18)$ |

Table 5.12: Initial compact intervals, wind velocity data.

| $k$ | $\mathbf{A}_k$ | | $\boldsymbol{\sigma}_k^2$ | |
|---|---|---|---|---|
| 2 | $a_{11} \in (0.88,1)$ | $a_{21} \in (0.02,0.2)$ | $\sigma_1^2 \in (0.12,0.28)$ | $\sigma_2^2 \in (4,8.5)$ |
| 3 | $a_{11} \in (0.6,1)$<br>$a_{2j} \in (0,1)$<br>$a_{32} \in (0,0.55)$ | $a_{12} \in (0,0.4)$<br>$j = 1,2$<br>$a_{33} \in (0.45,1)$ | $\sigma_1^2 \in (0.09,0.24)$<br>$\sigma_3^2 \in (4,14.7)$ | $\sigma_2^2 \in (0.1,4)$ |
| 4 | $a_{11} \in (0.1,1)$<br>$a_{13} \in (0,0.9)$<br>$a_{ij} \in (0,1)$ | $a_{12} \in (0,0.9)$<br><br>$i = 2,3.\ j = 1,2,3$ | $\sigma_1^2 \in (0.09,0.23)$<br>$\sigma_3^2 \in (0.2,6)$ | $\sigma_2^2 \in (0.1,2.3)$<br>$\sigma_4^2 \in (4.7,14.7)$ |

Table 5.13: Significant region after 5 iterations, wind velocity data.

| $K$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Prior | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| Posterior | 0.000 | 0.0153 | 0.8055 | 0.1792 |

Table 5.14: Prior and approximated posterior distributions of $k$, wind velocity data.

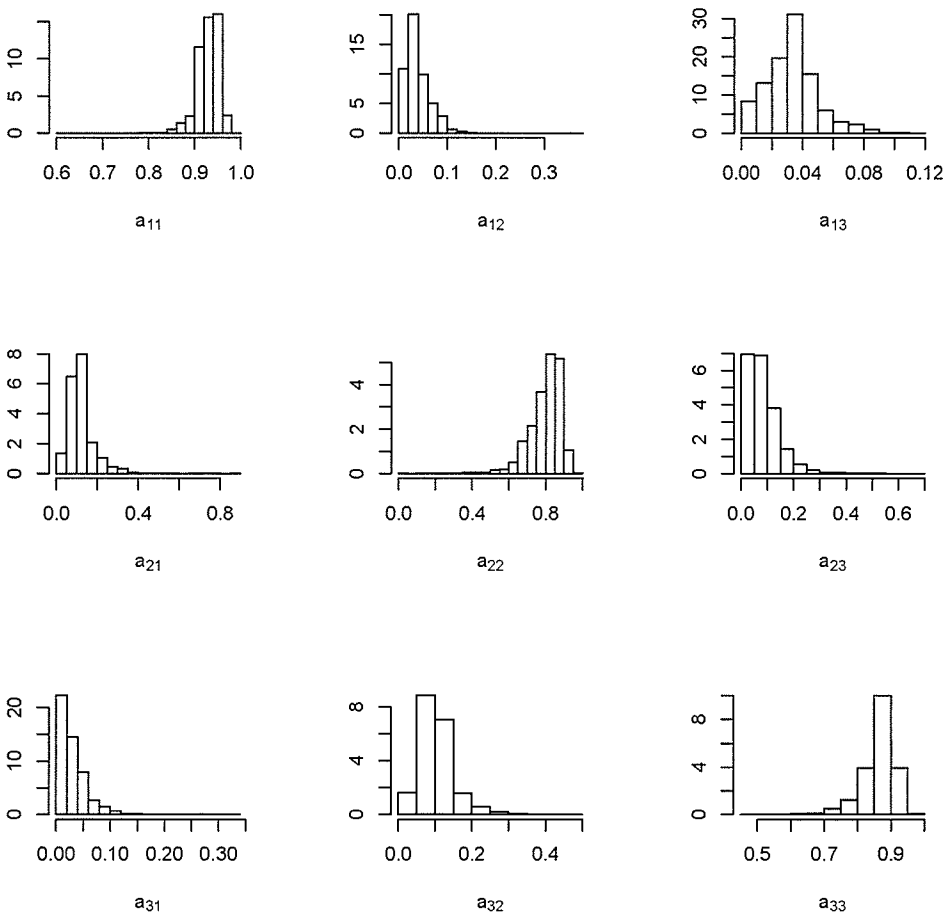


Figure 5.17: Approximate posterior distributions of $a_{3,ij}$, $(i, j = 1, 2, 3)$, wind velocity data.

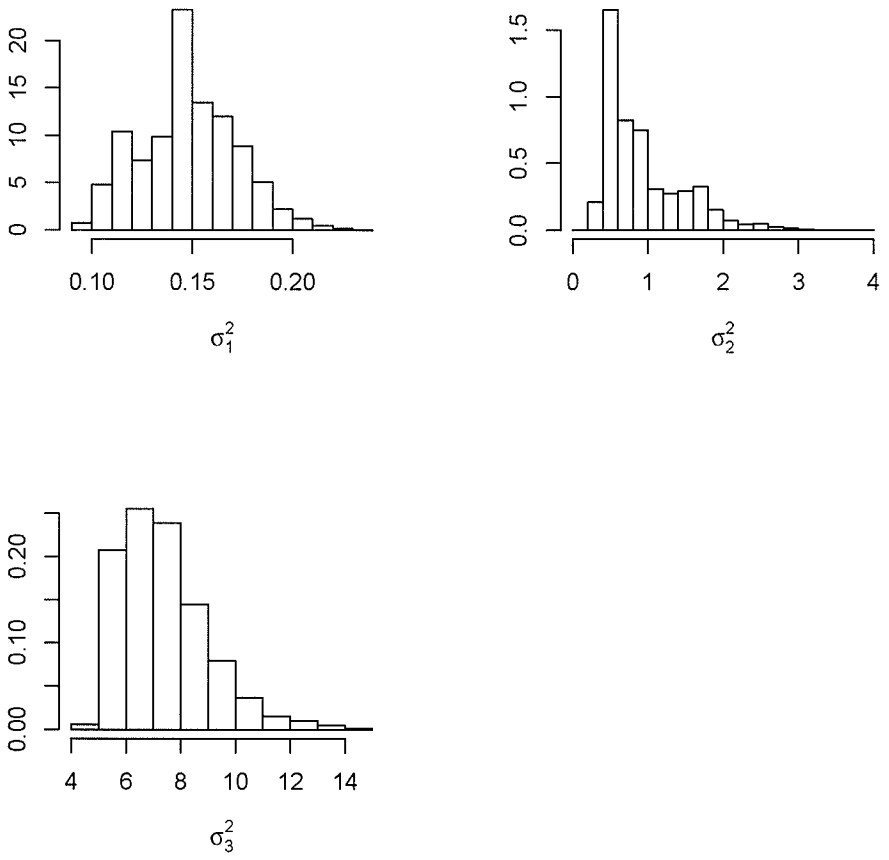


Figure 5.18: Approximate posterior distributions of $\sigma_{3i}^2$, $(i = 1, 2, 3)$, wind velocity data.

| | $\mathbf{A}$ | | | $\boldsymbol{\sigma}$ | | |
|---|---|---|---|---|---|---|
| Mean | 0.9286 | 0.0386 | 0.0328 | 0.1476 | 0.9353 | 7.3905 |
| | 0.1220 | 0.7957 | 0.0822 | | | |
| | 0.0297 | 0.1063 | 0.8640 | | | |
| Std. Dev. | 0.0243 | 0.0266 | 0.0168 | 0.0238 | 0.5332 | 1.5734 |
| | 0.0706 | 0.0942 | 0.0634 | | | |
| | 0.0252 | 0.0477 | 0.0519 | | | |
| AMLE | 0.9563 | 0.0212 | 0.0225 | 0.1602 | 1.7698 | 8.6775 |
| | 0.0692 | 0.8806 | 0.0502 | | | |
| | 0.0422 | 0.0824 | 0.8755 | | | |

Table 5.15: Posterior mean, standard deviation and approximate MLE of $\mathbf{A}_3$, and $\boldsymbol{\sigma}_3^2$, wind velocity data.

obtained in this thesis for fixed $k$ (*cf.* Section 5.1 and 5.2) are comparable to those available in the literature, using MLE and MCMC methodologies. For the simulated data set (*cf.* Section 5.3), our method not only correctly identified the number of hidden states, but also gave estimates that are very close to the true parameter values. For the wind velocity data (*cf.* Section 5.5), the conclusion of $k = 3$ agrees with the results of Robert *et al.* (2000). Finally, for the fetal lamb movement data (*cf.* Section 5.4), locating the significant regions was much more difficult than with the other examples, requiring the sampling of about twice as many random uniform points ($150 \times 10^6$ versus $70 \times 10^6$ for the other examples where $k$ is also unknown). Interestingly, the simulated data example even had more parameters to be estimated. One possible explanation for this is that the joint posterior distribution for the fetal lamb movement data example might be extremely flat with a few isolated highly-

peaked regions. Also, by inspecting the histograms of Figure 5.15, the posterior distributions of some of the elements of $\mathbf{A}$ seem to be highly concentrated close to 1. This could be problematic because generating uniformly distributed points over a $k$-dimensional simplex, where $k > 2$, makes it difficult to visit the "corners" of that simplex (i.e., regions corresponding to one element being close to one). Obviously, then, a lot more points are required if a good approximation to the significant region is to be obtained. However, the method still produced reasonable results, but with a considerable increase in computational effort.

# Chapter 6

# Conclusion

Reversible jump MCMC has been used for Bayesian inference in hidden Markov models with an unknown number of hidden states. However, the algebraic complexity of the method makes it very difficult to implement. This thesis adapted a direct sampling approach based on the algorithm of Fu & Wang (2002). This method allows one to sample over the significant region of the posterior distribution and ignore the insignificant part. It is suitable for Bayesian inference in hidden Markov models with both a known and unknown number of hidden states. It is easy to implement and the knowledge of the density function is required only up to a multiplicative constant. Unlike Gibbs sampling and reversible jump MCMC, this method does not required data-augmentation to handle the hidden states. Furthermore, we can simultaneously obtain the posterior expectation, posterior mode and approximate maximum likelihood estimate at no extra cost in terms of computational effort.

The method was used on both simulated data and real-life data. It turns out that this method not only correctly identified the unknown number of hidden states but also produced estimates of the model parameters that agrees with those previously obtained in the literature.

Therefore, with the availability of powerful computing devices, the simplicity of this method (and its ability to consider cases where the number of hidden states $k$ is unknown) make it a promising tool for inference in the HMM setup.

# Appendix A

# R Code for the Epileptic Seizure Count Series

```
y<-c(  0,3,0,0,0,0,1,1,0,2,1,1,2,0,0,
       1,2,1,3,1,3,0,4,2,0,1,1,2,1,2,
       1,1,1,0,1,0,2,2,1,2,1,0,0,0,2,
       1,2,0,1,0,1,0,1,0,0,0,0,0,0,0,
       0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,
       1,0,0,0,1,0,0,1,0,0,2,1,0,1,1,
       0,0,0,2,2,0,1,1,3,1,1,2,1,0,3,
       6,1,3,1,2,2,1,0,0,2,2,0,1,1,3,
       1,1,2,1,0,3,6,1,3,1,2,2,1,0,1,
       2,1,0,1,2,0,0,2,2,1,0,1,0,0,2,
       0,1,0,0,0,1,0,0,1,0,0,0,0,0,0,
       0,1,3,0,0,0,0,0,1,0,1,1,1,0,0,
       0,0,0,1,0,1,2,1,0,0,0,0,0,0,1,
       4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
       0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
```

```
logD <- function(a) {
  sum(lgamma(a)) - lgamma(sum(a))
}

ddirichlet<-function(x,alpha) {
  s<-sum((alpha-1)*log(x))
  exp(sum(s)-logD(alpha))
}

rdirichlet<-function(n,a) {
  l<-length(a)
  x<-matrix(rgamma(l*n,a),ncol=l,byrow=TRUE)
  sm<-x%*%rep(1,l)
  x/as.vector(sm)
```

```
}

hmm.sim.pars<-function(k,N){

  hmm<-list()
  hmm$tpm<-array(0,c(k,k,N))
  hmm$theta<-matrix(0,N,k)
  hmm$post<-rep(0,N)

  n<-length(y)
  c<-rep(0,n)
  alpha <- matrix( 0, n, k)

  hmm$mle_loglik<--Inf

  for (i in 1:N){
    a<-runif(1,0,1)
    b<-1-a
    c<-runif(1,0,1)
    d<-1-c
    tpm<-matrix(c(a,b,c,d),k,k,T)
    ssd<-abs(eigen(t(tpm))$vec[,1])
    ssd<-ssd/sum(ssd)
    theta<-c(runif(1,0,6),runif(1,0,6))
    theta<-sort(theta)

    fmat<-sapply(1:k,function(k){dpois(y,theta[k])})

    alpha[1,] <- ssd * fmat[1,]
    c[1]<-1/sum(alpha[1,])
    alpha[1,]<-alpha[1,]*c[1]
    for( j in 2:n ){
      alpha[j,] <- (alpha[j-1,] %*% tpm) * fmat[j,]
      c[j]<-1/sum(alpha[j,])
      alpha[j,]<-alpha[j,]*c[j]
    }
    loglik<--sum(log(c))

    d<-sapply(1:k,function(k){ddirichlet(tpm[k,],rep(1,2))})
    g<-dgamma(theta,1,0.0001)
```

```
    post<-loglik+sum(log(d))+sum(log(g))
    hmm$post[i]<-post
    hmm$tpm[,,i]<-tpm
    hmm$theta[i,]<-theta

    if (loglik>hmm$mle_loglik){
      hmm$mle_loglik<-loglik
      hmm$mle_tpm<-tpm
      hmm$mle_theta<-theta
    }
  }

  hmm$tpm<-hmm$tpm[,,order(hmm$post,decreasing=T)]
  hmm$theta<-hmm$theta[order(hmm$post,decreasing=T),]
  save(hmm,file="hmmpb")
  hmm
}

hmm.sample.contours<-function(m,k,N,c){

  sample<-list()
  hmm<-hmm.sim.pars(k,N)

  hmmpost<-exp(sort(hmm$post-max(hmm$post),decreasing=T))

  contours<-function(i,N,c){
    sum(hmmpost[((i-1)*(N/c)+1):(i*(N/c))])
  }

  i<-c

  C<-sapply(1:i,function(i){contours(i,N,c)})
  p<-C/sum(C)

  s<-sample(1:c,m,replace=T,prob=p)
  ss<-rep(0,m)

  l<-0

  for (i in 1:max(s)){
```

```
  if(length(s[s==i])!=0)
  ss[((l+1):(l+length(s[s==i])))]<-sample(((i-1)*(N/c)+1):(i*(N/c))
                                    ,length(s[s==i]),replace=T)
  l<-l+length(s[s==i])
}

sample$tpm<-hmm$tpm[,,ss]
sample$theta<-hmm$theta[ss,]
save(sample,file="sampleb")
sample

windows()
par(mfrow=c(1,2))
for (i in 1:k)
  hist(sample$theta[,i], main="")
  windows()
  par(mfrow=c(k,k))
    for (i in 1:k){
      for (j in 1:k)
        hist(sample$tpm[i,,][j,], main="")
    }

Meantheta<-apply(sample$theta,2,mean)
stdevtheta<-apply(sample$theta,2,sd)
Mean=matrix(sapply(1:k,function(k){apply(sample$tpm[k,,],1,mean)})
                                    ,k,k,T)
stdev=matrix(sapply(1:k,function(k){apply(sample$tpm[k,,],1,sd)})
                                    ,k,k,T)

min_theta<-apply(sample$theta,2,min)
max_theta<-apply(sample$theta,2,max)
min_tpm<-matrix(sapply(1:k,function(k){apply(sample$tpm[k,,],1,min)
                                      }),k,k,T)
max_tpm<-matrix(sapply(1:k,function(k){apply(sample$tpm[k,,],1,max)
                                      }),k,k,T)

return(list(Mean=Mean, std.dev=stdev,Meantheta=Meantheta,
            std.dev_theta=stdevtheta,mode_tpm=hmm$tpm[,,1],
            mode_theta=hmm$theta[1,],mle_tpm=hmm$mle_tpm,
            mode_post=sort(hmm$post,decreasing=T)[1],
```

```
                mle_theta=hmm$mle_theta,mle_loglik=hmm$mle_loglik,
                Min_theta=min_theta,Max_theta=max_theta,
                Min_tpm=min_tpm,Max_tpm=max_tpm))
}

##hmm.sample.contours(2000,2,1000000,100000)
```

# Bibliography


[1] Albert, P.S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, **47**, 1371-1381.

[2] Albert, J.H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, **11**, 1-15.

[3] Baras, J.S. and Finesso, L. (1992). Consistent estimation of the order of hidden Markov chains. In *Stochastic Theory and Adaptive Control (Lawrence, KS, 1991)*, 26-39. Springer-Verlag, Berlin.

[4] Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A Maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164-171.

[5] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second Edition, Springer, New York.

[6] Bickel, P.J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, **26**, 1614-1635.

[7] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer, New York.

[8] Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Second Edition, Chapman & Hall/CRC, Boca Raton.

[9] Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler *The American Statistician*, **46**, 167-174.

[10] Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79-97.

[11] Churchill, G.A. (1989). Stochastic models for heterogeous DNA sequences. *Bulletin of Mathematical Biology*, **51**, 79-94.

[12] Devijver, P.A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, **3**, 369-373.

[13] Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **7**, 381-420.

[14] Dempster, A., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[15] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363-375.

[16] Evans, M., and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press, New York.

[17] Francq, C. and Roussignol, M. (1995). On white noises driven by hidden Markov chains. *Journal of Time Series Analysis*, **18** 553-578

[18] Fredkin, D.R. and Rice, J.A. (1992). Bayesian restoration of single channel patch clamp recordings. *Biometrics*, **48**, 427-448.

[19] Fu, J.C. and Wang, L. (2002). A random-discretization based Monte Carlo sampling method and its applications. *Methodology and Computing in Applied Probability*, **4**, 5-25.

[20] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

[21] Gelman, A., Carlin, J.B., and Stern, H.S. (2003). *Bayesian Data Analysis*. Second Edition, Chapman & Hall/CRC, Boca Raton.

[22] Giudici, P., Rydén, T. and Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics*, **56**, 742-747.

[23] Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

[24] Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman and Hall, New York.

[25] Juang, B.H. and Rabiner, L.R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**, 251-272.

[26] Le, N.D., Leroux, B.G. and Puterman, M.L. (1992). Reader reaction: Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, **48**, 317-323

[27] Leroux, B.G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, **40**, 127-143.

[28] Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545-558.

[29] MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series.* Chapman and Hall, New York.

[30] MacKay, R.J. (2002). Estimating the order of a hidden Markov model. *The Canadian Journal of Statistics*, **30**, 573-589.

[31] Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257-286.

[32] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Rayal Statistical Society, Series B*, **59**, 731-792.

[33] Robert, C.P. (2001). *The Bayesian Choice.* Second Edition, Springer, New York.

[34] Rober, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Second Edition, Springer, New York.

[35] Robert, C.P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics & Probabiliy Letters*, **16**, 77-83.

[36] Robert, C.P., Rydén, T. and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, **62**, 57-75.

[37] Robert, C.P. and Titterington, D.M. (1998). Reparametrization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, **8**, 145-158.

[38] Ross, S.M. (2007). *Introduction to Probability Models.* Ninth Edition, Academic Press, San Diego.

[39] Rydén, T. (1995). Estimating the order of hidden Markov models. *Statistics*, **26**, 345-354.

[40] Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, **13**, 217-244.

[41] Taylor, H.M. and Karlin, S. (1998). *An Introduction to Stochastic Modeling.* Third Edition, Academic Press, San Diego.

[42] Scott, S.L. (2002). Bayesian methods for hidden Markov models: recursive computing in the $21^{th}$ century. *Journal of the American Statistical Association*, **97**, 337-351.

[43] Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Third Edition, Springer, New York.

[44] Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**, 260-269.

[45] Wang, L. and Fu, J.C. (2007). A practical sampling approach for a Bayesian mixture model with unknown number of components. *Statistical Papers*, **48**, 631-653.

[46] Xue, L., Fu, J. C., Wang, F. and Wang, L. (2005). A mixture model approach to analyzing major element chemistry data of the Changjiang (Yangtze River). *Environmetrics*, **16**, 305-318.