A HIDDEN MARKOV RENEWAL MODEL

with Applications to Financial Time Series

A Dissertation

Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Benjamin Preston

December 2015

A HIDDEN MARKOV RENEWAL MODEL

with Applications to Financial Time Series

Benjamin Preston

APPROVED:

Dr. Edward Kao (Committee Chair)
Professor of Mathematics
University of Houston

Dr. Giles Auchmuty
Professor of Mathematics
University of Houston

Dr. Raul Susmel
Associate Professor of Finance
University of Houston

Dr. Ziding Feng
Professor of Biostatistics
University of Texas
MD Anderson Cancer Center

Dr. Shanyu Ji
Professor of Mathematics
University of Houston

Dean, College of Natural Sciences and Mathematics

Dedicated to the loving memory of
Janyce Preston

November 22, 1946 – June 2, 2014

Acknowledgments

I first and foremost wish to express deep gratitude towards my advisor, Professor Edward Kao. It was an honor work with him in this endeavor; I particularly appreciate that he has encouraged my independent exploration. This has engendered a feeling of ownership and mastery that I do not believe would be possible otherwise. Additionally, I am a beneficiary of his strong commitment to teaching and sharing his extensive knowledge of mathematics and computing.

I am also profoundly grateful to the members of advisory committee. As a teacher, Professor Auchmuty imparted useful knowledge and skills that will serve me throughout my professional career. Professor Susmel's work was used in this dissertation; I am honored to have him on the committee. I am grateful to Professor Ziding Feng for taking time from his important work at the MD Anderson Cancer Center on my behalf. Finally, I thank Professor Shanyu Ji for his guidance and support throughout this effort.

Many friends and colleagues have provided interesting, thoughtful, conversations about math, science, and computing. Brian Brock, who mentored me through my first technical job, particularly shaped my interests and passion for the scientific method and learning. Others who I wish to thank are John Haas, Carlos Ortiz, Amy Harris, Anh Tranh, Bill Kilgore, Joel Getchius, James Blakeslee, Matt Euler, and Sam Welsh.

A sincere thanks is owed to my family. I am grateful to my siblings Beky, Ethan, and Doug, who have been continually loving and supportive. Throughout this endeavor Rose, Mac, Ashley, and Gordon have provided constant encouragement – and more

meals than I can count.

I cannot overstate my gratitude towards my parents, Rich and Jan. Throughout my life they instilled a reverence for education and perseverance.

Finally, my wife Kate has been selflessly patient, kind, and encouraging throughout this endeavor. I am deeply grateful for this.

A HIDDEN MARKOV RENEWAL MODEL

with Applications to Financial Time Series

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Benjamin Preston

December 2015

Abstract

Hidden semi-Markov models (HSMMs) are a powerful class of statistical model that have been applied to a wide range of areas such as speech recognition, protein structure prediction, Internet-traffic modeling, financial time-series modeling, and classification of music. Three basic problems of hidden Markov model inference are: Computation of the likelihood, computation of the maximum likelihood-estimate of the model parameters, and computation of the maximum a posteriori estimate of the hidden state sequence. We address these inference problems for a set of models closely related to HSMMs.

Our contributions are: (i) We extend the HSMM to allow observations to depend not only on the current underlying hidden state, but on the next underlying hidden state also. This extension can be used to model behavior whereby the observed data gradually transitions between states, rather than abruptly. (ii) We formulate the hidden portion of the model as a Markov renewal process. This allows us to naturally perform inference on models with hidden events other than state changes, e.g., jumps. (iii) We show that by augmenting the state space of our hidden Markov renewal model (HMRM), we can perform inference on an even larger class of phenomena, including models with stochastic volatility. Hence our HMRM can address three key areas of modern financial time series: regime-switching, jumps, and stochastic volatility.

We develop algorithms to solve the three basic problems of inference for the HMRM. We validate the algorithms by performing inference on simulated data.

We apply our model to two real-world datasets appearing in previously published

analyses. The first dataset contains the log-returns of four European sector indices. Specifications of the HMRM improve the modeling of the auto-correlation function of squared returns compared to the HSMMs used in this first analysis. The second dataset consists of weekly returns from a weighted portfolio of NYSE stocks. Another specification of the HMRM gives improved volatility forecasts compared to the regime-switching GARCH models published in the second analysis.

Contents

List of Figures

1. Introduction

We develop a model we call the hidden Markov renewal model (HMRM). It is based on a powerful class of models of called hidden semi-Markov models (HSMMs), which are an extension to the popular hidden Markov model (HMM). Like their HMM predecessor, HSMMs have been successfully applied to many areas such as speech recognition, protein structure prediction, Internet traffic modeling, financial time series modeling, and classification of music. A longer list of applications can be found in the survey by Yu [36].

As a prelude, we start with two well-known models of increasing complexity: the finite mixture model (MM) and the aforementioned HMM. In the next chapter we extend their potential applicability by introducing the hidden Markov renewal model HMRM.

Each one of these model assumes that there are two processes. The first process produces a sequence of random values that are not observed. This unobserved sequence of random values affects the second process, which produces a sequence of random values that *are* observable. The behavior of both of these processes, the unobserved and observed, is governed by a set of model parameters. In practice, these parameters

are not known, and to be estimated.

The relationship among the random variables and parameters in these types of models can be depicted with a *directed graphical model* (DGM), also known as a *Bayesian network*. DGMs are a powerful formalism that can be used to determine independence properties of a model's variables, thereby aiding in the derivation of inference algorithms [3]. Our use of DGMs, however, will be restricted to providing notional descriptions of the models. Fig. 1.1 on the following page depicts a DGM for a mixture model. There, $\{s_1, s_2, \ldots s_T\}$ is the unobserved sequence, $\{x_1, x_2, \ldots x_T\}$ is the observed sequence. The unknown model parameters are α and ε.

We will show in detail how to estimate both the model parameters and unobserved sequence using the observed sequence.

Notation and Conventions

A complete listing of the notation, symbols, etc. used in this dissertation is given in the nomenclature section at the end of the document. We mention the conventions we employ and some of the most commonly used variable names in our introductory models. First, we use ":" to compactly express a sequence or a vector, e.g., $y_{1:T} = \{y_1, y_2, \ldots, y_T\}$. We extend this notation to allow for multiple indices, e.g., for a $m \times n$ matrix:

$$M_{1:m,1:n} = \begin{pmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{m1} & \cdots & M_{mn} \end{pmatrix}$$

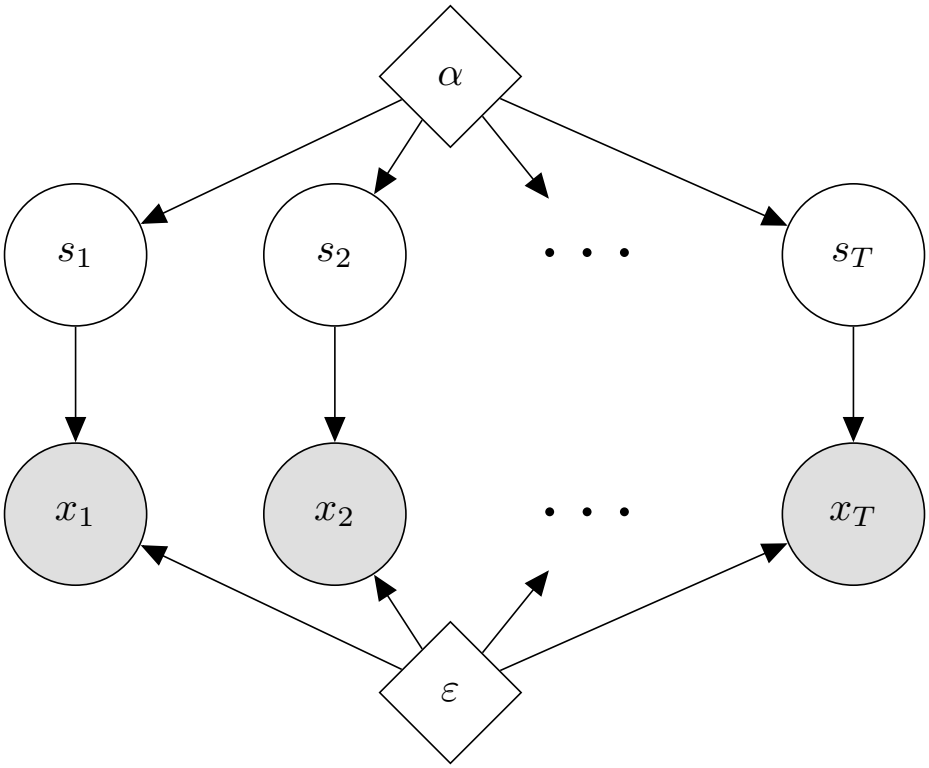


Figure 1.1.: A directed graphical model representing the MM. The shaded nodes represent observed values, unshaded are unobserved. Random variables are represented by circles, while diamond nodes are fixed, but possibly unknown, parameters. The model parameters are α, ε.

We denote $s_{1:T}$ as the unobserved sequence; we sometimes call this the *hidden state* sequence. The observed sequence is $x_{1:T}$, it is said to be emitted from the hidden state sequence, so the observations are sometimes called *emissions*. Each s_t takes a value in the *state space,* $\{1, \ldots, K\}$.

We denote the complete set of a model's parameters as θ; this is an aggregation of multiple sets of parameters. For example in an MM, the parameters $\alpha_{1:K}$ describe the distribution of the state sequence, and $\varepsilon_{1:K}$ describe the distribution of the observations (see Fig. 1.1). So for an MM, $\theta = \{\alpha_{1:K}, \varepsilon_{1:K}\}$. We frequently omit the

parameter set θ when it does not affect a derivation.

name	**meaning**
$s_{1:T}$	The hidden state sequence, s_t is the value of this sequence at time t.
$x_{1:T}$	The observation sequence, x_t is the value of this sequence at time t.
T	The number of observations.
K	The number of values each s_t can take, i.e., $s_t \in \{1, \ldots, K\}$.
θ	The complete set of model's parameters.

We employ a convention of using Greek letters to name the model parameters. Capital Roman letters are used to name variables used in the EM algorithm[1]. Script letters are used in the Viterbi algorithms. We use a hat to denote a maximizer, e.g., $\hat{\theta}$ is the value that maximizes $p(x_{1:T}; \theta)$ over θ. Finally, although it is not always possible, we try to use names that correspond to their meaning, e.g., $\varepsilon_{1:K}$ are the **e**mmissions distributions, $\iota_{1:K}$ is the **i**nitial distribution, $\eta_{1:K,1:K}$ are the **h**olding-time distributions, $\tau_{1:K,1:K}$ is the **t**ransition probability matrix, F_k^t is a **f**orward probability, B_k^t is a **b**ackward probability, and $S_{j,k}^{t,d}$ is a posterior **s**ojourn probability. These terms are explained in the coming sections.

We adopt the same compact notation for expressing probabilities, densities, and mass functions that is use by Gelman et al. [14, pg. 6]. An expression of the form $p(\cdot|\cdot)$ denotes a conditional probability distribution with the arguments determined by the context; similarly $p(\cdot)$ denotes a marginal distribution. We use the terms 'distribution' and 'density' interchangeably; we do this for 'distribution' and 'mass function' also. This allows us to us the same notation for continuous density functions and discrete

[1] The only exception is in the HMRM, where ϕ is **f**orward sojourn probability, and β is the **b**ackward sojourn probability.

probability mass functions. To distinguish parameters from random variables, we place parameters to the right of a semi-colon, e.g. $p(x_t|s_t;\theta)$ is the distribution of x_t given s_t under the parameter set θ.

Specific probability distributions are denoted by a leading calligraphic character, e.g. $\mathcal{N}(\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2. We denote the density or mass function of a distribution $\mathcal{D}(\theta)$ by $f_{\mathcal{D}}(x;\theta)$. When a random variable x is distributed according to $\mathcal{D}$, we write $x \sim \mathcal{D}$. A distribution used frequently in our models is the *categorical* distribution, $\mathcal{C}at(\alpha_{1:K})$. It is defined for all probability vectors $\alpha_{1:K}$, i.e., the elements of $\alpha_{1:K}$ are non-negative and sum to one. We say that $x \sim \mathcal{C}at(\alpha_{1:K})$ if $p(x=k) = \alpha_k$ for $k = 1,\ldots,K$. When $K = 2$, the categorical distribution reduces to the Bernoulli distribution.

Frequently and implicitly we use three results from elementary probability theory: The equivalent definitions of conditional independence, the chain rule of probability, and the law of total probability. A short review of these results follows.

We say, "a set of random variables A is *conditionally independent* of a set of random variables B, given the set of random variables C" if any of the following hold:

$$\begin{aligned} p(A,B|C) = p(A|C)p(B|C) &\iff \\ p(A|B,C) = p(A|C) &\iff \\ p(B|A,C) = p(B|C) & \end{aligned}$$

In such cases, we use the notation $A \perp\!\!\!\perp B|C$ to express the indicated conditional independence. For example, Theorem 1.1 states that in a hidden Markov model

$x_{1:t-1}, s_{1:t-1} \perp\!\!\!\perp x_{t:T}, s_{t+1:T} | s_t$. This is done by showing that $p(x_{t:T}, s_{t+1:T} | x_{1:t-1}, s_{1:t}) = p(x_{t:T}, s_{t+1:T} | s_t)$.

The *chain rule of probability* (CRP)

$$p(A, B) = p(A|B)p(B)$$

is generalized to read:

$$p(A_1, \ldots, A_N) = \prod_{n=1}^{N} p(A_n | A_{1:n-1})$$

The *law of total probability* (LTP) says that the marginal distribution of a set of random variables A can be obtained by summing over all possible values of set of discrete random variables B

$$\sum_B p(A, B) = p(A)$$

and in the case where B are continuous:

$$\int_B p(A, B) = p(A)$$

When we apply this rule we say that B has been *summed out*, or *integrated out*, respectively. We may say B has been *marginalized out* to include either case.

Chapter Outline

The next two sections present the MM and HMM. In each of these sections, we first describe the generation of the hidden state sequence $s_{1:T}$ and observation sequence $x_{1:T}$. We present plots that contain nearly all of the information used to generate the observation sequence. We formally define each model by specifying its probabilistic independence structure. We show how the EM algorithm, which is particularly suited to problems with unobserved values [8], can be used to estimate the model's parameters. We present an algorithm that estimates the hidden state sequence. We conclude each of these sections by demonstrating the model's and corresponding algorithms' feasibility. This is done by simulating data according to the model and estimating the parameters and hidden state sequence using the simulated observation sequence. In the next chapter we introduce the HMRM.

1.1. A Mixture Model

In a finite mixture model [29], the hidden state sequence is generated by iid draws from a $\mathcal{C}at(\alpha_{1:K})$ distribution:

$$p(s_t = k) \quad = \quad \alpha_k$$

We call $\alpha_{1:K}$, the *state distribution*. Consider a mixture model with $K = 3$ possible states, $T = 50$ observations, and state distribution:

$$\alpha_{1:3} = \begin{pmatrix} 0.2 & 0.3 & 0.5 \end{pmatrix} \tag{1.1}$$

The first 10 elements of a state sequence drawn from this model are

$$2, 1, 3, 3, 1, 1, 1, 2, 2, 3, \ldots$$

This state sequence is depicted in Fig. 1.2a on page 10. At each time $t = 1, \ldots, 50$, there is a rectangle whose color corresponds to the value of s_t. For example, since $s_1 = 2$ the left-most rectangle is green. The next rectangle is red, since $s_2 = 1$, and so forth. We will discuss the other elements of Fig. 1.2a, namely the rectangle heights and vertical positions, shortly.

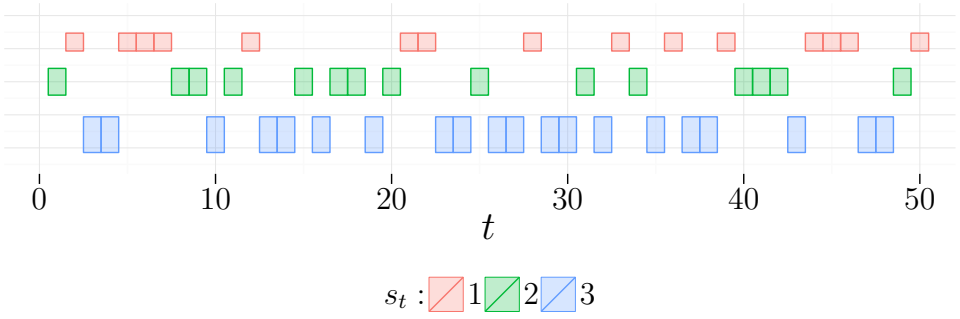
Now we describe how the observation sequence is generated, and how it is affected by the hidden state sequence. Besides the state distribution $\alpha_{1:K}$, the other set of parameters in a mixture model are K probability distributions, called *emission* distributions (or observation distributions). These emission distributions are denoted by $\varepsilon_{1:K}$. Once the state sequence $s_{1:T}$ has been generated, each x_t is drawn from the distribution ε_{s_t}, i.e., if $s_t = k$, then $x_t \sim \varepsilon_k$ where $\varepsilon_k = \mathcal{N}(\mu_k, \sigma_k^2)$.

The observation sequence in Fig. 1.2b was generated from the state sequence depicted in Fig. 1.2a, and emission distributions

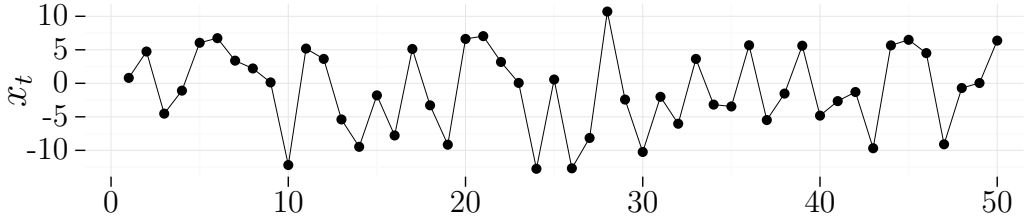
$$\varepsilon_{1:3} = \begin{pmatrix} \mathcal{N}(6, 2^2) & \mathcal{N}(0, 3^2) & \mathcal{N}(-8, 4^2) \end{pmatrix} \tag{1.2}$$

We now address the rectangle heights and vertical positions in Fig. 1.2a. The bottom and top of each rectangle are the first and third quartiles of the corresponding emission distribution in (1.2). The vertical center of each rectangle is the mean of the corresponding emission distribution in (1.2). Fig. 1.2c combines the state sequence depicted in Fig. 1.2a with the observation sequence depicted in Fig. 1.2b. It contains all the information about the parameters of this model and the sequences it generated, with the exception of the state distribution $\alpha_{1:3}$.

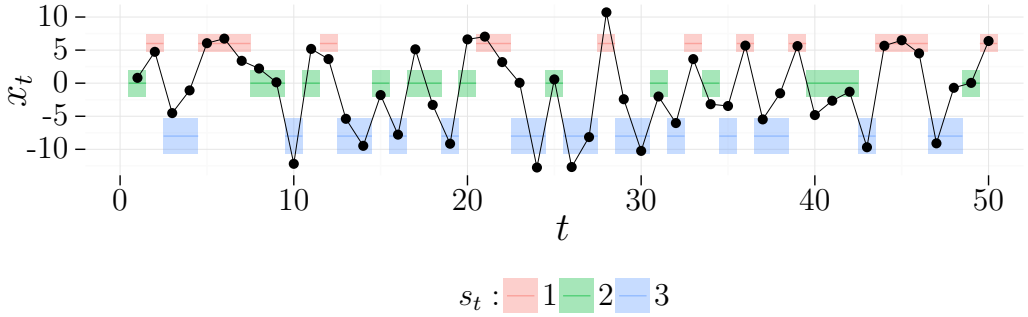
We will present many figures like Fig. 1.2c. These images of rectangles provide an intuitive way of thinking about these models and will provide a visualization of the more elaborate models to be introduced in the sequel.



(a) A state sequence drawn from an MM with $\alpha_{1:3} = \begin{pmatrix} 0.2 & 0.3 & 0.5 \end{pmatrix}$.



(b) An observation sequence generated from the state sequence shown in (a), and the emission distributions specified in (1.2) on the previous page.



(c) The state sequence with the observation sequence. Compared to (a), we have removed the outline of each rectangle, and added a line through the rectangles to indicate the mean of the corresponding emission distribution.

Figure 1.2.: (a) The state sequence from an MM (b) The corresponding observation sequence (c) Both sequences

1.1.1. Formal Definition

The MM is formally defined by equations (1.3). Given s_t, each x_t is distributed according to the emission distribution indexed by s_t. And each x_t is conditionally independent of all other variables in the model.

$$x_t|s_t \quad \sim \quad \varepsilon_{s_t} \qquad \text{for } [t=1,\ldots,T] \quad (1.3a)$$

$$x_t \quad \perp\!\!\!\perp \quad s_{\setminus t}, x_{\setminus t} \quad | \quad s_t \qquad \text{for } [t=1,\ldots,T] \quad (1.3b)$$

Here $v_{\setminus t}$ means all variables in the sequence $v_{1:T}$, excluding the t^{th} value, i.e., $v_{\setminus t} = v_{1:T} \setminus \{v_t\}$. The $s_{1:T}$ are iid according to a $\mathcal{C}at(\alpha_{1:K})$ distribution.

$$s_t \quad \sim \quad \mathcal{C}at(\alpha_{1:K}) \qquad \text{for } [t=1,\ldots,T] \quad (1.3c)$$

$$s_t \quad \perp\!\!\!\perp \quad s_{\setminus t} \qquad \text{for } [t=1,\ldots,T] \quad (1.3d)$$

The complete set of mixture model parameters is $\theta = \{\varepsilon_{1:K}, \alpha_{1:K}\}$. The emission distributions $\varepsilon_{1:K}$ need not be normal, although in this dissertation they will either be normal or a variant of the normal distribution.

1.1.2. Inference

We generated the observations in Fig. 1.2b using the parameters specified in (1.1), (1.2), and the state sequence in Fig. 1.2a. In actuality, we want to do the reverse: we have only the observations, and want to estimate the model parameters and the hidden state sequence producing the observations. We recover the model parameters

by finding a value that (locally) maximizes the likelihood of the observed data. That is, we find the θ that maximizes $p(x_{1:T};\theta)$. Such a θ is called a maximum likelihood estimator and by convention is denoted by:

$$\hat{\theta} \quad \triangleq \quad \arg\max_{\theta} p(x_{1:T};\theta)$$

Dempster et al. [10] showed that the iteration (1.4) converges to a local maximizer of $p(x_{1:T};\theta)$. This iteration is known as the expectation maximization (EM) algorithm .

$$\theta^{(n+1)} \quad \leftarrow \quad \arg\max_{\theta} E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log p(x_{1:T}, s_{1:T};\theta)\right] \tag{1.4}$$

The subscript on the expectation operator indicates that the expectation is to be taken under the posterior probability $p(s_{1:T}|x_{1:T};\theta^{(n)})$. Because the state space is finite, this expectation can be written as a sum over all possible state sequences:

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log p(x_{1:T}, s_{1:T};\theta)\right] \quad = \quad \sum_{s_{1:T}} \log p(x_{1:T}, s_{1:T};\theta) p(s_{1:T}|x_{1:T};\theta^{(n)})$$

Given the model parameters, we can compute the posterior probability of any state sequence, i.e., $p(s_{1:T}|x_{1:T};\theta)$. We will see that for a mixture model, it is straightforward to find the state sequence that maximizes this probability. This state sequence is known as the *maximum a posterior* (MAP) estimate.

1.1.2.1. EM Algorithm

In order to apply the EM algorithm we must find an expression for the expectation in (1.4). The complete data log likelihood (CDLL) can be expressed:

$$\log p(x_{1:T}, s_{1:T}; \theta) \overset{(A.1)}{=} \sum_{t=1}^{T} \log f_{\varepsilon_{s_t}}(x_t) + \sum_{t=1}^{T} \log \alpha_{s_t} \tag{1.5}$$

The expression we derive for (1.5) is based on the *posterior state probabilities*; e.g. the probability that $s_t = k$ after having made the observations $x_{1:T}$. We denote this probability as A_k^t, and find:

$$\begin{aligned} A_k^t &\triangleq p(s_t = k | x_{1:T}; \theta^{(n)}) && \text{for } \begin{bmatrix} t=1,\dots,T \\ k=1,\dots,K \end{bmatrix} \\ &\overset{(A.2)}{=} \frac{f_{\varepsilon_k^{(n)}}(x_t)\, \alpha_k^{(n)}}{\sum_{j=1}^{K} f_{\varepsilon_j^{(n)}}(x_t)\, \alpha_j^{(n)}} \end{aligned}$$

The expectation of each summand in (1.5) is:

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log f_{\varepsilon_{s_t}}(x_t) \right] \overset{(A.3)}{=} \sum_{k=1}^{K} \log f_{\varepsilon_k}(x_t) A_k^t \qquad \text{for } [t=1,\dots,T] \tag{1.6a}$$

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log \alpha_{s_t} \right] \overset{(A.3)}{=} \sum_{k=1}^{K} \log \alpha_k A_k^t \qquad \text{for } [t=1,\dots,T] \tag{1.6b}$$

so the expectation can be written:

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log p(x_{1:T}, s_{1:T}; \theta) \right] \overset{(1.5)\,(1.6)}{=} \underbrace{\sum_{k=1}^{K} \sum_{t=1}^{T} \log f_{\varepsilon_k}(x_t) A_k^t}_{\triangleq Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})} + \underbrace{\sum_{k=1}^{K} \log \alpha_k \sum_{t=1}^{T} A_k^t}_{\triangleq Q_\alpha(\alpha_{1:K}; \theta^{(n)})} \tag{1.7}$$

We note that the quantities $A_{1:K}^{1:T}$ are computed using the EM algorithm's previous iteration's parameter set, $\theta^{(n)} = \{\alpha_{1:K}^{(n)}, \varepsilon_{1:K}^{(n)}\}$, which is to be distinguished from $\theta = \{\alpha_{1:K}, \varepsilon_{1:K}\}$. The value of $\theta = \{\alpha_{1:K}, \varepsilon_{1:K}\}$ that maximizes (1.7) becomes $\theta^{(n+1)}$.

In (1.7) we have defined $Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})$ and $Q_\alpha(\alpha_{1:K}; \theta^{(n)})$. These two quantities partition the expression into two summands. Because $Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})$ does not contain any α terms and $Q_\alpha(\alpha_{1:K}; \theta^{(n)})$ does not contain any ε terms, maximizing each separately maximizes the entire expectation (1.7). The maximizer for $Q_\alpha(\alpha_{1:K}; \theta^{(n)})$ is:

$$\hat{\alpha}_k \overset{(A.4)}{=} \frac{\sum_{t=1}^{T} A_k^t}{T} \qquad \text{for } [k=1,\ldots,K] \quad (1.8)$$

The maximizer for $Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})$ is depedendent on the distributional assumption of ε. In the specific case of the normal distribution, $\hat{\varepsilon}_k$ is $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ with:

$$\begin{aligned} \hat{\mu}_k &\overset{(A.5a)}{=} \frac{\sum_{t=1}^{T} x_t A_k^t}{\sum_{t=1}^{T} A_k^t} && \text{for } [k=1,\ldots,K] \\ \hat{\sigma}_k^2 &\overset{(A.5b)}{=} \frac{\sum_{t=1}^{T} (x_t - \hat{\mu}_k)^2 A_k^t}{\sum_{t=1}^{T} A_k^t} && \text{for } [k=1,\ldots,K] \quad (1.9) \end{aligned}$$

Equations (1.8) and (1.9) are called *update formulas*. Together with (A.2), they form the essential part of the EM algorithm for the MM.

One iteration of the EM algorithm amounts to computing $\hat{\alpha}_{1:K}, \hat{\varepsilon}_{1:K}$ under the parameter set $\theta^{(n)}$, and then setting:

$$\theta^{(n+1)} \quad \leftarrow \quad \{\hat{\alpha}_{1:K}, \hat{\varepsilon}_{1:K}\}$$

1.1.2.2. MAP State Sequence

Because of the MM's independence properties (1.3), the MAP sequence $\hat{s}_{1:T}$ is easily found. The posterior of $s_{1:T}$ factors as follows

$$p(s_{1:T}|x_{1:T}) \overset{(A.1)}{\propto} \prod_{t=1}^{T} p(s_t|x_t)p(x_t) \overset{(1.3b)}{\propto} \prod_{t=1}^{T} p(s_t|x_{1:T})$$

and so each s_t can be maximized separately. The maximizer $\hat{s}_t$ is simply the state that yields the maximum posterior probability:

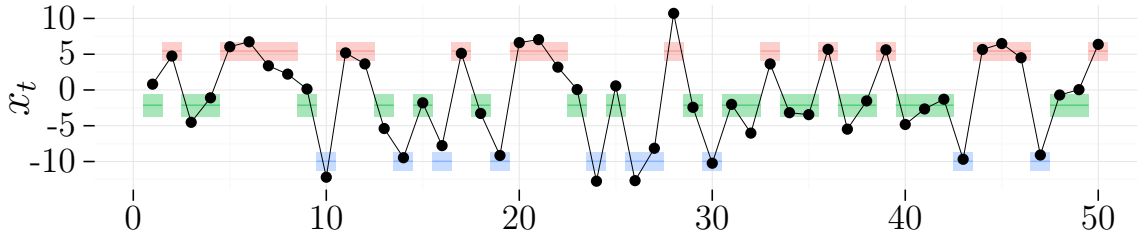
$$\hat{s}_t \quad \leftarrow \quad \arg\max_k A_k^t \qquad \text{for } [t=1,\ldots,T]$$

1.1.3. An Example

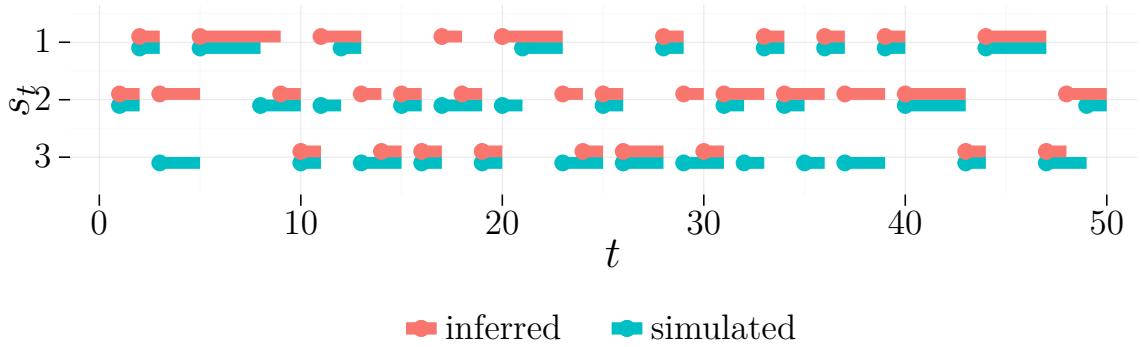
We apply the aforementioned procedures to an observation sequence. Fig. 1.3a is based on the parameter values and state sequence resulting from the inference algorithms applied to the data from Fig. 1.2b. The inferred parameters are

$$\begin{aligned} \hat{\alpha}_{1:3} &= \begin{pmatrix} 0.37 & 0.43 & 0.2 \end{pmatrix} \\ \hat{\varepsilon}_{1:3} &= \begin{pmatrix} \mathcal{N}(5.44, 1.87^2) & \mathcal{N}(-2.14, 2.3^2) & \mathcal{N}(-9.97, 1.93^2) \end{pmatrix} \end{aligned} \tag{1.10}$$

Fig. 1.3b permits a comparison the simulated and inferred state sequences.



(a) The data from Fig. 1.2 with the inferred parameters and state sequence overlayed. The inferred parameters are in (1.10).



(b) Simulated vs. inferred state sequences.

Figure 1.3.: (a) The inferred parameters and state sequence using the data from Fig. 1.2 (b) A comparison of the actual and inferred state sequences.

1.2. A Hidden Markov Model

The difference between the MM and HMM has lies with the hidden state sequence. The MM assumes that the elements of $s_{1:T}$ are iid, whereas the HMM assumes that $s_{1:T}$ is drawn from a finite state Markov chain [30]. After defining a finite state Markov chain, this section proceeds analogously to Section 1.1.

Suppose we have defined for each $t \in \mathbb{N} = \{1, 2, \ldots\}$ a random variable s_t taking values in $\{1, \ldots, K\}$. The process $\{s_t\}_{t \in \mathbb{N}}$ is said to be a Markov chain with state

space $\{1, \ldots, K\}$ provided that

$$p(s_{t+1} = k | s_{1:t}) \quad = \quad p(s_{t+1} = k | s_t) \qquad \text{for } \left[\begin{smallmatrix} t \in \mathbb{N} \\ k \in \{1,\ldots,K\} \end{smallmatrix}\right] \quad (1.11a)$$

Equation (1.11a) is known as the *Markov* property. We assume that $\{s_t\}_{t \in \mathbb{N}}$ is *time-homogeneous*, that the RHS of (1.11a) does not depend on t.

$$\tau_{j,k} \quad \triangleq \quad p(s_{t+1} = k | s_t = j) \qquad \text{for } \left[\begin{smallmatrix} j \in \{1,\ldots,K\} \\ k \in \{1,\ldots,K\} \end{smallmatrix}\right] \quad (1.11b)$$

The probabilities $\tau_{1:K,1:K}$ form the *transition probability matrix* (TPM). In this dissertation, we assume that s_1 is distributed according to an *initial distribution*, $\iota_{1:K}$:

$$\iota_k \quad \triangleq \quad p(s_1 = k) \qquad \text{for } \left[k \in \{1,\ldots,K\}\right] \quad (1.11c)$$

An MC is parameterized by its TPM and and initial distribution. If $s_{1:T}$ is drawn from a process satisfying (1.11), we write

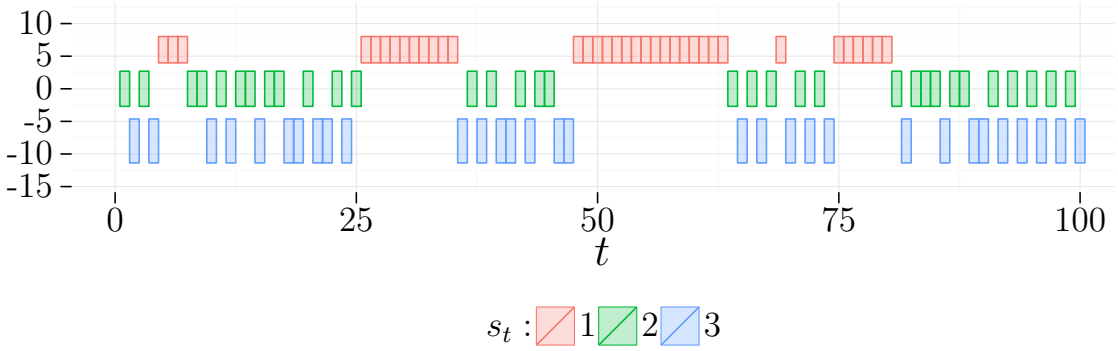
$$s_{1:T} \quad \sim \quad \mathcal{MC}(\iota_{1:K}, \tau_{1:K,1:K})$$

Fig. 1.4a on the following page shows a state sequence drawn from a Markov chain with the following parameters:

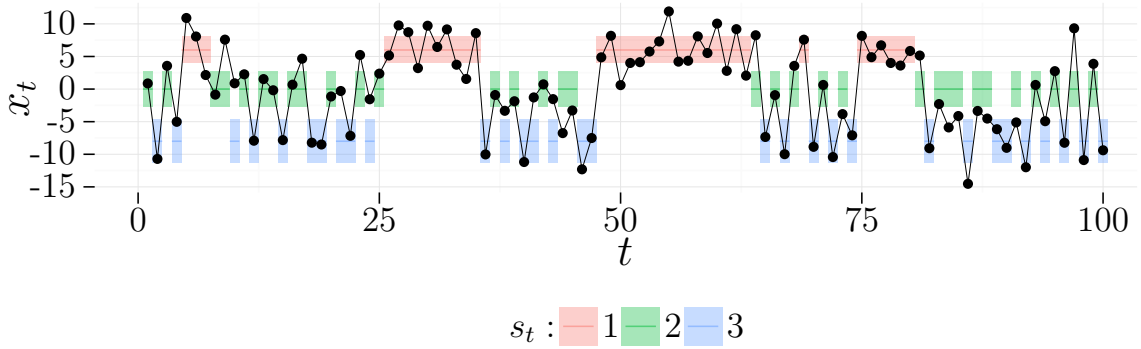
$$\iota_{1:K} = \begin{bmatrix} 0.20 & 0.30 & 0.50 \end{bmatrix} \qquad \tau_{1:K,1:K} = \begin{bmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.10 & 0.80 \\ 0.10 & 0.80 & 0.10 \end{bmatrix} \qquad (1.12)$$

Once the state sequence $s_{1:T}$ is generated, the observation sequence $x_{1:T}$ is produced in the same manner as in the MM: each x_t is drawn from ε_{s_t}. Fig. 1.4b shows an observation sequence generated using $\varepsilon_{1:K}$ below, and the state sequence from Fig. 1.4a.

$$\varepsilon_{1:K} = \begin{bmatrix} \mathcal{N}(6, 3^2) & \mathcal{N}(0, 4^2) & \mathcal{N}(-8, 5^2) \end{bmatrix} \tag{1.13}$$



(a) A state sequence drawn from $\mathcal{MC}(\iota_{1:K}, \tau_{1:K,1:K})$, with parameters as in (1.12).



(b) Observation sequence generated using $\varepsilon_{1:K}$ as in (1.13), and the state sequence as in (a).

Figure 1.4.: (a) A state sequence drawn from a Markov chain (b) The emitted observation sequence

1.2.1. Formal Definition

Equations (1.14) characterize an HMM. Given the state sequence, the observations are distributed as they are in the MM. Each observation is conditionally independent of all other variables given knowledge of its state. Stated succinctly, we have

$$x_t|s_t \quad \sim \quad \varepsilon_{s_t} \qquad \text{for } [t=1,\ldots,T] \quad (1.14a)$$

$$x_t \quad \perp\!\!\!\perp \quad s_{\setminus t}, x_{\setminus t} \quad | \quad s_t \qquad \text{for } [t=1,\ldots,T] \quad (1.14b)$$

In the MM definition (1.3), the $s_{1:T}$ are independent and identically distributed according to $\mathcal{C}at(\alpha_{1:K})$. Whereas in an HMM, $s_{1:T}$ adheres to (1.11). And in this case we write:

$$s_{1:T} \quad \sim \quad \mathcal{MC}(\iota_{1:K}, \tau_{1:K,1:K}) \qquad (1.14c)$$

The entire set of HMM parameters is $\theta = \{\varepsilon_{1:K}, \iota_{1:K}, \tau_{1:K,1:K}\}$.

1.2.2. Inference

As one might expect, inference of both the parameters θ and the hidden state sequence $s_{1:T}$ is more complicated than it is in the MM case. Because the states $s_{1:T}$ are not iid, the posterior state probabilities $p(s_t = k|x_{1:T}; \theta)$ are no longer efficiently calculated as simply as in the MM case, i.e., (A.2).

Baum et al. [2] presented a computationally feasible method for computing these probabilities. Their algorithm has become known as the *Baum-Welch* algorithm, or the *forward-backward* algorithm [28].

We present a theorem that states: Given knowledge of s_t, all the model's random variables before time t are conditionally independent of all the random variables after (or at) time t. We appeal to this theorem frequently as we derive the forward, backward, and other inference algorithms.

Theorem 1.1. *The following conditional independence property holds in an HMM:*

$$x_{1:t-1}, s_{1:t-1} \quad \perp\!\!\!\perp \quad x_{t:T}, s_{t+1:T} \quad | \quad s_t \tag{1.15}$$

Proof. We apply the chain rule of probability and the independence properties of the HMM:

$$\begin{aligned}
&p(x_{t:T}, s_{t+1:T} | x_{1:t-1}, s_{1:t}) \\
&\overset{(CRP)}{=} p(x_{t:T} | x_{1:t-1}, s_{1:T})\, p(s_{t+1:T} | x_{1:t-1}, s_{1:t}) \\
&\overset{(CRP)}{=} \prod_{u=t}^{T} p(x_u | x_{1:u-1}, s_{1:T}) \prod_{u=t+1}^{T} p(s_u | x_{1:t-1}, s_{1:u-1}) \\
&\overset{\substack{(1.14b)\\(1.11a)}}{=} \prod_{u=t}^{T} p(x_u | s_{t:T}) \prod_{u=t+1}^{T} p(s_u | s_{t:u-1}) \\
&\overset{(CRP)}{=} p(x_{t:T}, s_{t+1:T} | s_t)
\end{aligned}$$

□

In the parlance of DGMs (e.g., see [3]), Theorem 1.1 is equivalent to saying that s_t *blocks*, or *d-separates*, any path between $s_{1:t-1}, x_{1:t-1}$ and $s_{t+1:T}, x_{t:T}$. Fig. 1.5 shows the DGM for an HMM.

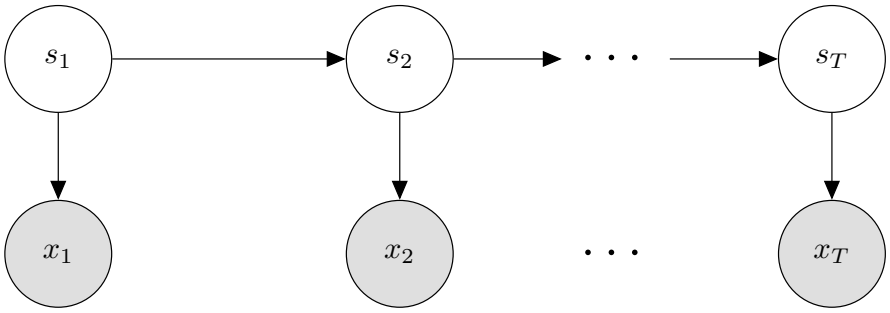


Figure 1.5.: A directed graphical model representing the HMM

Consider the computation of the posterior state probabilities, $p(s_t = k|x_{1:T})$. These probabilities are of intrinsic interest; we may wish to know what state the system was in at time t. They are also needed for the EM algorithm. Using Theorem 1.1, it can be shown that (e.g., see Zucchini and MacDonald [37, (4.10)]):

$$p(s_t = k|x_{1:T}) \overset{(A.7)}{\propto} \overbrace{p(x_{t:T}|s_t = k)}^{\triangleq B_k^t} \overbrace{p(s_t = k, x_{1:t-1})}^{\triangleq F_k^t}$$

This probability is proportional to the forward and backward probabilities. We will show that all of these probabilities can be computed in $\mathcal{O}(TK^2)$ time.

Forward Algorithm The key insight the forward algorithm exploits is that the forward probability $F_k^t \triangleq p(s_t = k, x_{1:t-1})$ can be computed from the previous forward probabilities $F_{1:K}^{t-1}$. The algorithm starts by setting

$$F_k^1 \quad \leftarrow \quad \iota_k \qquad\qquad \text{for } [k=1,\ldots,K]$$

Then for $t = 2, \ldots, T$, we apply the following identity

$$\begin{aligned} F_k^t &\triangleq p(s_t = k, x_{1:t-1}) \\ &= \sum_{j=1}^{K} p(s_t = k, s_{t-1} = j, x_{1:t-1}) \\ &\overset{(CRP)}{=} \sum_{j=1}^{K} p(s_t = k | s_{t-1} = j, x_{1:t-1}) p(x_{t-1} | s_{t-1} = j, x_{1:t-2}) p(s_{t-1} = j, x_{1:t-2}) \\ &\overset{(1.15)}{=} \sum_{j=1}^{K} p(s_t = k | s_{t-1} = j) p(x_{t-1} | s_{t-1} = j) p(s_{t-1} = j, x_{1:t-2}) \\ &= \sum_{j=1}^{K} \tau_{j,k} f_{\varepsilon_j}(x_{t-1}) F_j^{t-1} \end{aligned} \tag{1.16}$$

and set

$$F_k^t \leftarrow \sum_{j=1}^{K} \tau_{j,k} f_{\varepsilon_j}(x_{t-1}) F_j^{t-1} \qquad \text{for } \left[k=1,\ldots,K\right]$$

The forward algorithm executes in $\mathcal{O}(TK^2)$ time; there are $\mathcal{O}(TK)$ many forward probabilities and each requires a sum of K terms.

Fig. 1.6 on the next page illustrates the forward recursion. The black node represents $s_{t+1} = k$. There are K^t many paths to the black node; computing the probability of this many paths is intractable. Any path to the black node must pass through one of the blue nodes. So once we know the probability of each blue node, the probability of the black node can be computed by summing over just the blue nodes – the gray nodes need not be considered. This explains why the forward algorithm is relatively more efficient. The lines following the nodes represent a recursion of this figure; i.e., the forward probability corresponding to each node was computed just as the forward

probability corresponding to the black node is computed in this figure. The following is the correspondence between the computation of F_k^{t+1} and the trajectories shown in Fig. 1.6

$$\underbrace{F_k^{t+1}}_{\text{black node}} = \sum_{j=1}^{K} \underbrace{\tau_{j,k} f_{\varepsilon_j}(x_t)}_{\text{black line}} \times \underbrace{F_j^t}_{\text{blue node}}$$

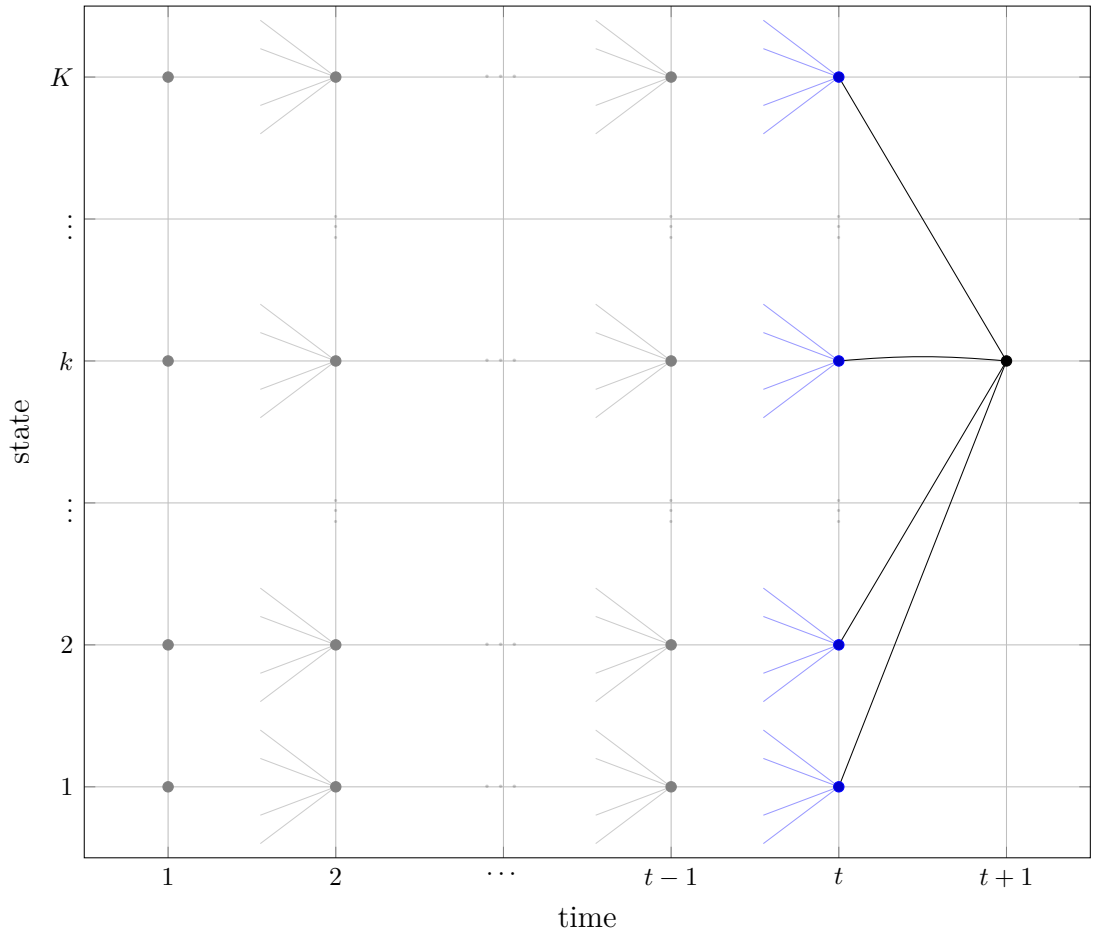


Figure 1.6.: An illustration of the forward and Viterbi HMM algorithms

Backward Algorithm Like the forward algorithm, the backward algorithm computes the backward probabilities using previously computed values. The backward algorithm, however, starts at T and goes backwards in time. It starts by setting:

$$B_T^k \quad \leftarrow \quad \varepsilon_k(x_T) \qquad \text{for } [k=1,\ldots,K]$$

Then for $t = T-1,\ldots,1$, we apply the following identity

$$\begin{aligned}
B_j^t &\triangleq p(x_{t:T}|s_t = j) \\
&\overset{(1.15)}{=} p(x_t|s_t = j)p(x_{t+1:T}|s_t = j) \\
&= p(x_t|s_t = j)\sum_{j=1}^{K} p(x_{t+1:T}, s_{t+1} = k|s_t = j) \\
&\overset{(CRP)}{=} p(x_t|s_t = j)\sum_{j=1}^{K} p(x_{t+1:T}|s_{t+1} = k, s_t = j)p(s_{t+1} = j|s_t = j) \\
&\overset{(1.15)}{=} p(x_t|s_t = j)\sum_{j=1}^{K} p(x_{t+1:T}|s_{t+1} = k)p(s_{t+1} = k|s_t = j) \\
&= f_{\varepsilon_j}(x_t)\sum_{j=1}^{K} B_j^{t+1}\tau_{j,k}
\end{aligned}$$

and set

$$B_j^t \quad \leftarrow \quad f_{\varepsilon_j}(x_t)\sum_{j=1}^{K} B_j^{t+1}\tau_{j,k} \qquad \text{for } [j=1,\ldots,K]$$

1.2.2.1. EM Algorithm

Recall the EM algorithm (1.4). In the case of the HMM, the CDLL is:

$$\log p(x_{1:T}, s_{1:T}; \theta) \overset{(A.6)}{=} \sum_{t=1}^{T} \log f_{\varepsilon_{s_t}}(x_t) + \log \iota_{s_1} + \sum_{t=2}^{T} \log \tau_{s_{t-1}, s_t} \tag{1.17}$$

In addition to the posterior state probabilities $A_{1:T}^{1:K}$, the EM algorithm requires the *posterior transition* probabilities. They are defined:

$$N_{j,k}^{t} \triangleq p(s_t = k | s_{t-1} = j, x_{1:T}; \theta^{(n)}) \qquad \text{for } \begin{bmatrix} t=1,\dots,T \\ j=1,\dots,K \\ k=1,\dots,K \end{bmatrix}$$

The expectation of each summand in (1.17) is:

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log f_{\varepsilon_{s_t}}(x_t) \right] \overset{(A.3)}{=} \sum_{k=1}^{K} \log f_{\varepsilon_k}(x_t) A_k^t \qquad \text{for } [t=1,\dots,T] \tag{1.18a}$$

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log \iota_{s_1} \right] \overset{(A.3)}{=} \sum_{k=1}^{K} \log \iota_k A_k^1 \tag{1.18b}$$

$$E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log \tau_{s_t, s_{t+1}} \right] \overset{(A.3)}{=} \sum_{j=1}^{K} \sum_{k=1}^{K} \log \tau_{j,k} N_{j,k}^{t} \qquad \text{for } [t=2,\dots,T] \tag{1.18c}$$

Combining (1.17) and (1.18), the entire expectation can be written:

$$\begin{aligned} & E_{s_{1:T}|x_{1:T};\theta^{(n)}} \left[\log p(x_{1:T}, s_{1:T}; \theta) \right] \\ &= \underbrace{\sum_{k=1}^{K} \sum_{t=1}^{T} \log f_{\varepsilon_k}(x_t) A_k^t}_{\triangleq Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})} + \underbrace{\sum_{k=1}^{K} \log \iota_k A_k^1}_{\triangleq Q_\iota(\iota_{1:K}; \theta^{(n)})} + \underbrace{\sum_{j=1}^{K} \sum_{k=1}^{K} \log \tau_{j,k} \sum_{t=2}^{T} N_{j,k}^{t}}_{\triangleq Q_\tau(\tau_{1:K,1:K}; \theta^{(n)})} \end{aligned}$$

Because none of the parameters $\{\varepsilon_{1:K}, \iota_{1:K}, \tau_{1:K,1:K}\}$ are shared among $Q_\varepsilon(\varepsilon_{1:K}; \theta^{(n)})$,

$Q_\iota(\iota_{1:K};\theta^{(n)})$, $Q_\tau(\tau_{1:K,1:K};\theta^{(n)})$, maximizing each separately maximizes the entire expectation. The maximizer for $Q_\iota(\iota_{1:K};\theta^{(n)})$ is:

$$\hat{\iota}_k \overset{(A.4)}{=} A^1_k \qquad \text{for } [k=1,\ldots,K] \quad (1.19)$$

The maximizer for $Q_\tau(\tau_{1:K,1:K};\theta^{(n)})$ is

$$\hat{\tau}_{j,k} \overset{(A.9)}{=} \frac{\sum_{t=2}^{T} N^t_{j,k}}{\sum_{l=1}^{K}\sum_{t=2}^{T} N^t_{j,l}} \qquad \text{for } \begin{bmatrix} j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix} \quad (1.20)$$

Since $Q_\varepsilon(\varepsilon_{1:K};\theta^{(n)})$ is the same expression as in the MM case, the maximizers $\hat{\varepsilon}_k$ are also the same. See e.g., (1.9) for the case where $\varepsilon_{1:K}$ are normal distributions. One iteration of the EM algorithm amounts to computing $\hat{\iota}_{1:K}, \hat{\tau}_{1:K,1:K}, \hat{\varepsilon}_{1:K,1:K}$ under the parameter set $\theta^{(n)}$, and then setting:

$$\theta^{(n+1)} \quad \leftarrow \quad \{\hat{\iota}_{1:K}, \hat{\tau}_{1:K,1:K}, \hat{\varepsilon}_{1:K}\}$$

1.2.2.2. MAP State Sequence

Unlike the MM case, the sequence of states with maximum a posteriori probability cannot be obtained by simply maximizing $p(s_t|x_{1:T})$ separately for each s_t. This naive approach often leads to a state sequence very similar to the maximizer of $p(s_{1:T}|x_{1:T})$ [37]. However, such an approach can also lead to an impossible sequence (see Rabiner

[28]). As an example, consider:

$$\iota_{1:3} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \qquad \tau_{1:3,1:3} = \begin{pmatrix} 0 & .5 & .5 \\ .5 & 0 & .5 \\ .5 & .5 & 0 \end{pmatrix}, \qquad \varepsilon_{1:3} = \begin{pmatrix} \mathcal{N}(-1,1) \\ \mathcal{N}(0,1) \\ \mathcal{N}(1,1) \end{pmatrix}$$

Then if $x_{1:2} = \{-1, -1\}$ is observed, $A_1^t = .45$, $A_2^t = .42$, $A_3^t = .13$ for each $t = 1, 2$. Since A_1^t is the maximum of $A_{1:K}^t$ for each of $t = 1, 2$, the naive approach with these parameters and observations leads to $\hat{s}_{1:2} = \{1, 1\}$, which is impossible.

The *Viterbi* algorithm [13] is a dynamic programming algorithm that computes this optimal state sequence. First we define $\mathscr{M}_k^t$ to be the maximum a posteriori probability of all state subsequences $s_{1:t}$ that end with $s_t = k$:

$$\mathscr{M}_k^t \triangleq \max_{s_{1:t-1}} p(s_t = k, s_{1:t-1} | x_{1:T}) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ k=1,\ldots,K \end{bmatrix} \quad (1.21)$$

Similarly to the forward algorithm, we compute $\mathscr{M}_k^t$ efficiently by using the previously computed values $\mathscr{M}_{1:K}^{t-1}$. We define $\mathscr{S}_k^{t-1}$ to be the value of s_{t-1} in (1.21). Equation (1.22) shows how $\mathscr{M}_k^t$ and $\mathscr{S}_k^{t-1}$ can be computed using the previously computed values $\mathscr{M}_{1:K}^{t-1}$.

$$\begin{aligned} \mathscr{M}_k^t &\triangleq \max_{s_{1:t-1}} p(s_t = k, s_{1:t-1} | x_{1:T}) && \text{for } \begin{bmatrix} k=1,\ldots,K \\ t=1,\ldots,T \end{bmatrix} \\ &= \max_{j=1,\ldots,K} \max_{s_{1:t-1}} p(s_t = k, s_{t-1} = j, s_{1:t-2} | x_{1:T}) \\ &\overset{(CRP)}{=} \max_{j=1,\ldots,K} \max_{s_{1:t-2}} p(s_t = k | s_{t-1} = j, s_{1:t-2}, x_{1:T}) \\ &\qquad \times p(s_{t-1} = i, s_{1:t-2} | x_{1:T}) \end{aligned}$$

$$
\begin{aligned}
&\overset{(1.15)}{=} \max_{j=1,\ldots,K} p(s_t = k | s_{t-1} = j, x_{1:T}) \\
&\qquad\qquad \times \max_{s_{1:t-2}} p(s_{t-1} = j, s_{1:t-2} | x_{1:T}) \\
&= \max_{j=1,\ldots,K} \frac{N_{j,k}^{t-1}}{A_j^{t-1}} \mathscr{M}_j^{t-1} \qquad (1.22) \\
\text{with } \mathscr{S}_k^{t-1} &\triangleq \arg\max_{j=1,\ldots,K} \frac{N_{j,k}^{t-1}}{A_j^{t-1}} \mathscr{M}_j^{t-1}
\end{aligned}
$$

So the algorithm begins by setting

$$
\mathscr{M}_{1:K}^1 \leftarrow \iota_{1:K} \qquad \text{for } [k=1,\ldots,K]
$$

then for each $t = 2, \ldots, T$, $\mathscr{M}_k^t$ is computed according to (1.22), the value of j in this equation is stored in $\mathscr{S}_k^{t-1}$:

$$
\begin{aligned}
\mathscr{M}_k^t &\leftarrow \max_{j=1,\ldots,K} \frac{N_{j,k}^t}{A_j^t} \mathscr{M}_j^t \qquad \text{for } [k=1,\ldots,K] \\
\mathscr{S}_k^{t-1} &\leftarrow \arg\max_{j=1,\ldots,K} \frac{N_{j,k}^t}{A_j^t} \mathscr{M}_j^t \qquad \text{for } [k=1,\ldots,K]
\end{aligned}
$$

Once $\mathscr{M}_{1:K}^T$ and $\mathscr{S}_{1:K}^{T-1}$ are computed, we set $\mathscr{S}^T \leftarrow \arg\max_k \mathscr{M}_k^T$. Then the MAP state sequence $\hat{s}_{1:T}$ is constructed in reverse order, starting with $\hat{s}_T \leftarrow \mathscr{S}^T$, and then for each $t = T, \ldots, 2$, $\hat{s}_{t-1} \leftarrow \mathscr{S}_{\hat{s}_t}^{t-1}$.

The Viterbi algorithm shares the same structure as the forward algorithm. It can be

related to the nodes and lines in Fig. 1.6 on page 23 as follows:

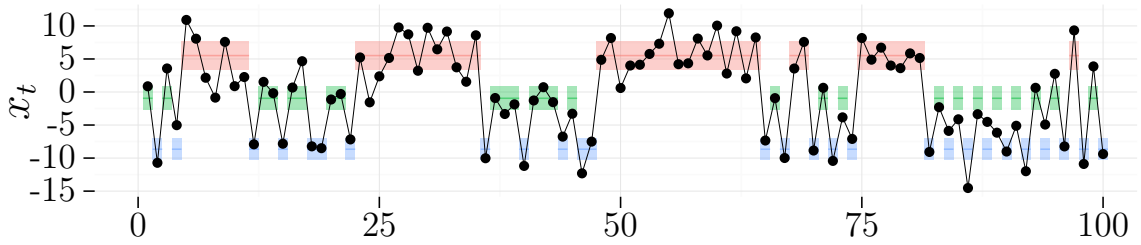
$$\underbrace{\mathscr{M}_k^{t+1}}_{\text{black node}} = \max_{j=1,\ldots,K} \underbrace{\frac{N_{j,k}^t}{A_j^t}}_{\text{black line}} \times \underbrace{\mathscr{M}_j^t}_{\text{blue node}}$$

1.2.3. An Example

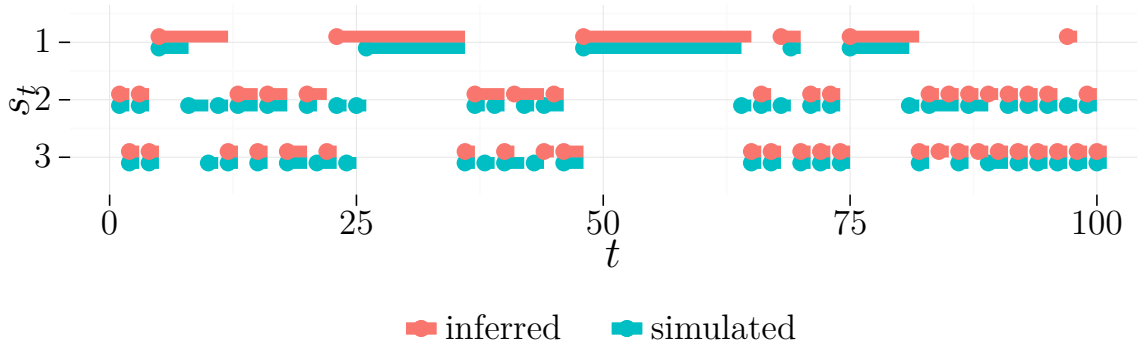
Fig. 1.7a on the next page shows the result of our HMM inference algorithms applied to the simulated observation sequence in Fig. 1.4a on page 18. The inferred parameters are:

$$\hat{\iota}_{1:3} = \begin{bmatrix} 0.00 \\ 1.00 \\ 0.00 \end{bmatrix} \qquad \hat{\varepsilon}_{1:3} = \begin{bmatrix} \mathcal{N}(5.5, 3.2^2) \\ \mathcal{N}(-0.9, 2.7^2) \\ \mathcal{N}(-8.7, 2.4^2) \end{bmatrix}$$
$$\hat{\tau}_{1:3,1:3} = \begin{bmatrix} 0.90 & 0.00 & 0.10 \\ 0.00 & 0.30 & 0.70 \\ 0.30 & 0.60 & 0.10 \end{bmatrix} \tag{1.23}$$

Fig. 1.7b shows the actual and inferred state sequences.



(a) The observation sequence from Fig. 1.4b. The colored rectangles are based on the inferred parameters and state sequence. The inferred parameters are specified in (1.23) on the preceding page.



(b) Simulated vs. inferred state sequences.

Figure 1.7.: (a) The inferred HMM and state sequence using the observation sequence in Fig. 1.4. (b) A comparison of the simulated and inferred state sequences.

2. A Hidden Markov Renewal Model

Hidden Markov models have two major limitations. We develop a model that addresses these limitations by using a Markov renewal process (MRP) in place of the HMM's Markov chain. We call this model a hidden Markov renewal model (HMRM).

The HMRM is based on a class of models, which also address these HMM limitations, called hidden semi-Markov models (HSMMs). These models are based on an hidden semi-Markov process (SMP). In contrast to most (if not all) HSMM authors, we explicitly develop our model as having a hidden Markov renewal process – the term "semi-Markov process" does not appear in our development. The primary reason for this decision is that, compared to an Markov renewal process, the associated semi-Markov process loses information. In particular, an MRP retains the times of all state changes, including times where a state transitions to itself. Such self-transitions are lost when an SMP is used. By exploiting these self-transitions, we can naturally model behavior other than state changes, e.g., the Jump model we present in Section 3.2 and the Stochastic Volatility model of Section 3.3. Our formulation progresses

naturally from the HMM because the MRP is characterized by a property similar to Markov chain's Markov property.

The HMRM extends the capability of the HSMM by allowing observations to depend on adjacent (super)states[1]. This, for example, allows us to model behavior where the observations transition gradually, rather than abruptly, between states. We do this with our Bridging-Means model of Section 3.1. To our knowledge, previous incantations of the HSMM have only allowed for observations to depend on a single (super)state.

The next section elaborates on the aforementioned HMM limitations and describes how they can be rectified by the HMRM. After defining and describing the MRP, we formally define the HMRM and show how we can perform inference for this model. We conclude the chapter with a section on hidden semi-Markov models (HSMMs) and relate the HMRM to some HSMMs we encountered in the literature. In the next chapter we specify some sub-models of our HMRM.

2.1. Two HMM Limitations

HMMs exhibit two major limitations. The first is the modeling of the *holding-time*, which is the contiguous amount of time spent in a state [28]. The holding-time is equivalent to the width of the rectangles in e.g., Fig. 1.4b on page 18. As (2.1) shows,

[1] a *super*state is defined shortly

this quantity is geometrically distributed with parameter $(1 - \tau_{k,k})$ for each state k.

$$\begin{aligned} & p(s_{t+1:t+d-1} = k, s_{t+d} \neq k | s_t = k) \\ & \overset{(1.11a)}{=} \prod_{\delta=1}^{d-1} p(s_{t+\delta} = k | s_{t+\delta-1} = k) p(s_{t+d} \neq k | s_{t+d-1} = k) \\ & = (\tau_{k,k})^{d-1} (1 - \tau_{k,k}) \end{aligned} \tag{2.1}$$

Fig. 2.1 on this page is a particular example of a dataset that the geometric distribution fails to model. It is the word lengths in the English language.

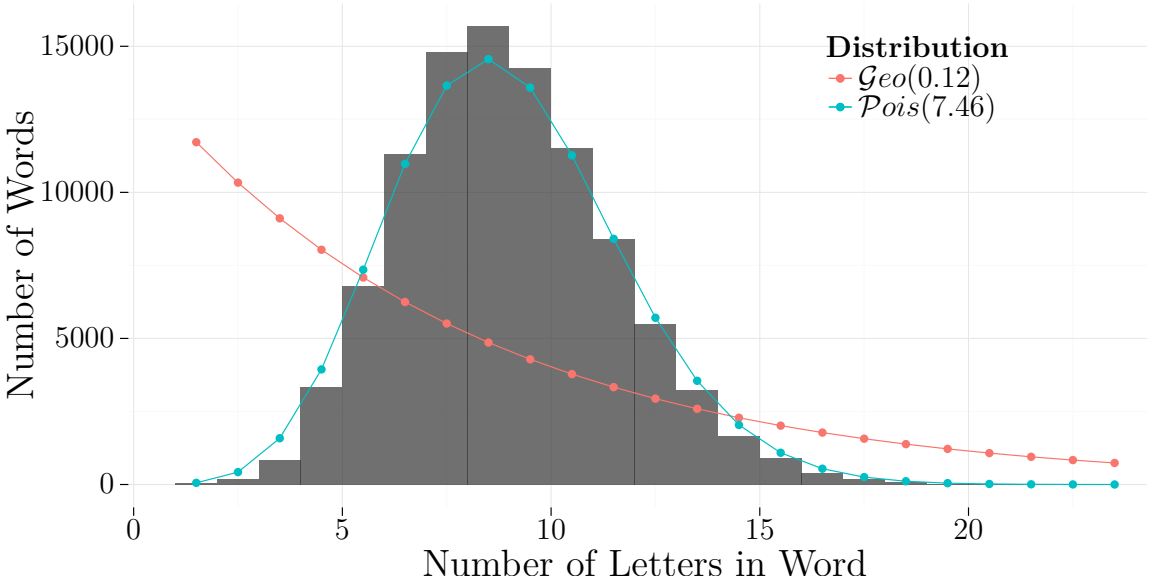


Figure 2.1.: Histogram of word lengths in the English Linux dictionary `/usr/share/dict/words`. Overlayed are the MLEs for the geometric and Poisson distributions.

A second major limitation of the HMM is the strong independence assumption it makes on the observations. Because each observation x_t is associated with a single state s_t, the observations are all conditionally independent, given the underlying state

sequence $s_{1:T}$.

Next we define and describe a Markov renewal process. Then we show how the HMRM addresses the limitations of the HMM.

2.2. Markov Renewal Process

We define a (finite space, discrete time) Markov renewal process (see Çınlar [9], Howard [17], Kao [19], Janssen and Manca [18]). Suppose we have defined for each $r \in \mathbb{N} = \{1, 2, \ldots\}$, a random variable z_r taking values in $\{1, \ldots, K\}$ and a random variable t_r taking values in $\mathbb{N}$ such that $1 = t_1 < t_2 < t_3 < \cdots$. The process $\{(z_r, t_r)\}_{r\in\mathbb{N}}$ is said to be an MRP with state space $\{1, \ldots, K\}$ provided that

$$\begin{aligned} & p(z_{r+1} = k, t_{r+1} - t_r = d | z_{1:r}, t_{1:r}) \\ & \qquad = \quad p(z_{r+1} = k, t_{r+1} - t_r = d | z_r) \qquad\qquad \text{for } \begin{bmatrix} r\in\mathbb{N}, \\ t_{r+1}-t_r\in\mathbb{N} \\ z_r\in\{1,\ldots,K\} \end{bmatrix} \end{aligned} \tag{2.2a}$$

We always assume that $\{(z_r, t_r)\}_{r\in\mathbb{N}}$ is *time-homogeneous*, that the RHS of (2.2a) does not depend on r. Define (as in Çınlar [9, pg. 314]):

$$\kappa^d_{j,k} \quad \triangleq \quad p(z_{r+1} = k, t_{r+1} - t_r = d | z_r = j) \qquad\qquad \text{for } \begin{bmatrix} z_r\in\{1,\ldots,K\} \\ z_{r+1}\in\{1,\ldots,K\} \\ d\in\mathbb{N} \end{bmatrix} \tag{2.2b}$$

The family of probabilities $\kappa^{\mathbb{N}}_{1:K,1:K}$ is called the *semi-Markov kernel*, which is well-

defined because of time-homogeneity. Each of these probabilities can be factored

$$\begin{aligned}\kappa_{j,k}^{d} &= p(z_{r+1}=k, t_{r+1}-t_r=d|z_r=j) \\ &\overset{(CRP)}{=} \underbrace{p(t_{r+1}-t_r=d|z_r=j,z_{r+1}=k)}_{\triangleq f_{\eta_{j,k}}(d)} \underbrace{p(z_{r+1}=k|z_r=j)}_{\triangleq \tau_{j,k}}\end{aligned}$$

We define the *holding-time* distribution $\eta_{j,k}$ in terms of its mass function $f_{\eta_{j,k}}$. We also define the transition probability $\tau_{j,k}$:

$$f_{\eta_{j,k}}(d) \triangleq p(t_{r+1}-t_r=d|z_r=j,z_{r+1}=k) \quad \text{for } \begin{bmatrix} r\in\mathbb{N}, d\in\mathbb{N} \\ z_r\in\{1,\ldots,K\} \\ z_{r+1}\in\{1,\ldots,K\}\end{bmatrix} \tag{2.2c}$$

$$\tau_{j,k} \triangleq p(z_{r+1}=k|z_r=j) \quad \text{for } \begin{bmatrix} r\in\mathbb{N} \\ z_r\in\{1,\ldots,K\}, \\ z_{r+1}\in\{1,\ldots,K\}\end{bmatrix} \tag{2.2d}$$

It remains to specify how z_1, t_1 are distributed. We assume that Markov renewal processes are not *delayed* [9], and that z_1 is distributed according to an initial distribution $\iota_{1:K}$:

$$p(t_1=1) = 1 \tag{2.2e}$$

$$\iota_k \triangleq p(z_1=k) \quad \text{for } [k\in\{1,\ldots,K\}] \tag{2.2f}$$

When a sequence $(z_{1:R+1}, t_{1:R+1})$ is drawn from a process satisfying each of (2.2), and $t_R \leq T < t_{R+1}$, we write:

$$(z_{1:R+1}, t_{1:R+1}) \sim \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, T)$$

A draw from an MRP is shown in Fig. 2.2 on the next page. While we have assumed

that $t_1 = 1$ for all our Markov renewal processes, we do not assume that T, the time of the last observation, coincides with any of $t_{1:R+1}$. For example in Fig. 2.2, $T = 20$.

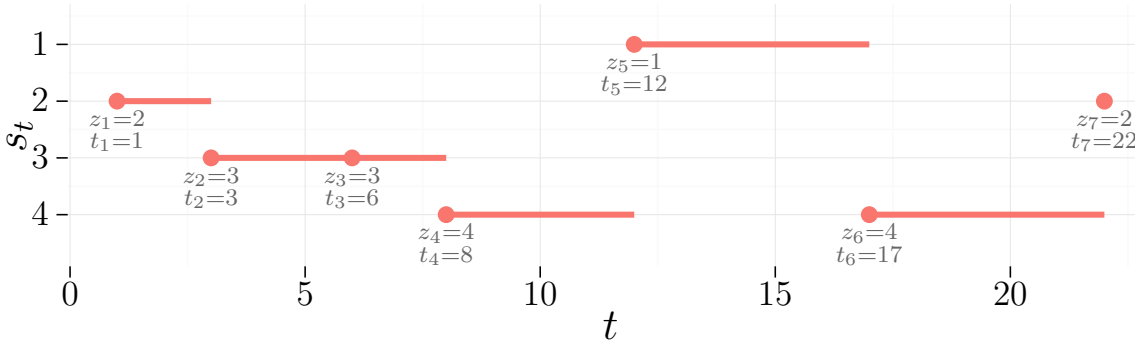


Figure 2.2.: The first 7 renewals of a draw from an MRP, e.g. $(z_{1:7}, t_{1:7}) \sim \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, 20)$.

In the HMM (1.2) each element s_t of the hidden state sequence is associated with a single time. We call $z_{1:R+1}$ a *super*state sequence because each z_r is associated with possibly multiple times, namely $\{t_r, t_r+1, \ldots, t_{r+1}-1\}$. The sequence $t_{1:R+1}$ accounts for time; it is called the *renewal-time* sequence. We further define the pair (z_r, t_r) to be the r^{th} *renewal*, and call adjacent renewals a *sojourn*. For each $r = 1, \ldots, R$ we define $d_r \triangleq t_{r+1} - t_r$ to be the r^{th} *holding-time*[2]. In the context of an MRP, we let $s_{1:T}$ denote the value of the superstate sequence at time t, i.e., $s_t = z_r$ where $t_r \leq t < t_{r+1}$, and call $s_{1:T}$ the *state* sequence. Later, when we incorporate the MRP into the HMRM, we will see that the superstates emit *observation subsequences*. We summarize this new nomenclature in Tab. 2.1 on the following page.

Recall the two limitations of the hidden Markov model: It has implicit geometric holding-time distributions, and an excessively strong independence assumption with

[2]or "sojourn **d**uration"

word	**meaning**	**expression (for r^{th})**
superstate	a hidden value associated with an observation subsequence	z_r
holding-time	the length of an observation subsequence	d_r
renewal-time	the starting time of an observation subsequence	t_r
renewal	a superstate paired with its renewal-time	(z_r, t_r)
sojourn	a pair of adjacent renewals	(z_r, t_r), (z_{r+1}, t_{r+1})
observation subsequence	the observations generated during a sojourn	$x_{t_r:t_{r+1}-1}$
state sequence	the value of the superstate associated with time t	s_t

Table 2.1.: A summary of nomenclature introduced for the HMRM.

regards to the observations. We now develop a model that addresses these limitations. The main idea is that the hidden process now emits subsequences of observations rather than a single observation. The distribution of the lengths of these subsequences need not be geometric. Fig. 2.5 on page 54 shows an HSMM with Poisson distributed holding-times. The HMRM allows us to model each observation subsequence however we wish, in particular, the observations need not be independent of each other within a subsequence. For example, the observation subsequences in Fig. 2.7 on page 57 are drawn from a Wiener process.

We describe how the hidden and observed sequences are generated. First, we draw a renewal sequence

$$(z_{1:R+1}, t_{1:R+1}) \quad \sim \quad \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, T)$$

and then observations

$$x_{t_r:t_{r+1}-1} \sim \varepsilon_{z_r,z_{r+1}} \text{ for } r = 1, \ldots, R-1, \text{ and } x_{t_R:T} \sim \varepsilon_{z_R,z_{R+1}}$$

where $\varepsilon_{1:K,1:K}$ are the emission distributions. In a HSMM, the emission distributions

depend on one superstate, z_r. We have allowed them to depend on both z_r and z_{r+1}.

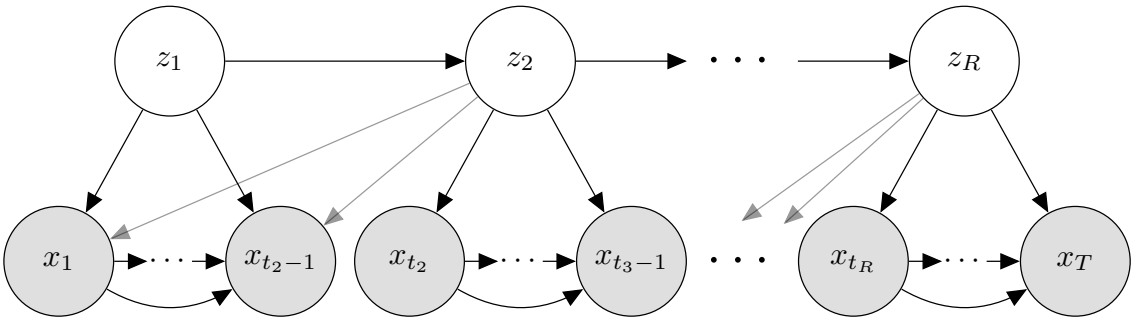


Figure 2.3.: A notional representation of an HMRM. (This is not a valid directed graphical model because the t_r are random, so the structure is not fixed [23, pg. 3]). Whereas the HMM in (see Fig. 1.5) emits a single observation x_t, the HSMM emits a subsequence of observations $x_{t_r:t_{r+1}-1}$. Each subsequence of observations is fully connected, indicating that no independence assumption is made within each subsequence. The gray edges represent our extension.

2.3. Formal HMRM Definition

Equations (2.3) formally define our model. Equations (2.3a) and (2.3b) are analogous to (1.14a) and (1.14b) in the HMM. Equation (2.3a) specifies that $\varepsilon^d_{j,k}$ is the emission distribution for an observation subsequence emitted during a sojourn of length d from superstate j to k. The subscript "$t_{r+1} - 1 \wedge T$" on x allows for the last observation subsequence to get cut off at T, even though t_{R+1} may go past $T+1$. Equation (2.3b) specifies that every observation subsequence is conditionally independent of every other model variable, given the subsequence's sojourn. Recall that the holding-times

are defined $d_r \triangleq t_{r+1} - t_r$.

$$x_{t_r:(t_{r+1}-1)\wedge T} | z_{r:r+1}, t_{r:r+1} \quad \sim \quad \varepsilon^{d_r}_{z_r, z_{r+1}} \qquad \text{for } \left[r \in \{1,\ldots,R\}\right] \tag{2.3a}$$

$$x_{t_r:(t_{r+1}-1)\wedge T} \quad \perp\!\!\!\perp \quad x_{t'}, t_{r'}, z_{r'} \quad | \quad z_{r:r+1}, t_{r:r+1} \qquad \text{for } \left[\begin{smallmatrix} r' \notin \{r, r+1\} \\ t' \notin \{t_r, \ldots, t_{r+1}-1\} \end{smallmatrix}\right] \tag{2.3b}$$

The line below specifies that the renewal sequence is drawn from an MRP with specified parameters, and that $t_R \leq T < t_{R+1}$. Equivalent to this line are the equations (2.2), which define an MRP.

$$(z_{1:R+1}, t_{1:R+1}) \quad \sim \quad \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, T) \tag{2.4}$$

The entire set of HMRM parameters is $\theta = \{\varepsilon_{1:K,1:K}, \iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}\}$.

In order to perform inference on an HMRM, we must additionally make a technical assumption that the holding-time distributions have finite support:

$$\exists D \, s.t. \, d > D \implies f_{\eta_{j,k}}(d) = 0 \quad \text{for each } k = 1, \ldots, K, \, j = 1, \ldots, , K$$

This condition can typically be satisfied practically even for parametric holding-time distributions that do *not* have finite support. This is done by limiting the support to a value D such that $\sum_{d=D+1}^{\infty} f_{\eta_{j,k}}(d) < \epsilon$ for each $j = 1, \ldots, K$, $k = 1, \ldots, K$. We found that setting ϵ to the *machine epsilon* [26, pg. 49] was sufficient to perform the

inference presented in this dissertation. A computer's machine epsilon[3] is the smallest positive number ϵ such that $1 + \epsilon > 1$.

2.4. Inference

Inference with the HMRM resembles inference with the HMM. We follow same outline as the HMM case, first presenting a useful theorem that shows how the model's random variables can be split into two conditionally independent sets. We call on this theorem frequently as we develop our inference algorithms.

An important insight used in our inference algorithms is that the probability that a renewal from superstate j occurs at time t can be computed by summing over all possible sojourns from superstate j beginning at time t to all immediate successor superstates after holding in j for all possible lengths:

$$\underbrace{p(\exists r \, s.t. \, z_r = j, t_r = t)}_{\text{renewal probability}} = \sum_{k=1}^{K} \sum_{d=1}^{D} \underbrace{p(\exists r \, s.t. \, z_r = j, z_{r+1} = k, t_r = t, d_r = d)}_{\text{sojourn probability}} \tag{2.5}$$

Renewal probabilities can, in turn, be used to compute sojourn probabilities:

$$p(\exists r \, s.t. \, z_r = j, z_{r+1} = k, t_r = t, d_r = d) \overset{\substack{(CRP) \\ (A.10)}}{=} \underbrace{p(z_{r+1} = k, d_r = d | z_r = j, t_r = t)}_{\triangleq \eta_{j,k}(d)\,\tau_{j,k}} \underbrace{p(\exists r \, s.t. \, z_r = j, t_r = t)}_{\text{renewal probability}} \tag{2.6}$$

[3]The computer we used had a machine epsilon of $2^{-52} \approx 2.22 \times 10^{-16}$.

The "$\exists$" qualifier in (2.5)-(2.6) provides algorithmic efficiency. We will describe precisely how this qualifier promotes a more efficient forward algorithm after presenting the algorithm.

Analogous to the HMM's Theorem 1.1, Theorem 2.1 shows that given knowledge of a renewal into superstate z_r at time t_r, all the model's random variables before t_r are conditionally independent of all the random variables after (or at) time t_r. This theorem helps motivate the forward/backward algorithm. We cite it frequently in our development of the HMRM inference algorithms.

Theorem 2.1. *The following conditional independence property holds in an HMRM:*

$$x_{1:t_r-1}, z_{1:r-1}, t_{1:r-1} \perp\!\!\!\perp x_{t_r:T}, z_{r+1:R+1}, t_{r+1:R+1} | z_r, t_r \tag{2.7}$$

Proof. Apply the CRP and properties (2.3b), (2.2a).

$$
\begin{aligned}
& p\left(x_{t_r:T}, \begin{smallmatrix} t_{r+1:R+1} \\ z_{r+1:R+1} \end{smallmatrix} \mid x_{1:t_r-1}, \begin{smallmatrix} t_{1:r-1} & t_r \\ z_{1:r-1} & z_r \end{smallmatrix}\right) \\
&\overset{\substack{(2.3b)\\(CRP)}}{=} p\left(x_{t_r:T} \mid \begin{smallmatrix} t_{r+1:R+1} & t_{1:r-1} & t_r \\ z_{r+1:R+1} & z_{1:r-1} & z_r \end{smallmatrix}\right) \\
&\qquad \times p\left(\begin{smallmatrix} t_{r+1:R+1} \\ z_{r+1:R+1} \end{smallmatrix} \mid \begin{smallmatrix} t_{1:r-1} & t_r \\ z_{1:r-1} & z_r \end{smallmatrix}\right) \\
&\overset{(CRP)}{=} \prod_{u=r+1}^{R} p\left(x_{t_u:(t_{u+1}-1)\wedge T} \mid x_{1:t_u-1}, \begin{smallmatrix} t_{r+1:R+1} & t_{1:r-1} & t_r \\ z_{r+1:R+1} & z_{1:r-1} & z_r \end{smallmatrix}\right) \\
&\qquad \times \prod_{u=r+1}^{R+1} p\left(\begin{smallmatrix} t_u \\ z_u \end{smallmatrix} \mid \begin{smallmatrix} t_{1:t_u-1} & t_{1:r-1} & t_r \\ z_{1:t_u-1} & z_{1:r-1} & z_r \end{smallmatrix}\right) \\
&\overset{\substack{(2.3b)\\(2.2a)}}{=} \prod_{u=r+1}^{R} p\left(x_{t_u:(t_{u+1}-1)\wedge T} \mid x_{1:t_u-1}, \begin{smallmatrix} t_{r+1:R+1} & t_r \\ z_{r+1:R+1} & z_r \end{smallmatrix}\right) \\
&\qquad \times \prod_{u=r+1}^{R+1} p\left(\begin{smallmatrix} t_u \\ z_u \end{smallmatrix} \mid \begin{smallmatrix} t_{1:t_u-1} & t_r \\ z_{1:t_u-1} & z_r \end{smallmatrix}\right) \\
&\overset{(CRP)}{=} p\left(x_{t_r:T} \mid \begin{smallmatrix} t_{r+1:R+1} & t_r \\ z_{r+1:R+1} & z_r \end{smallmatrix}\right) \times p\left(\begin{smallmatrix} t_{r+1:R+1} \\ z_{r+1:R+1} \end{smallmatrix} \mid, \begin{smallmatrix} t_r \\ z_r \end{smallmatrix}\right) \\
&\overset{(CRP)}{=} p\left(x_{t_r:T}, \begin{smallmatrix} t_{r+1:R+1} \\ z_{r+1:R+1} \end{smallmatrix} \mid \begin{smallmatrix} t_r \\ z_r \end{smallmatrix}\right)
\end{aligned}
$$

□

To see how this theorem motivates the choice of forward and backward probabilities, consider the posterior probability that a renewal into superstate k occurs at time t, i.e., $p(\exists r\, s.t.\, z_r = k, t_r = t | x_{1:T})$. Theorem 2.1 tells us that this probability can be

written:

$$\begin{aligned} & p(\exists r \, s.t. \, z_r = k, t_r = t | x_{1:T}) \\ & \overset{(A.16)}{\propto} \underbrace{p(\exists r \, s.t. \, z_r = k, t_r = t, x_{1:t-1})}_{=F_k^{t-1}} \underbrace{p(x_{t:T} | z_r = k, t_r = t)}_{=B_k^t} \end{aligned}$$

The probabilities F_k^{t-1} and B_k^t are the forward and backward probabilities (which are defined in (2.8a) and (2.9a), respectively). We will show that these probabilities can be computed in $\mathcal{O}(K^2TD)$ time.

2.4.1. Forward-Backward Algorithm

The forward and backward algorithms each use two types of probabilities: forward/backward sojourn probabilities, and forward/backward renewal probabilities.

2.4.1.1. Forward Algorithm

The forward renewal and sojourn probabilities are defined in (2.8).

$$F_k^t \triangleq p(\exists r \, s.t. \, z_r = k, \, t_r = t+1, \, x_{1:t}) \qquad \text{for } \begin{bmatrix} t=0,\ldots,T \\ k=1,\ldots,K \end{bmatrix} \quad (2.8a)$$

$$\phi_{j,k}^{t,d} \triangleq p\left(\exists r \, s.t. \, {}^{t_r=t+1-d}_{z_r=j}, \, {}^{t_{r+1}=t+1}_{z_{r+1}=k}, \, x_{1:t}\right) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ d=1,\ldots,t\wedge D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix} \quad (2.8b)$$

The upper limit on d in (2.8b) is needed to ensure that $t_r \geq 1$ in this equation. The algorithm starts by setting

$$F_{1:K}^0 \leftarrow \iota_{1:K}$$

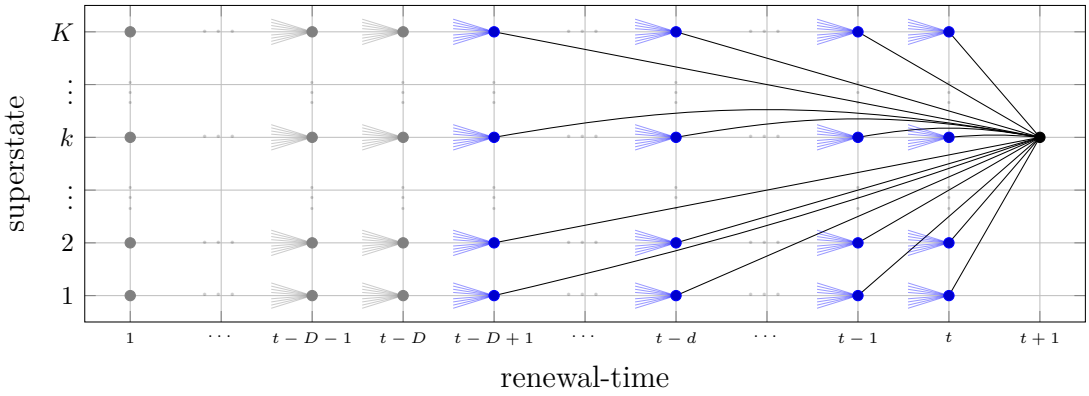


Figure 2.4.: An illustration of the forward HMRM algorithm, and the Viterbi algorithm.

Then for each $t = 1, \ldots, T$, the forward algorithm first computes $\phi_{1:K}^{t,1:t\wedge D}$, and then $F_{1:K}^{t}$.

$$\phi_{j,k}^{t,d} \overset{(A.11)}{\leftarrow} f_{\varepsilon_{j,k}^{d}}(x_{t+1-d:t})\, f_{\eta_{j,k}}(d)\, \tau_{j,k} \times F_{j}^{t-d} \qquad \text{for } \begin{bmatrix} d=1,\ldots,t\wedge D \\ k=1,\ldots,K \end{bmatrix}$$

$$F_{k}^{t} \overset{(A.12)}{\leftarrow} \sum_{j=1}^{K} \sum_{d=1}^{t\wedge D} \phi_{j,k}^{t,d} \qquad \text{for } \begin{bmatrix} k=1,\ldots,K \end{bmatrix}$$

We illustrate how $\phi_{j,k}^{t+1,d}$, and F_{k}^{t+1} correspond to Fig. 2.4 on the current page:

$$\underbrace{F_{k}^{t+1}}_{\text{black node}} = \sum_{j=1}^{K} \sum_{d=1}^{D\wedge t+1} \overbrace{\underbrace{f_{\varepsilon_{j,k}^{d}}(x_{t-d:t+1})\, f_{\eta_{j,k}}(d)\, \tau_{j,k}}_{\text{black line}} \times \underbrace{F_{k}^{t+1-d}}_{\text{blue node}}}^{=\phi_{j,k}^{t+1,d}}$$

In Fig. 2.4, the black node represents the renewal $z_r = k$, $t_r = t + 1$. The blue nodes represent all possible previous adjacent renewals. F_{k}^{t+1} is computed by summing over all these adjacent renewals. The lines following the nodes represent a recursion of this figure; i.e., the forward probability corresponding to each node was computed

just as the black node in this figure. Because holding-times are restricted to be no greater than D, it is impossible for a sojourn starting at a gray node to end at the black node.

An important point that still remains is the use of the "$\exists$" qualifier in the definition of the forward probabilities. By including the "$\exists$" qualifier, we sum over all possible values of r such that $z_r = k$, $t_r = t + 1$

$$p(\exists r \, s.t. \, z_r = k, \, t_r = t + 1) = \sum_{r=1}^{t} p(z_r = k, \, t_r = t + 1)$$

and collapse $\mathcal{O}(t)$ many probabilities into a single probability. Were we to omit the qualifier from our forward sojourn probabilities, instead defining them as

$$\phi_{j,k}^{t,d,r} \triangleq p\left(\begin{smallmatrix} z_r=j \\ t_r=t+1-d \end{smallmatrix}, \begin{smallmatrix} z_{r+1}=k \\ t_{r+1}=t+1 \end{smallmatrix}, x_{1:t}\right) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ r=2\wedge t,\ldots,t \\ d=1,\ldots,t\wedge D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix}$$

there would be $\mathcal{O}(T^2DK^2)$ many forward sojourn probabilities, whereas there are only $\mathcal{O}(TDK^2)$ many with the qualifier.

2.4.1.2. Backward Algorithm

The backward renewal and sojourn probabilities are defined in (2.9).

$$B_j^t \triangleq p(x_{t:T} | z_r = j, \, t_r = t) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ k=1,\ldots,K \end{bmatrix} \quad (2.9a)$$

$$\beta_{j,k}^{t,d} \triangleq p\left(x_{t:T}, \begin{smallmatrix} t_{r+1}=t+d \\ z_{r+1}=k \end{smallmatrix} \mid \begin{smallmatrix} t_r=t \\ z_r=j \end{smallmatrix}\right) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ d=1,\ldots,D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix} \quad (2.9b)$$

Neither B nor β is indexed by r, so there is concern that these probabilities may not be well-defined. But Theorem A.1 shows that these probabilities are constant with respect to r; so indeed, they are well-defined. For each $t = T, \ldots, 1$ the backward algorithm computes $\beta_{1:K}^{t,1:D}$, and then $B_{1:K}^{t}$.

$$\beta_{j,k}^{t,d} \overset{\text{(A.13a)}}{\leftarrow} B_j^{t+d} \times f_{\varepsilon_{j,k}^{d}}(x_{t:t+d-1}) \times f_{\eta_{j,k}}(d)\, \tau_{j,k} \qquad \text{for } \begin{bmatrix} d=1,\ldots,(T-t)\wedge D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix}$$

$$\beta_{j,k}^{t,d} \overset{\text{(A.13b)}}{\leftarrow} f_{\varepsilon_{j,k}^{d}}(x_{t:T}) \times f_{\eta_{j,k}}(d)\, \tau_{j,k} \qquad \text{for } \begin{bmatrix} d=T-t+1,\ldots,D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix}$$

$$B_k^t \overset{\text{(A.14)}}{\leftarrow} \sum_{k=1}^{K} \sum_{d=1}^{D} \beta_{j,k}^{t,d} \qquad \text{for } \begin{bmatrix} k=1,\ldots,K \end{bmatrix}$$

2.4.1.3. Likelihood and Posterior Probabilities

Our EM algorithm for the HMRM uses the likelihood, posterior renewal probabilities, and posterior sojourn probabilities. These are defined in (2.10).

$$L \triangleq p(x_{1:T}) \tag{2.10a}$$

$$E_k^t \triangleq p(\exists r\ s.t.\ z_r = k,\ t_r = t | x_{1:T}) \qquad \text{for } \begin{bmatrix} t=0,\ldots,T \\ k=1,\ldots,K \end{bmatrix} \tag{2.10b}$$

$$S_{j,k}^{t,d} \triangleq p(\exists r\ s.t.\ z_r = j,\ z_{r+1} = k,\ t_r = t,\ t_{r+1} = t+d | x_{1:T}) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T \\ d=1,\ldots,D \\ j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix} \tag{2.10c}$$

These quantities can be computed using the forward-backward probabilities, as we show in (2.11). We first compute the likelihood

$$L \overset{(A.17)}{\leftarrow} \sum_{k=1}^{K} B_k^1 F_k^0 \tag{2.11a}$$

the posterior renewal and sojourn probabilities can then be computed in any order:

$$E_k^t \overset{(A.18)}{=} B_k^t F_k^{t-1} / L \qquad \text{for } \begin{bmatrix} t=0,\dots,T \\ k=1,\dots,K \end{bmatrix} \tag{2.11b}$$

$$S_{j,k}^{t,d} \overset{(A.19)}{=} F_j^{t-1} \beta_{j,k}^{t,d} / L \qquad \text{for } \begin{bmatrix} t=1,\dots,T \\ d=1,\dots,D \\ j=1,\dots,K \\ k=1,\dots,K \end{bmatrix} \tag{2.11c}$$

2.4.2. EM Algorithm

Recall the EM algorithm (1.4). In the case of the HMRM, the CDLL is

$$\begin{aligned} \log p(x_{1:T}, z_{1:R+1}, t_{1:R+1}) \overset{(A.15)}{=} & \log \iota_{z_1} + \sum_{r=1}^{R} \left[\log \tau_{z_r, z_{r+1}} + \log f_{\eta_{z_r, z_{r+1}}}(d_r) \right] \\ & + \sum_{r=1}^{R} \log f_{\varepsilon_{z_r, z_{r+1}}^{d_r}}(x_{t_r:(t_{r+1}-1)\wedge T}) \end{aligned} \tag{2.12}$$

The expectation of each summand is:

$$
\begin{aligned}
Q_\varepsilon(\varepsilon_{1:K,1:K};\theta^{(n)}) &\triangleq E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[\sum_{r=1}^{R}\log f_{\varepsilon^{d_r}_{z_r,z_{r+1}}}(x_{t_r:(t_{r+1}-1)\wedge T})\right] \\
&\overset{(A.20)}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\log f_{\varepsilon^d_{j,k}}(x_{t:(t-1+d)\wedge T})S^{t,d}_{j,k} \qquad (2.13a) \\
Q_\iota(\iota_{1:K};\theta^{(n)}) &\triangleq E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[\sum_{r=1}^{R}\log \iota_{z_1}\right] \\
&\overset{(A.21)}{=} \sum_{k=1}^{K}\iota_k E^1_k \qquad (2.13b) \\
Q_\tau(\tau_{1:K,1:K};\theta^{(n)}) &\triangleq E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[\sum_{r=1}^{R}\log \tau_{z_r,z_{r+1}}\right] \\
&\overset{(A.20)}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\log\tau_{j,k}\sum_{t=1}^{T}\sum_{d=1}^{D}S^{t,d}_{j,k} \qquad (2.13c) \\
Q_\eta(\eta_{1:K,1:K};\theta^{(n)}) &\triangleq E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[\sum_{r=1}^{R}\log f_{\eta_{z_r,z_{t+1}}}(d_r)\right] \\
&\overset{(A.20)}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\log f_{\eta_{j,k}}(d)S^{t,d}_{j,k} \qquad (2.13d)
\end{aligned}
$$

So the entire expectation can be written:

$$
\begin{aligned}
&E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[\log p(x_{1:T},z_{1:R+1},t_{1:R+1})\right] \\
&\overset{(2.12)}{\overset{(2.13)}{=}} Q_\iota(\iota_{1:K};\theta^{(n)}) + Q_\varepsilon(\varepsilon_{1:K,1:K};\theta^{(n)}) + Q_\tau(\tau_{1:K,1:K};\theta^{(n)}) + Q_\eta(\eta_{1:K,1:K};\theta^{(n)})
\end{aligned}
$$

Because none of the parameters $\{\varepsilon_{1:K,1:K},\iota_{1:K},\tau_{1:K,1:K},\eta_{1:K,1:K}\}$ are shared among $Q_\varepsilon(\varepsilon_{1:K,1:K};\theta^{(n)})$, $Q_\iota(\iota_{1:K};\theta^{(n)})$, $Q_\tau(\tau_{1:K,1:K};\theta^{(n)})$, $Q_\eta(\eta_{1:K,1:K};\theta^{(n)})$, maximizing each

separately maximizes the entire expectation. The maximizer for $Q_\iota(\iota_{1:K}; \theta^{(n)})$ is:

$$\hat{\iota}_k \overset{(A.4)}{=} E_k^1 \qquad \text{for } [k=1,\ldots,K] \quad (2.14)$$

The maximizer for $Q_\tau(\tau_{1:K,1:K}; \theta^{(n)})$ is:

$$\hat{\tau}_{j,k} \overset{(A.9)}{=} \frac{\sum_{t=1}^{T} \sum_{d=1}^{D} S_{j,k}^{t,d}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{d=1}^{D} S_{j,k}^{t,d}} \qquad \text{for } \begin{bmatrix} j=1,\ldots,K \\ k=1,\ldots,K \end{bmatrix} \quad (2.15)$$

The maximizer of $Q_\varepsilon(\varepsilon_{1:K,1:K}; \theta^{(n)})$ varies with the form of ε choosen; in the next chapter we present some interesting forms for ε. The maximizer for $Q_\eta(\eta_{1:K,1:K}; \theta^{(n)})$ also varies with the form of $\eta_{j,k}$. Barbu and Limnios [1] show how to find $\hat{\eta}_{j,k}$ for a non-parametric distribution. In the case where the holding-time distributions are dependent only on the current superstate, i.e., $\eta_{j,k} = \eta_j$ for each k (a simpler independence structure), Bulla [5] shows how to find $\hat{\eta}_j$ for geometric and negative binomial distributions, and Ferguson [12] finds it for the Poisson distribution.

One iteration of the EM algorithm amounts to computing $\hat{\varepsilon}_{1:K,1:K}$, $\hat{\iota}_{1:K}$, $\hat{\tau}_{1:K,1:K}$, $\hat{\eta}_{1:K,1:K}$ under the parameter set $\theta^{(n)}$, and then setting:

$$\theta^{(n+1)} \quad \leftarrow \quad \{\hat{\varepsilon}_{1:K,1:K}, \hat{\iota}_{1:K}, \hat{\tau}_{1:K,1:K}, \hat{\eta}_{1:K,1:K}\}$$

A potential mitigator for the increased complexity required by this model is that much of it can be done in parallel, e.g., for each t, all the values of $\phi_{1:K,1:K}^{t,1:D}$ can be computed in any order. This also true for $\beta_{1:K,1:K}^{t,1:D}$, and $S_{1:K,1:K}^{t,1:D}$.

2.4.3. Viterbi Algorithm

The algorithm to find the maximum a posteriori renewal sequence is analogous to the HMM case. There, to find the optimal partial path ending with $s_t = k$, we had to consider all possible previous adjacent states, $s_{t-1} = 1, \ldots, K$. Here, to find the optimal partial path ending with $z_r = k$, $t_r = t$, we must consider all possible previous adjacent renewals, $z_{r-1} = 1, \ldots, K$, $t_{r-1} = t-1, \ldots, t-D$.

We define $\mathscr{M}_k^t$ to be the maximum a posteriori probability of all partial renewal sequences ending with a renewal in superstate k at time t:

$$\mathscr{M}_k^t \quad \triangleq \quad \max_{\substack{r=1,\ldots,t \\ z_{1:r-1}, t_{1:r-1}}} p(z_r = k,\, t_r = t,\, z_{1:r-1},\, t_{1:r-1} | x_{1:T}) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T+D \\ k=1,\ldots,K \end{bmatrix} \quad (2.16)$$

To facilitate recovery of the optimal renewal sequence, we further keep a record of best predecesors [12] $\mathscr{Z}_k^t$, $\mathscr{T}_k^t$. $\mathscr{Z}_k^t$ is the value of z_{r-1} in (2.16) and $\mathscr{T}_k^{t-1}$ is the value of t_{r-1}. (2.17) shows how $\mathscr{M}_k^t$, $\mathscr{Z}_k^t$, and $\mathscr{T}_k^t$ can be computed using the previously computed values $\mathscr{M}_{1:K}^{t-1}$.

$$
\begin{aligned}
\mathscr{M}_k^t &\triangleq \max_{\substack{r=1,\dots,t \\ z_{1:r-1},t_{1:r-1}}} p\left(\begin{smallmatrix} z_r=k \\ t_r=t \end{smallmatrix}, \begin{smallmatrix} z_{1:r-1} \\ t_{1:r-1} \end{smallmatrix} \middle| x_{1:T}\right) \\
&= \max_{\substack{j=1,\dots,K \\ d=1\vee t,\dots,(t-1)\wedge D}} \max_{\substack{r-1=1,\dots,t-d \\ z_{1:r-2},t_{1:r-2}}} p\left(\begin{smallmatrix} z_r=k \\ t_r=t \end{smallmatrix}, \begin{smallmatrix} z_{r-1}=j \\ t_{r-1}=t-d \end{smallmatrix}, \begin{smallmatrix} z_{1:r-2} \\ t_{1:r-2} \end{smallmatrix} \middle| x_{1:T}\right) \\
&\overset{\substack{(CRP) \\ (2.7)}}{=} \max_{\substack{j=1,\dots,K \\ d=1\vee t,\dots,t-1\wedge D}} p\left(\begin{smallmatrix} z_r=k \\ t_r=t \end{smallmatrix} \middle| \begin{smallmatrix} z_{r-1}=j \\ t_{r-1}=t-d \end{smallmatrix}, x_{t:T}\right) \\
&\qquad \times \max_{\substack{r-1=1,\dots,t-d \\ z_{1:r-2},t_{1:r-2}}} p\left(\begin{smallmatrix} z_{r-1}=j \\ t_{r-1}=t-d \end{smallmatrix}, \begin{smallmatrix} z_{1:r-2} \\ t_{1:r-2} \end{smallmatrix} \middle| x_{1:T}\right) \\
&= \max_{\substack{j=1,\dots,K \\ d=1\vee t,\dots,t-1\wedge D}} \frac{\beta_{j,k}^{t-d,d}}{B_j^{t-d}} \times \mathscr{M}_j^{t-d} \qquad (2.17) \\
\textit{with } (\mathscr{Z}_k^t,\, \mathscr{T}_k^t) &\triangleq \operatorname*{arg\,max}_{\substack{j=1,\dots,K \\ t'=1\vee t-D,\dots,(t-1)\wedge T}} \frac{\beta_{j,k}^{t',t-t'}}{B_j^{t'}} \mathscr{M}_j^{t'}
\end{aligned}
$$

The algorithm begins by setting

$$
\mathscr{M}_{1:K}^1 \leftarrow \iota_{1:K}
$$

then for each $t = 2, \dots, T + D$, $\mathscr{M}_k^t$ is computed according to (2.17). The value of j in this equation is stored in $\mathscr{Z}_k^t$. The value of d is used to compute $t - d$, which is

stored in $\mathscr{T}_k^t$.

$$\begin{aligned}
\mathscr{M}_k^t &\leftarrow \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D}} \frac{\beta_{j,k}^{t-d,d}}{B_j^{t-d}} \mathscr{M}_j^{t-d} && \text{for } [k=1,\ldots,K] \\
(\mathscr{Z}_k^t,\, d_k) &\leftarrow \operatorname*{arg\,max}_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D}} \frac{\beta_{j,k}^{t-d,d}}{B_j^{t-d}} && \text{for } [k=1,\ldots,K] \\
\mathscr{T}_k^t &\leftarrow t - d_k && \text{for } [k=1,\ldots,K]
\end{aligned}$$

Once $\mathscr{M}_{1:K}^{T+D}$ are computed, the MAP state sequence $\hat{z}_{1:R+1}, \hat{t}_{1:R+1}$ is constructed in reverse order, starting with the last renewal, which must occur after time T:

$$(\hat{z}_{R+1},\, \hat{t}_{R+1}) \leftarrow \operatorname*{arg\,max}_{\substack{k=1,\ldots,K \\ t=T+1,\ldots,T+D}} \mathscr{M}_k^t$$

And then for each r:

$$(\hat{z}_r,\, \hat{t}_r) \leftarrow (\mathscr{Z}_{\hat{z}_{r+1}}^{\hat{t}_{r+1}},\, \mathscr{T}_{\hat{z}_{r+1}}^{\hat{t}_{r+1}})$$

The algorithm stops when $\hat{t}_r = 1$. As in the HMM case, the Viterbi algorithm shares the same structure as the forward algorithm. It and can be related to the nodes and lines in Fig. 2.4 on page 44 as follows:

$$\underbrace{F_k^{t+1}}_{\text{black node}} = \sum_{j=1}^{K} \sum_{d=1}^{D} \overbrace{\underbrace{f_{\varepsilon_k^d}(x_{t-d:t+1})\, f_{\eta_{j,k}}(d)\, \tau_{j,k}}_{\text{black line}} \times \underbrace{F_k^{t+1-d}}_{\text{blue node}}}^{=\phi_{j,k}^{t+1,d}}$$

2.5. Relation to HSMMs

A variety of models fall under the umbrella of "hidden semi-Markov model". Yu [36] gives a survey of many of these variations. We describe HSMMs using the terminology introduced previously in this chapter. We identify three characterizers that distinguish the different variants of HSMMs and give graphical examples. The section is concluded with a short review of the HSMM literature.

2.5.1. Characterizers of the HSMM

There are three characterizers of an HSMM; they are described in the subsections below. Each has to do with the assumptions that the particular model makes.

2.5.1.1. Independence Assumption Among Renewals

The first characterizer has to do with the independence assumption among the renewals. In the case of the HMRM, this independence assumption is specified by (2.2). According to Yu [36, pg. 226], one of the simplest, most popular dependence structures used in applications is

$$\begin{aligned} z_{r+1}, d_r &\perp\!\!\!\perp z_{1:r-1}, t_{1:r} \mid z_r \\ z_{r+1} &\perp\!\!\!\perp d_r \mid z_r \\ p(z_{r+1} \neq z_r) &= 1 \end{aligned} \tag{2.18}$$

The above is stronger than the MRP independence assumption (2.2) that we use. It assumes that the holding-time depends only on its superstate, and that there is 0 probability of self-transition. This is the independence assumption used in an HSMM called the explicit duration HMM (EDHMM). The data in Fig. 2.5 on the current page was simulated from an EDHMM.

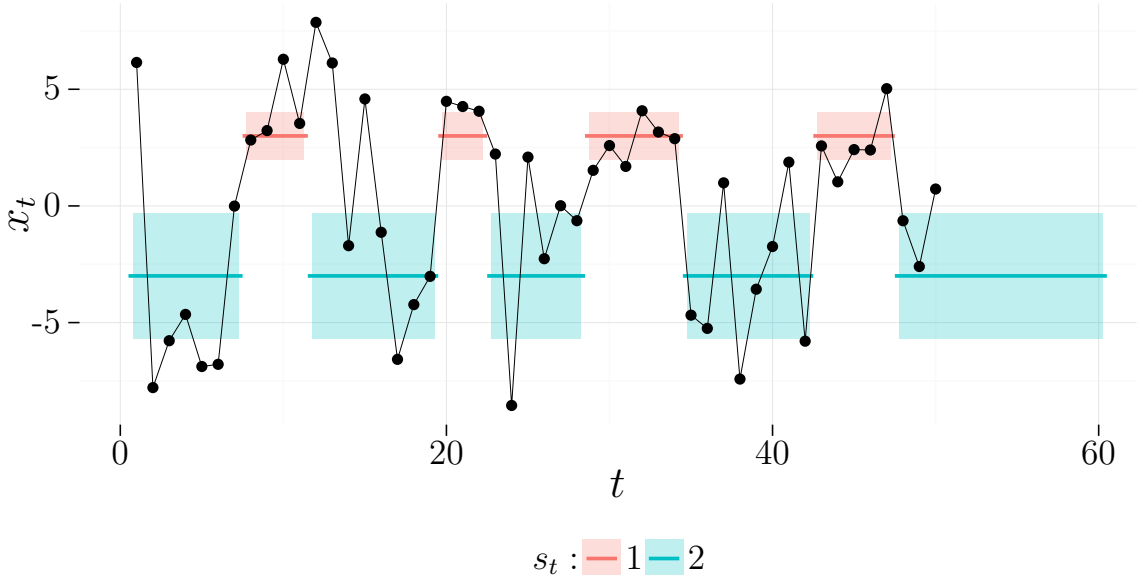


Figure 2.5.: An HSMM with Poisson distributed holding-times; unlike an HMM, the width of each rectangle is not geometrically distributed. Here we have set the diagonal of the transition probability matrix to 0, so the system never transitions back to the same superstate. Each emission distribution assumes independence within the observation subsequence, so the rectangle heights are constant.

In contrast, Fig. 2.6 on the following page shows a simulation using the MRP independence assumption. The diagonal of the transition probability matrix used to generate this figure is set to 0.8, allowing superstate self-transitions. This is manifested by adjacent rectangles of the same color. This weaker assumption also allows the holding-times to depend on their superstate and the next superstate also. Here we

have set the mean holding-time for self-transitions to be 0.1. The mean holding-time for a transition $1 \to 2$ is 3, and $2 \to 1$ is 4. This is reflected in the tendency of wider rectangles to precede a change in color.

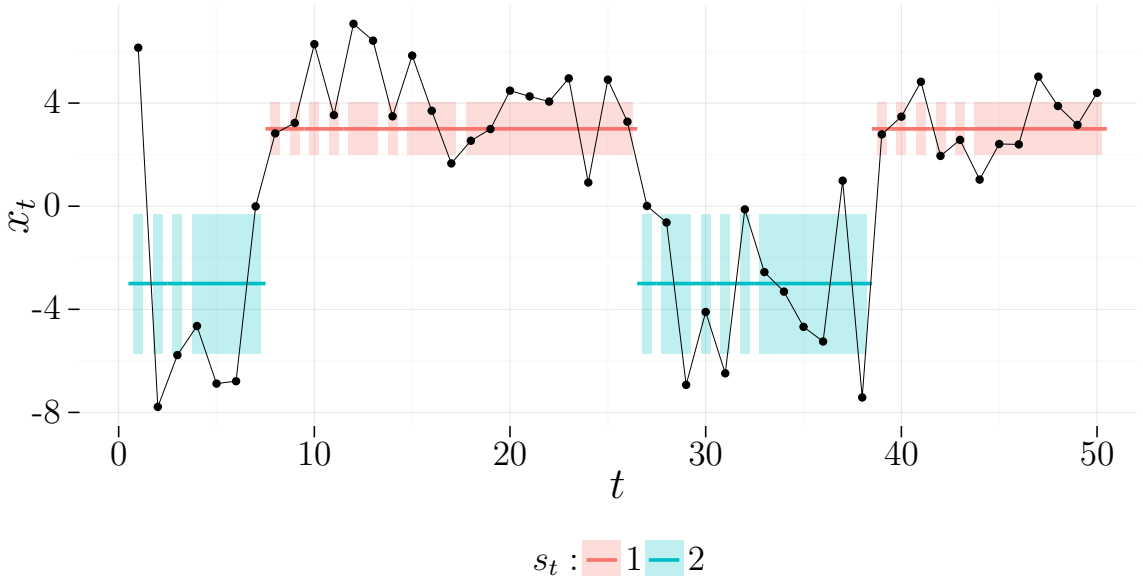


Figure 2.6.: An HSMM with Poisson distributed holding-times with a more elaborate renewal dependence structure than was used in Fig. 2.5. Again we have used emission distributions that assume independence within an observation subsequence.

Not all HSMMs use an underlying SMP to define the hidden process – at least not as it is defined by several authors: Çınlar [9], Howard [17], Kao [19], Janssen and Manca [18] and Pyke [25]. Weaker independence assumptions, leading to more sophisticated independence structures, such as the general model of Yu [36] can be used. Stronger assumptions with less sophisticated independence structures like that of Guédon [15] are also used. So the name "hidden *semi-Markov* model" is perhaps a slight misnomer. The common idea binding models dubbed "HSMM" is that superstates emit subsequences of observations and the length of these subsequences is random.

2.5.1.2. Emission Distribution Assumptions

The 2nd characterizer of an HSMM involves the assumptions made on the emissions distributions ε_k. One possible assumption is that the observations are iid within a observation subsequence, e.g.

$$\varepsilon_k(x_{t:t+d-1}) \quad = \quad \prod_{\delta=0}^{d-1} f_{\mathcal{N}}(x_{t+\delta}; \mu_k, \sigma_k^2)$$

Such distributional assumptions are common. They correspond to Fig. 2.5 and Fig. 2.6. More elaborate emissions distributions are possible. For example, if the observation subsequences follow a Wiener process starting at some point μ_k (for each superstate k) then the emission distribution would be:

$$\varepsilon_k^d(x_{t:t+d-1}) \quad = \quad f_{\mathcal{N}}(x_t; \mu_k, \sigma_k^2) \prod_{\delta=1}^{d-1} f_{\mathcal{N}}(x_{t+\delta}; x_{t+\delta-1}, \sigma_k^2)$$

Such emissions distributions were used for Fig. 2.7 on the next page.

It is with respect to this characterizer that our model differs substantively from previous HSMMs. Our model allows an observation subsequence to depend on both the coinciding superstate and the next. This requires that the emission distributions be indexed by two superstates, i.e., $\varepsilon_{j,k}$.

2.5.1.3. Censoring

The last characterizer of an HSMM is called *censoring*. If we do not assume that the first observation coincides with a renewal time, we say the model is *left cen-*

sored. Similarly, if we do not assume that the last observation immediately precedes a renewal-time, we say the model is *right censored*. Right censoring is illustrated in Fig. 2.5, the overlay indicates that had they been observed, emissions 51 to 60 would have been from superstate 2. Our HMRM is right censored but not left censored.

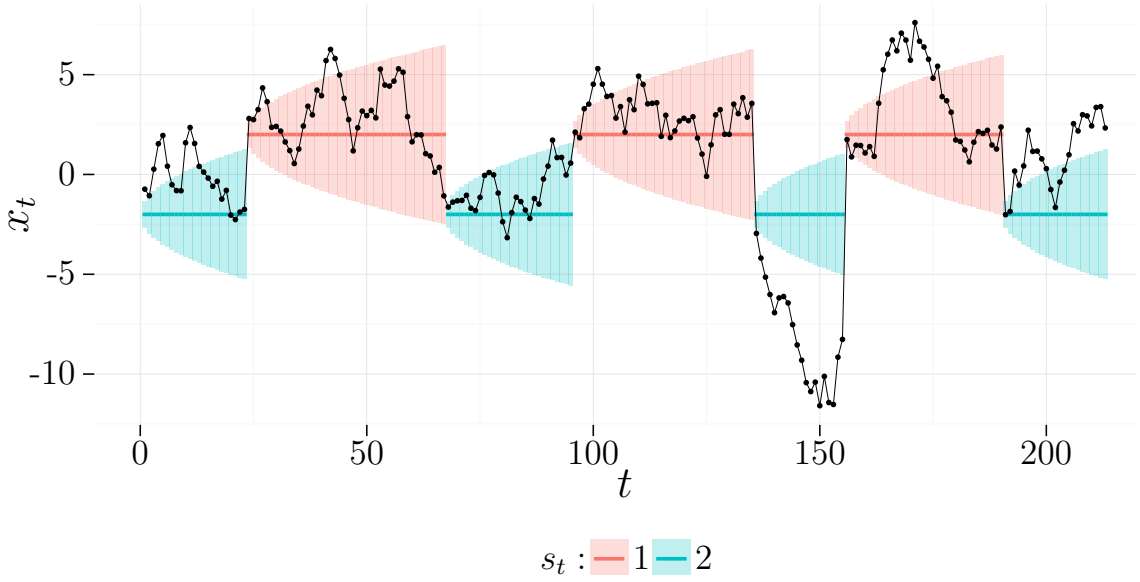


Figure 2.7.: An HSMM with Poisson-distributed holding times. The emission distributions in this case do not assume independence of the observation subsequence, rather these observation subsequences are draws from a Wiener process with starting means at either 2 or -2. The parabolic overlays are determined by the .1, .9 quantiles at the start of each observation subsequence.

2.5.2. Literature Review

Here we review some of the papers that most shaped this dissertation. We describe each paper's model in terms of the three characterizers above: the renewal indendence structure, the emissions distributions, and any censoring assumptions. For each paper

we include a table summarizing these characterizers. We also enumerate any errors we identified during our study of each paper.

2.5.2.1. Ferguson [12]

The "variable duration HMM" of Ferguson [12] was the first appearance of what is now called an HSMM. The renewal independence structure (2.18) was used. No assumption regarding the observation distributions were made in the development of the forward-backward algorithm. However, in his development of the EM algorithm, Ferguson assumed that the observations are iid within an observation subsequence. The development did not address the third characterizer, censoring, except to say "it is messy, but possible, to handle". So the model assumes no censoring.

Ferguson notes that the HMM can be recovered from the HSMM in two ways. This is important because it allows us to use existing HMM routines to test our HMRM routines. The first way[4]the HMM can be recovered from the HSMM is by setting each holding-time distribution to be degenerate at 1, and reusing the transition probability matrix of the HMM to be replicated:

$$p(d_r = d|z_r = j) := \delta_1(d), \qquad \tau_{j,k}^{HSMM} := \tau_{j,k}^{HMM} \qquad \forall j = 1, \ldots, K,\ k = 1, \ldots, K$$

The second way is by setting the hold-time distributions to the geometric distribution with parameters determined by the diagonal of the transition probability matrix in the HMM, and replicating the rest of the HMM:

[4]Inference using this method ends up being substantially faster in practice; performance is comparable to the standard HMM inference algorithms.

$$p(d_r = d|z_r = j) := f_{\mathcal{G}eo}(d; 1 - \tau_{j,j}^{HMM}), \qquad \tau_{j,k}^{HSMM} := \tau_{j,k}^{HMM} \qquad \begin{matrix}\forall j=1,\ldots,K \\ k \neq j\end{matrix}$$

Conversely, Ferguson notes that an HSMM can be embedded in a larger HMM, "with considerable labor". This was done by Langrock and Zucchini [20].

He states that there are three basic problems in the study such models: (i) computation of the likelihood (ii) computation of the MLE (iiia) computation of the maximum a posteriori state sequence and (iiib) computation of the maximum a posteriori state at each time t. He gives solutions (i), (ii), and (iiia) using an adaptation of the forward/backward and EM algorithms of Baum et al. [2]. For (iiib) he develops an adaptation of the Viterbi algorithm [13].

Finally he gives exact reestimation formulas for Poisson and geometrically distributed holding-times, and normally distributed emissions.

characterizer	**model assumptions**
renewal structure	$p(z_{r+1}, t_{r+1}\|z_{1:r}, t_{1:r}) \quad = \quad p(t_{r+1}\|z_r, t_r = t)\, p(z_{r+1}\|z_r)$
emission distributions	No assumptions for forward-backward algorithm. Assumes iid emissions for EM algorithm.
censoring	None

2.5.2.2. Guédon [15]

This paper presented efficient, detailed algorithms for solving the inference problems outlined by Ferguson [12]. These algorithms serve as the basis for two freely available R [27] software packages, those of O'Connell et al. [24] and Bulla et al. [7]. The model

presented in this paper used the same renewal independence structure (2.18). Also assumed was that the observations were iid within an observation subsequence. The model allows for right censoring.

Another interesting aspect of this paper is that it presented an algorithm that is immune to numerical underflow problems. To do this, the algorithm exploits the assumption that observations were iid within an observation subsequence. Our model does not make this assumption and so does not employ Guédon's technique[5].

In the process of testing our implementation of the backward algorithm presented in this paper, we found that it was susceptible to "divide by 0" errors. This can occur in the backward recursion, specifically line -7, pg. 637:

$$G_j(t+1) := G_j(t+1) + L1_j(t+u)\, Observ\, d_j(u)/F_j(t+u)$$

as the value $F_j(t+u)$ can be 0. The problem is alleviated if instead the algorithm stores the quantity, e.g. $L1 dividedByF_j(t) := L1_j(t) + G_k(t+1)p_{jk}$ on line 2, pg. 638, and replaces the computation of $G_j(t+1)$ above with:

$$G_j(t+1) := G_j(t+1) + L1 dividedByF_j(t+u)\, Observ\, d_j(u)$$

Our inspection of the source code in Bulla et al. [7]'s and O'Connell et al. [24]'s R packages indicated that they used some other technique to mitigate this "divide by 0" issue.

[5]Rather we worked with our probabilities in log domain, as in [23], and found that a double-precision machine was sufficiently accurate.

characterizer	**model assumptions**
renewal structure	$p(z_{r+1}, t_{r+1}\|z_{1:r}, t_{1:r}) = p(t_{r+1}\|z_r, t_r = t)\, p(z_{r+1}\|z_r)$ and $p(z_r \neq z_{r+1}) = 1$
emission distributions	Assumes iid emissions.
censoring	Right-censoring

2.5.2.3. Yu [36]

Yu developed a unified model that is the most general we encountered. The model made the weakest renewal independence structure assumption that we saw

$$\begin{aligned} p(z_{r+1}, d_{r+1}|z_{1:r}, d_{1:r}) &= p(z_{r+1}, d_{r+1}|z_r, d_r) \\ p(z_{r+1} \neq z_r) &= 1 \end{aligned}$$

It made no assumptions regarding the emission distribution, and allows for left and/or right censoring.

The paper enumerated several categories that HSMMs can fall into. It showed how these categories arouse by making particular assumptions on the unified model. This paper is a survey, and included over 200 references.

One error we noticed is on line 16 pg. 225. For the first equality there to be true one of the two assumptions must be made: (1) there is a subsequence boundary between t and $t+1$, (2) the observations are independent with a subsequence.

characterizer	**model assumptions**
renewal structure	$p(z_{r+1}, d_{r+1} \mid z_{1:r}, d_{1:r}) \quad = \quad p(z_{r+1}, d_{r+1} \mid z_r, d_r)$ and $p(z_r \neq z_{r+1}) \quad = \quad 1$
emission distributions	Makes no assumption.
censoring	Left and right censoring.

2.5.2.4. Murphy [23]

Murphy's paper began by making the same weak renewal independence structure assumption of Yu – and additionally allowed for superstate self-transitions – but eventually imposed the stronger assumption (2.18). He made no assumptions about the emissions distributions. He noted that this allows the emissions to themselves be an HMM or state space model. By introducing additional variables, he showed how a valid DGM for HSMMs can be constructed. He showed how numerical underflow in HSMM computations, an inherent problem when multiplying many probabilities, can be alleviated by working in log space. Censoring was not addressed in this paper.

characterizer	**model assumptions**
renewal structure	$p(z_{r+1}, t_{r+1} \mid z_{1:r}, t_{1:r}) \quad = \quad p(t_{r+1} \mid z_r, t_r = t)\, p(z_{r+1} \mid z_r)$
emission distributions	Makes no assumption.
censoring	Not addressed.

2.5.2.5. Barbu and Limnios [1] and Malefaki et al. [22]

Malefaki et al. [22] presented an efficient EM algorithm for inference on a model previously developed by Barbu and Limnios [1]. The renewal independence structure in this model is the same as our MRP, although these authors do not allow for self-superstate transitions, as we do. There independence structure allows the holding-time distribution to depend on both the current and next superstates.

characterizer	**model assumptions**
renewal structure	$p(z_{r+1}, d_{r+1} \mid z_{1:r}, d_{1:r}) = p(z_{r+1}, d_{r+1} \mid z_r, d_r)$ and $p(z_r \neq z_{r+1}) = 1$
emission distributions	Assumes emissions are non-parametric, discrete. Barbu and Limnios allow for auto-regressions in the observation subsequences. Malefaki et al. assume observation subsequences are iid.
censoring	Left and right censoring.

3. Some Examples of HMRM Based Models

Chapter 2 presented the hidden Markov renewal model as a general model, and left certain components unspecified. For the HMRM to be of any practical use, we must further specify the following four things: The observation distributions $\varepsilon_{1:K,1:K}$, the holding-time distributions $\eta_{1:K,1:K}$, and corresponding formulas for the maximizers of Q_ε and Q_η. We call such specifications *sub-models*. By specifying different emission and holding-time distributions, different sub-models suitable for different problems can be generated. But the core of the model can remain unchanged.

We present three sub-models, each according to the following template: We specify ε and η, and correspondingly derive maximizers for Q_ε and Q_η. We then simulate data from the sub-model. Finally, we use our inference routines to recover the model's parameters and hidden renewal sequence.

3.1. Bridging-Means Sub-model

One of our primary contributions is an extension of the HSMM that allows the observation subsequences to depend on both current and future superstates. This sub-model highlights that contribution. In particular, each observation is drawn from a normal random variable whose mean is determined by (the mean of) the current superstate and (the mean of) the next superstate. Equation (3.1) specifies the precise manner in which this determination is made.

$$f_{\varepsilon^d_{j,k}}\left(x_{t:(t+d-1\wedge T)}\right) \quad \triangleq \quad \prod_{\delta=0}^{(d-1)\wedge(T-t)} f_{\mathcal{N}}\left(x_{t+\delta}; \frac{d-\delta}{d}\mu_j + \frac{\delta}{d}\mu_k, \sigma^2_{j,k}\right) \tag{3.1}$$

The mean for any observation is a convex combination of the current and future mean. The weight of the combination is determined by the proximity of the observation to the adjacent renewal-times. For an observation occurring in the exact middle of the adjacent renewals, the weights will each be 1/2. Contrastingly, an observation coinciding with a renewal will have the same mean as the associated superstate, with no contribution from the future superstate's mean. Because it leads to a non-analytic update form, we have chosen to develop the model such that the variance does not "bridge" in the same way as the mean[1]. The variance depends on both adjacent superstates, but it is constant for the duration of each observation subsequence.

[1]It may be possible to allow the variance to bridge in this manner but our analysis indicates numerical methods would be needed to solve for each $\hat{\sigma}_{j,k}$.

3.1.1. Efficient Computation of ε

The inference algorithms require the computation of $f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)})$ for each $d = 1,\ldots,D$, $t = 1,\ldots,T$. The naive computation of each $f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)})$ is $\mathcal{O}(d)$, and so the total computational complexity of these ε would be $\mathcal{O}(K^2D^2T)$, rendering our algorithm unfeasible for larger datasets.

We develop a method to compute $f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)})$, for each j,k,t,d, in $\mathcal{O}(K^2DT)$. This is done by storing auxiliary quantities obtained during the computation of $f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)})$, which allows us to compute $f_{\varepsilon^{d+1}_{j,k}}(x_{t:(t+d\wedge T)})$ in $\mathcal{O}(1)$ time. In essence, the improved efficiency relies on the simple idea that $\sum_{\delta=0}^{d+1} x_\delta$ can be computed in $\mathcal{O}(1)$ time once $\sum_{\delta=0}^{d} x_\delta$ has been computed, since $\sum_{\delta=0}^{d+1} x_\delta = x_d + \sum_{\delta=0}^{d} x_\delta$.

From (3.1) and the density of a normal random variable we have:

$$\begin{aligned}\log f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)}) \quad = \quad & -\frac{(d\wedge T-t+1)}{2}(\log 2\pi + 2\log\sigma_{jk}) \\ & -\frac{1}{2\sigma^2_{jk}}\sum_{\delta=0}^{d-1\wedge T-t}(x_{t+\delta}-\frac{d-\delta}{d}\mu_j-\frac{\delta}{d}\mu_k)^2 \qquad (3.2)\end{aligned}$$

With the exception of the sum, (3.2) can be computed in $\mathcal{O}(1)$ time. We rewrite the sum as:

$$\frac{1}{d^2}\sum_{\delta=0}^{(d-1)\wedge(T-t)}\left((d-\delta)\left[x_{t+\delta}-\mu_j\right]+\delta\left[x_{t+\delta}-\mu_k\right]\right)^2$$

Defining $y_{t,j} \triangleq x_t - \mu_j$, this becomes

$$\frac{1}{d^2}\left[\sum_{\delta=0}^{(d-1)\wedge(T-t)}(d-\delta)^2y^2_{t+\delta,j} + 2\sum_{\delta=0}^{(d-1)\wedge(T-t)}(d-\delta)\delta y_{t+\delta,j}y_{t+\delta,k} + \sum_{\delta=0}^{(d-1)\wedge(T-t)}\delta^2y^2_{t+\delta,k}\right]$$

Lemma A.5 on page 133 shows that this is

$$\frac{1}{d^2}\left[U_{t,d}(y_{1:T,j} * y_{1:T,j}) + 2C_{t,d}(y_{1:T,j} * y_{1:T,k}) + V_{t,d}(y_{1:T,k} * y_{1:T,k})\right] \tag{3.3}$$

where "$*$" denotes element-wise multiplication, and U, C, V are defined in Lemma A.5 on page 133.

3.1.2. Optimizing Q_ε

Recall that the EM algorithm for an HMRM requires maximization of:

$$Q_\varepsilon = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \log f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)}) S^{t,d}_{j,k}$$

We derive optimal values of μ, σ for ε defined as in (3.1).

3.1.2.1. Optimal μ

First, we find the optimality condition for each μ_j:

$$\begin{aligned}\frac{\partial Q_\varepsilon}{\partial \mu_j} \quad = \quad & \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \frac{d-\delta}{d^2\sigma^2_{j,k}}(dx_{t+\delta} - (d-\delta)\mu_j - \delta\mu_k) S^{t,d}_{j,k} \\ & + \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \frac{\delta}{d^2\sigma^2_{k,j}}(dx_{t+\delta} - \delta\mu_j - (d-\delta)\mu_k) S^{t,d}_{k,j}\end{aligned}$$

Isolating the μ_j:

$$
\begin{aligned}
= \quad & \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \frac{1}{d^2\sigma_{j,k}^2}[(d-\delta)dS_{j,k}^{t,d}x_{t+\delta} - (d-\delta)^2 S_{j,k}^{t,d}\mu_j - \delta(d-\delta)S_{j,k}^{t,d}\mu_k] \\
& + \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \frac{1}{d^2\sigma_{k,j}^2}[\delta dS_{k,j}^{t,d}x_{t+\delta} - \delta^2 S_{k,j}^{t,d}\mu_j - \delta(d-\delta)S_{k,j}^{t,d}\mu_k]
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d}\sum_{\delta=0}^{(d-1)\wedge(T-t)}(d-\delta)x_{t+\delta} + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d}\sum_{\delta=0}^{(d-1)\wedge(T-t)}\delta x_{t+\delta}\right] \\
& - \sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d^2\sigma_{j,k}^2}\sum_{\delta=0}^{(d-1)\wedge(T-t)}(d-\delta)^2 + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d^2}\sum_{\delta=0}^{(d-1)\wedge(T-t)}\delta^2\right]\mu_j \\
& - \sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d^2}\sum_{\delta=0}^{(d-1)\wedge(T-t)}\delta(d-\delta) + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d^2}\sum_{\delta=0}^{(d-1)\wedge(T-t)}\delta(d-\delta)\right]\mu_k
\end{aligned}
$$

We address the issue discussed in 3.1.1 by replacing the sums over δ with the equivalent from (A.22)-(A.30).

$$
\begin{aligned}
= \quad & \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d}D_{t,d}(x_{1:T}) + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d}E_{t,d}(x_{1:T})\right]}_{\triangleq b_j} \\
& - \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d^2}P_{t,d} + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d^2}Q_{t,d}\right]}_{\triangleq a_j}\mu_j \\
& - \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{T}\left[\frac{1}{\sigma_{j,k}^2}\sum_{d=1}^{D}\frac{S_{j,k}^{t,d}}{d^2}R_{t,d} + \frac{1}{\sigma_{k,j}^2}\sum_{d=1}^{D}\frac{S_{k,j}^{t,d}}{d^2}R_{t,d}\right]}_{\triangleq B_{j,k}}\mu_k
\end{aligned}
$$

$$= b_j - a_j\mu_j - \sum_{k=1}^{K} B_{j,k}\mu_k$$

So the optimal $\hat{\mu}$ satisfies the following matrix equation, with a, b, B defined as above:

$$\underbrace{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} B_{1,1}+a_1 & B_{1,2} & \cdots & B_{1,K} \\ B_{2,1} & B_{2,2}+a_2 & \cdots & B_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ B_{K,1} & B_{K,2} & \cdots & B_{K,K}+a_K \end{pmatrix}}_{B+\mathrm{diag}[a]} \underbrace{\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_K \end{pmatrix}}_{\hat{\mu}}$$

3.1.2.2. Optimal σ

Optimal σ is derived as follows:

$$\frac{\partial Q_\varepsilon}{\partial \sigma_{j,k}} \propto \sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \left[\left(x_{t+\delta} - \frac{d-\delta}{d}\mu_j - \frac{\delta}{d}\mu_k\right)^2 - \sigma_{j,k}^2\right] S_{j,k}^{t,d}$$

So the optimal value for $\sigma_{j,k}$ satisfies

$$\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{(d-1)\wedge(T-t)} \left(x_{t+\delta} - \frac{d-\delta}{d}\mu_j - \frac{\delta}{d}\mu_k\right)^2 S_{j,k}^{t,d} = \sigma_{j,k}^2 \sum_{t=1}^{T}\sum_{d=0}^{D} (d \wedge T - t + 1)\, S_{j,k}^{t,d}$$

Again, the naive computation of the sum over δ is problematic from a performance standpoint. This is readily addressed by employing (3.3):

$$\begin{aligned}\sum_{\delta=0}^{(d-1)\wedge(T-t)} &\left(x_{t+\delta} - \frac{d-\delta}{d}\mu_j - \frac{\delta}{d}\mu_k\right)^2 \\ &= \frac{1}{d^2}\left[U_{t,d}(y^2_{1:T}) + 2C_{t,d}(y_{1:T,j} * y_{1:T,k}) + V_{t,d}(y^2_{1:T,k})\right]\end{aligned}$$

So the optimal $\sigma^2_{j,k}$ is:

$$\hat{\sigma}^2_{j,k} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D}\frac{1}{d^2}\left[U_{t,d}(y_{1:T,j} * y_{1:T,j}) + 2C_{t,d}(y_{1:T,j} * y_{1:T,k}) + V_{t,d}(y_{1:T,k} * y_{1:T,k})\right] S^{t,d}_{j,k}}{\sum_{t=1}^{T}\sum_{d=1}^{D}(d \wedge T - t + 1)\, S^{t,d}_{j,k}}$$

3.1.3. The Holding-Time Distribution η

We define each $\eta_{j,k}$ to be the Poisson distribution with parameter $\lambda_{j,k}$:

$$\eta_{j,k} \quad \triangleq \quad \mathcal{P}ois(\lambda_{j,k}) \tag{3.4}$$

3.1.3.1. Optimizing Q_η

With the $\eta_{1:K,1:K}$ defined as in (3.4) we optimize Q_η over λ.

$$Q_\eta \quad = \quad \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{d=1}^{D}\log f_{\eta_{j,k}}(d)\sum_{t=1}^{T} S^{t,d}_{j,k}$$

Taking the partial derivative gives:

$$\frac{\partial Q_\eta}{\partial \lambda_{j,k}} \quad \propto \quad \sum_{d=1}^{D} \left[\frac{d}{\lambda_{j,k}} - 1 \right] \sum_{t=1}^{T} S_{j,k}^{t,d}$$

So the optimal value for λ is given by:

$$\hat{\lambda}_{j,k} \quad = \quad \frac{\sum_{d=1}^{D} d \sum_{t=1}^{T} S_{j,k}^{t,d}}{\sum_{d=1}^{D} \sum_{t=1}^{T} S_{j,k}^{t,d}}$$

3.1.4. Simulated Data

We first simulate a renewal sequence from an MRP with a Poisson holding-time distribution and the following parameters:

$$\iota \triangleq \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \qquad \tau \triangleq \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \qquad \lambda \triangleq \begin{bmatrix} 14 & 14 \\ 14 & 14 \end{bmatrix} \tag{3.5}$$

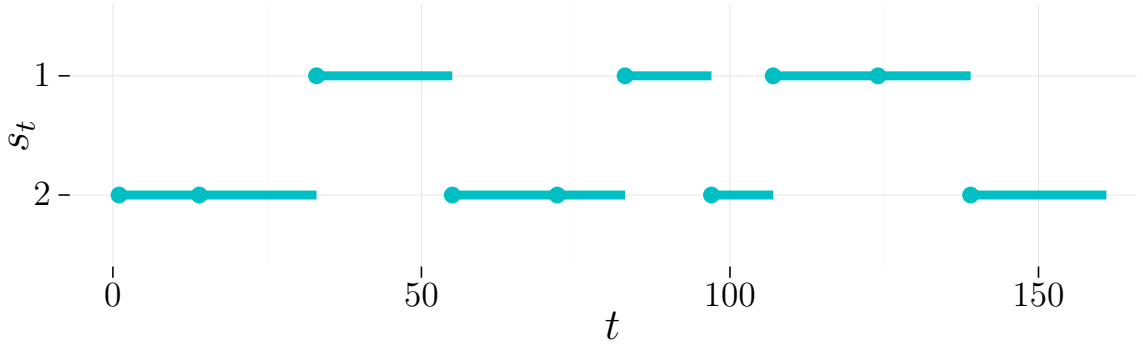
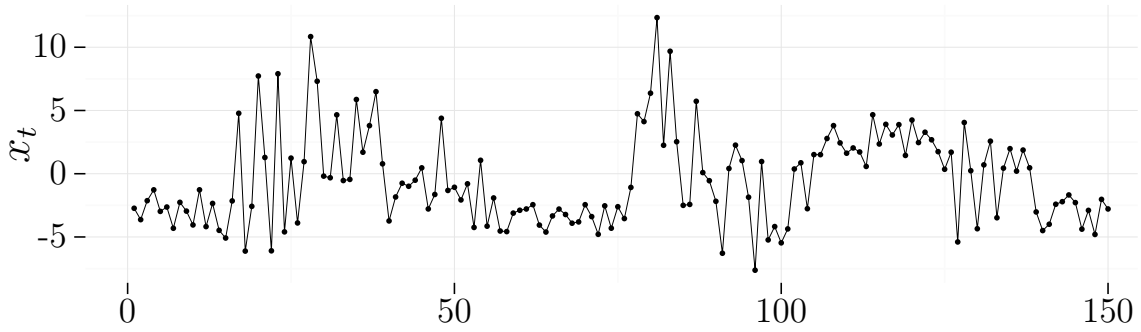


Figure 3.1.: A renewal sequence drawn from an MRP with ι, τ, and $\eta_{j,k} = \mathcal{P}ois(\lambda_{j,k})$ as in (3.5).

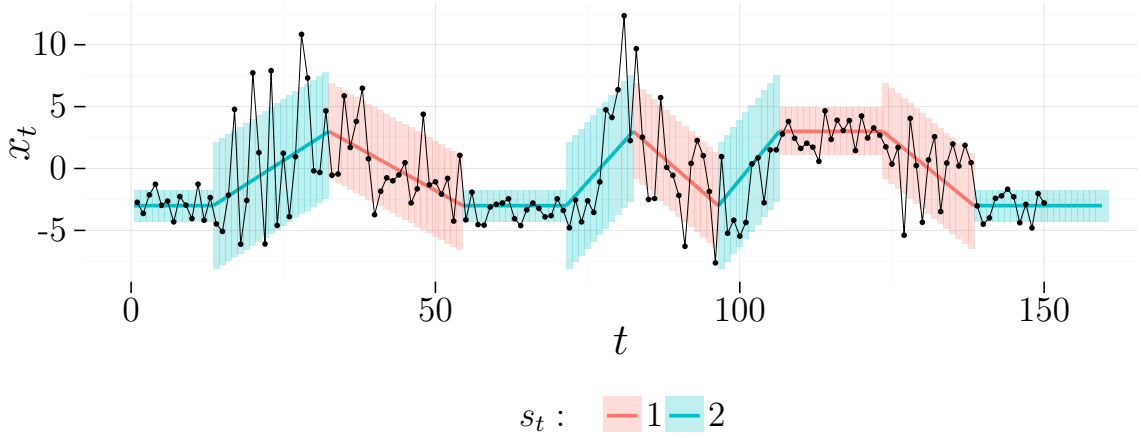
Using this renewal sequence, we simulate an observation sequence with ε as in (3.1) and parameters:

$$\mu \triangleq \begin{bmatrix} 3 \\ -3 \end{bmatrix}, \qquad \sigma \triangleq \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix} \tag{3.6}$$

Fig. 3.1 on the preceding page contains the renewal sequence, Fig. 3.2a on the next page contains the corresponding observation sequence. The information from these 2 plots and (3.6) is all contained in Fig. 3.2b. The parallelograms in the overlays of this figure elucidate our extension: the observations corresponding to these parallelograms depend on adjacent superstates, not a single superstate.



(a) A simulated observation sequence from a bridging-means sub-model with underlying sequence as in Fig. 3.1 and observation model (3.1) with parameters (3.6).



(b) The same observation sequence as in (a). The overlay represents the underlying renewal sequence and the means and variances of the observation sequence at any time. The bottom and top of the overlay at any time t are the .1 and .9 quantiles.

Figure 3.2.: An observation sequence realized from the bridging-means model (a) without overlay (b) with overlay.

The distance between the top and bottom of the overlay at any point is determined by $\sigma_{j,k}$, where j and k are the adjacent superstates. Because our model is right-censored, the end of the observations does not necessarily correspond with the final renewal, as in the case in Fig. 3.2b.

3.1.5. Inference

We attempted to find the MLE of the parameters of the model using the EM algorithm with 1000 different starting parameters. The parameters were chosen randomly in a manner similar to that used by Rydén [31] or Bishop et al. [3]. The starting parameter for ι was set to the uniform distribution. Each row of τ was drawn independently from a symmetric Dirichlet distribution. Each of $\lambda^{j,k}$ was drawn independently from a continuous uniform distribution on the interval $(0, \sqrt{T})$. The μ were sampled from the observations without replacement. Finally, each of $\sigma_{j,k}$ was drawn independently from a continuous uniform distribution on the interval $(0, \sqrt{\max x_{1:T} - \min x_{1:T}})$.

3.1.5.1. Parameters

The parameters we find using the EM algorithm are:

$$\hat{\iota} \triangleq \begin{bmatrix} 0.0 \\ 1.0 \end{bmatrix}, \qquad \hat{\tau} \triangleq \begin{bmatrix} 0.25 & 0.75 \\ 0.50 & 0.50 \end{bmatrix}, \qquad \hat{\lambda} \triangleq \begin{bmatrix} 17 & 19 \\ 8 & 17 \end{bmatrix}$$
$$\hat{\mu} \triangleq \begin{bmatrix} 3 \\ -3 \end{bmatrix}, \qquad \hat{\sigma} \triangleq \begin{bmatrix} 1.1 & 3.0 \\ 5.2 & 1.0 \end{bmatrix} \tag{3.7}$$

3.1.5.2. Renewal Sequence

We use the Viterbi algorithm to find the most likely hidden sequence. We use the inferred parameters listed in (3.7).

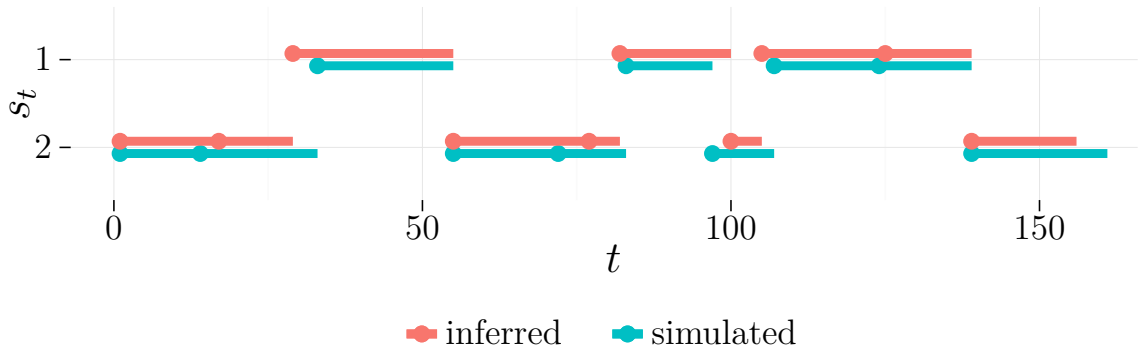


Figure 3.3.: Simulated and inferred renewal sequences for the bridging model (3.1).

3.1.6. Discussion

The bridging-means sub-model requires that we allow each observation subsequence to depend on adjacent superstates as opposed to a single superstate. This is something that cannot be done with a HSMM. So this model elucidates a contribution of our HMRM.

3.2. A Jump Sub-model

We have formulated our HMRM to have a hidden MRP rather than the HSMM's SMP. We now present a sub-model conceived in the context of this formulation. This sub-model illustrates how we can associate events other than superstate changes with a renewal.

The idea behind the model we present is simple: The first observation in each observation subsequence is normal distributed according to a *global* distribution. The remaining observations in the subsequence are normally distributed about this first

observation, with a different, *local*, variance. We can think of the first observation in each subsequence as a "jump" in the observation sequence. Fig. 3.5b on page 80 shows a simulated data set with 8 such jumps.

3.2.1. The Observation Model ε

The emission distributions $\varepsilon_{1:K,1:K}$ are defined:

$$f_{\varepsilon^d_{j,k}}(x_{t:(t+d-1\wedge T)}) \quad = \quad \overbrace{f_{\mathcal{N}}(x_t;\mu_{g(j,k)},\sigma^2_{g(j,k)})}^{\text{jump}} \prod_{\delta=1}^{(d-1)\wedge(T-t)} \overbrace{f_{\mathcal{N}}(x_{t+\delta};x_t,\sigma^2_{l(j,k)})}^{\text{locally distributed}} \quad (3.8)$$

Given that a sojourn from state to j to k begins at time t, $x_t| \sim \mathcal{N}(\mu_{g(j,k)},\sigma^2_{g(j,k)})$, and the remaining observations in the sojourn will be normally distributed about x_t with a variance of $\sigma^2_{l(j,k)}$. So each observation subsequence has it's own mean – the number of means is not restricted by the number of states K. The number of means is determined by the number of renewals in the hidden renewal sequence.

3.2.2. Optimizing Q_ε

We derive EM updates for $\mu_{g(j,k)}$, $\sigma_{g(j,k)}$, and $\sigma_{l(j,k)}$. Here Q_ε can be written:

$$
\begin{aligned}
Q_\varepsilon &= \sum_{j=1}\sum_{k=1}\sum_{t=1}^{T}\sum_{d=1}^{D} \log \varepsilon_{j,k}^{d}(x_{t:(t+d-1\wedge T)}) S_{j,k}^{t,d} \\
&= \sum_{j=1}\sum_{k=1}\sum_{t=1}^{T}\sum_{d=1}^{D} -S_{j,k}^{t,d}\left(\log \sigma_{g(j,k)} + \frac{1}{2}\log 2\pi + \frac{\left(x_t - \mu_{g(j,k)}\right)^2}{2\sigma_{g(j,k)}^2}\right) \\
&\qquad -S_{j,k}^{t,d}\sum_{\delta=0}^{d-1\wedge(T-t)}\left(\log \sigma_{l(j,k)} + \frac{1}{2}\log 2\pi + \frac{\left(x_{t+\delta} - x_t\right)^2}{2\sigma_{l(j,k)}^2}\right)
\end{aligned}
$$

Employing the first order optimality conditions for μ_{global} TODO gives:

$$
\frac{\partial Q_\varepsilon}{\partial \mu_{g(j,k)}} \propto \sum_{t=1}^{T}\sum_{d=1}^{D}\left(x_t - \mu_{g(j,k)}\right) S_{j,k}^{t,d}
$$

$$
\therefore \hat{\mu}_{g(j,k)} = \frac{\sum_{t=1}^{T} x_t \sum_{d=1}^{D} S_{j,k}^{t,d}}{\sum_{t=1}^{T}\sum_{d=1}^{D} S_{j,k}^{t,d}} \overset{\text{(A.18)}}{=} \frac{\sum_{t=1}^{T} x_t B_{j,k}^{t} F_{j,k}^{t-1}}{\sum_{t=1}^{T} B_{j,k}^{t} F_{j,k}^{t-1}}
$$

For $\sigma_{g(j,k)}$:

$$
\frac{\partial Q_\varepsilon}{\partial \sigma_{g(j,k)}} \propto \sum_{t=1}^{T}\left[\left(x_t - \mu_{g(j,k)}\right)^2 - \sigma_{g(j,k)}^2\right]\sum_{d=1}^{D} S_{j,k}^{t,d}
$$

$$
\therefore \hat{\sigma}_{g(j,k)}^2 = \frac{\sum_{t=1}^{T}\left(x_t - \mu_{g(j,k)}\right)^2 \sum_{d=1}^{D} S_{j,k}^{t,d}}{\sum_{t=1}^{T}\sum_{d=1}^{D} S_{j,k}^{t,d}} = \frac{\sum_{t=1}^{T}\left(x_t - \mu_{g(j,k)}\right)^2 B_{j,k}^{t} F_{j,k}^{t-1}}{\sum_{t=1}^{T} B_{j,k}^{t} F_{j,k}^{t-1}}
$$

For $\sigma_{l(j,k)}$:

$$
\frac{\partial Q_\varepsilon}{\partial \sigma_{l(j,k)}} \propto \sum_{t=1}^{T}\sum_{d=2}^{D} S_{j,k}^{t,d} \sum_{\delta=1}^{(d-1)\wedge(T-t)}\left[\left(x_{t+\delta} - x_t\right)^2 - \sigma_{l(j,k)}^2\right]
$$

$$\therefore \hat{\sigma}^2_{l(j,k)} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D} S_{j,k}^{t,d} \sum_{\delta=1}^{(d-1)\wedge(T-t)} (x_{t+\delta} - x_t)^2}{\sum_{t=1}^{T}\sum_{d=1}^{D} S_{j,k}^{t,d} [d-1 \wedge (T-t)]}$$

3.2.3. Emphasizing Renewals' Importance

To emphasize the importance of our formulation of the HSMM as having a hidden MRP rather than SMP, we present the case of our Jump sub-model were there is only a single superstate – hence superstate changes don't occur. Because there is no switching among multiple superstates, the hidden process is a renewal process (e.g., see Çınlar [9, Chapter 9]). We could call this a *hidden renewal model* (HRM). Fig. 3.4 on the next page depicts a draw from a renewal process. Note that the corresponding semi-Markov sequence contains no information because it is constant for all t. Renewals, particularly the times of the renewals, are indispensable in this model.

3.2.4. Simulated Data

Because $K = 1$ here, we omit the super-state subscripts for the rest of this section.

The values of the parameters we use to simulate data from this model are:

$$\lambda \triangleq 20, \qquad \mu_g \triangleq 0, \qquad \sigma_g \triangleq 10, \qquad \sigma_l \triangleq 8 \tag{3.9}$$

A realization from the renewal process underlying the observations is depicted in Fig. 3.4 on the following page.

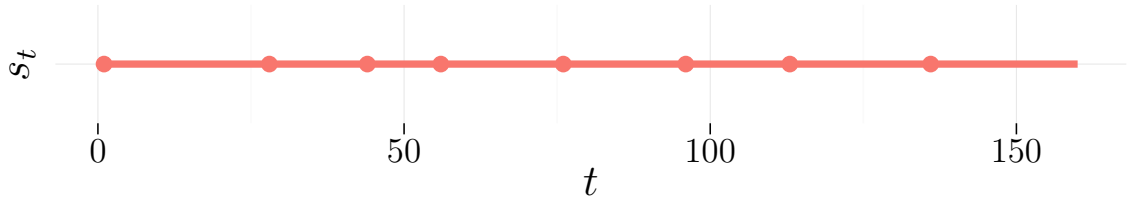
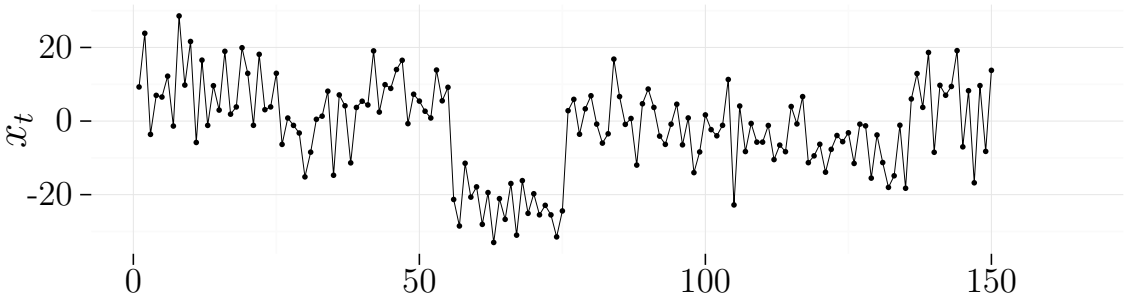
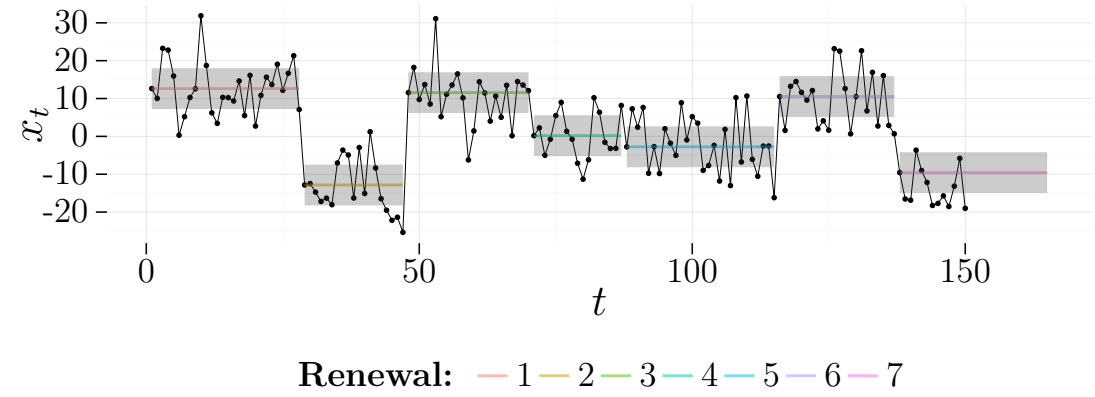


Figure 3.4.: A realization of a renewal process with $\lambda = 20$ as in (3.9) and $\eta = \mathcal{P}ois(\lambda)$.

This realization is used to generate the observation sequence Fig. 3.5a, with ε as in (3.8) and parameters as in (3.9). The bottom and top of the rectangles in Fig. 3.5b are the .25 and .75 quantiles.



(a) A simulated observation sequence from a jump sub-model with underlying renewal sequence Fig. 3.4 and observation model (3.8) with parameters (3.9).



(b) The same observation sequence as in (a). Each colored line is a local mean, and the surrounding rectangle represents the .25 and .75 quantiles for all but the first observation in each observation subsequence (conditioned on the first observation in each observation subsequence). The height of this rectangle is determined by σ_{local}. Note there is different local mean for each renewal.

Figure 3.5.: An observation sequence realized from the jump sub-model (a) without overlay (b) with overlay.

3.2.5. Inference

We again used 1000 sets of random starting parameters to fit our model. The starting parameters for λ were drawn from a uniform distribution on $[0, T]$. The

starting parameters for μ_{global} were drawn from a normal distribution with mean the same as the sample mean of $x_{1:T}$ and variance equal to the sample variance of $\{\min x_{1:T}, \max x_{1:T}\}$. The starting parameters for σ_{global} were drawn from a uniform distribution on $[0, sd(\{\min x_{1:T}, \max x_{1:T}\})]$, where $sd(y_{1:T})$ is the sample standard deviation of $y_{1:T}$. The starting parameters for σ_{local} were set to $sd(x_{1:T})$.

3.2.5.1. Parameters

The MLE parameters obtained from the EM algorithm are similar to the true values (3.9):

$$\lambda = 18.12, \quad \mu_g = -1.39, \quad \sigma_g = 9.34, \quad \sigma_l = 7.91 \tag{3.10}$$

3.2.5.2. Renewal Sequence

The result of the Viterbi algorithm using the MLE parameters (3.10) is shown in Fig. 3.6 on the current page.

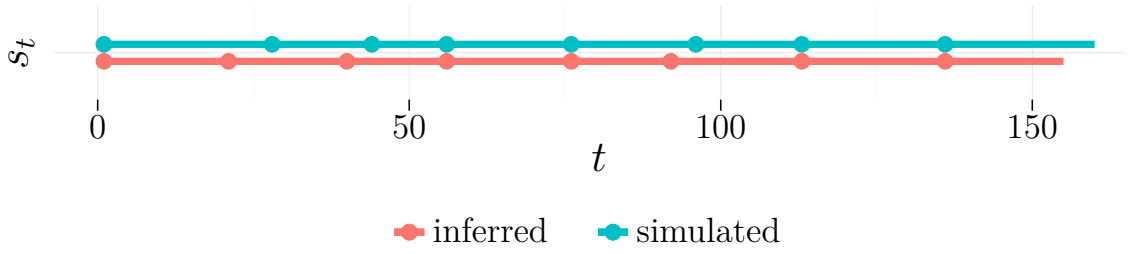


Figure 3.6.: Simulated and inferred renewal sequences for the jump sub-model (3.8).

3.2.6. Discussion

We have presented a jump sub-model, demonstrating that inference for a class of models we call *hidden renewal models* is readily performed using the framework developed in Chapter 2. This model illustrates an example of associating an event other than a state change with a renewal, something that is typically not thought of in the context of a HSMM.

3.3. A Stochastic Volatility Sub-model

We can use the HMRM framework to develop a stochastic volatility (SV) sub-model. Whereas our jump sub-model associated a mean with each observation subsequence, in this model, we instead associate an unobserved variance with each subsequence.

We do this by augmenting the hidden MRP (2.2) with a (inverse) variance sequence $v_{1:R}$. These variances are drawn from a distribution determined by the adjacent superstates, $z_{r:r+1}$. Each random variance is conditionally independent of all random variables outside its sojourn:

$$v_r | z_{r:r+1} \quad \sim \quad \mathcal{G}a(\nu_{z_r,z_{r+1}}/2, \nu_{z_r,z_{r+1}}/2) \tag{3.11}$$

$$v_r \quad \perp\!\!\!\perp \quad x_{\setminus t_r:t_{r+1}-1}, t_{\setminus r:r+1}, z_{\setminus r:r+1}, v_{\setminus r} \quad | \quad z_{r:r+1}, t_{r:r+1} \tag{3.12}$$

When

$$(z_{1:R+1}, t_{1:R+1}) \quad \sim \quad \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, T)$$

and

$$v_r|z_{r:r+1} \quad \sim \quad \mathcal{G}a(\nu_{z_r,z_{r+1}}/2, \nu_{z_r,z_{r+1}}/2)$$

for each $r = 1, \ldots, R$, we say that the sequence $(z_{1:R+1}, t_{1:R+1}, v_{1:R})$ is drawn from an *augmented* Markov renewal process, and write:

$$(z_{1:R+1}, t_{1:R+1}, v_{1:R}) \quad \sim \quad \mathcal{AMRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, \nu_{1:K,1:K}, T)$$

Fig. 3.8a shows a simulation from this model. In Fig. 3.8b, the height of each rectangle is inversely proportional to a value drawn from a $\mathcal{G}a(\nu/2, \nu/2)$ distribution. The bottom and top of these rectangles represent the .1 and .9 quantiles of observations drawn during the associated sojourn.

Liu and Rubin [21] show how the EM algorithm can be employed to find the MLE for a Student's t-distribution. A key difference with their analysis and ours is that they associate every unobserved variance with a single observation, whereas in our model, an unobserved variance is associated with a subsequence of the observations.

3.3.1. The Observation Model

Given $(z_{r:r+1}, t_{r:r+1}, v_r)$, which we call the the r^{th} *augmented* sojourn, the observations are distributed as iid normal, with variance inversely proportional to v_r:

$$\begin{aligned} &p(x_{t_r:(t_{r+1}-1)\wedge T}|z_{r:r+1}, t_{r:r+1}, v_r) \\ &\qquad = \prod_{t=t_r}^{(t_{r+1}-1)\wedge T} f_{\mathcal{N}}(x_t; \mu_{z_r,z_{r+1}}, \sigma^2_{z_r,z_{r+1}}/v_r) \qquad \text{for } [r=1,\ldots,R] \end{aligned} \tag{3.13}$$

Using (3.11) and (3.13), the distribution of $x_{t_r:t_{r+1}-1}, v_r | z_{r:r+1}, t_{r:r+1}$ can be expressed

$$\begin{aligned} & p(x_{t_r:t_{r+1}-1}, v_r | z_{r:r+1}, t_{r:r+1}) \\ & \overset{(CRP)}{=} p(x_{t_r:t_{r+1}-1} | z_{r:r+1}, t_{r:r+1}) p(v_r | z_{r:r+1}, t_{r:r+1}) \qquad (3.14) \\ & \overset{(3.13)}{\underset{(3.11)}{=}} \prod_{t=t_r}^{(t_{r+1}-1)\wedge T} f_{\mathcal{N}}(x_t; \mu_{z_r,z_{r+1}}, \sigma^2_{z_r,z_{r+1}}/v_r) \\ & \qquad \times f_{\mathcal{G}a}(v_r; \nu_{z_r,z_{r+1}}/2, \nu_{z_r,z_{r+1}}/2) \qquad (3.15) \end{aligned}$$

Then v_r can be integrated out to yield the emission distributions:

$$\begin{aligned} f_{\varepsilon_{j,k}}(x_{t:t+d-1}) & \triangleq p(x_{t:t+d-1} | z_r = j, z_{r+1} = k, t_r = t, t_{r+1} = t+d) \\ & = \int_0^\infty p(x_{t:t+d-1}, v_r | z_r = j, z_{r+1} = k, t_r = t, t_{r+1} = t+d)\, dv_r \\ & \overset{(3.15)}{=} (\pi \nu_{j,k} \sigma^2_{j,k})^{-d/2} \frac{\Gamma\left(\frac{d+\nu_{j,k}}{2}\right)}{\Gamma\left(\frac{\nu_{j,k}}{2}\right)} \left(\frac{\sum_{\delta=0}^{d-1} (x_{t+\delta} - \mu_{j,k})^2}{\nu_{j,k}\sigma^2_{j,k}} + 1 \right)^{-\frac{d+\nu_{j,k}}{2}} \end{aligned}$$

Whence the emission probabilities needed by the forward/backward probabilties can be computed.

3.3.2. Optimizing Q_ε

Recall the CDLL given in (2.12):

$$\begin{aligned} \log p(x_{1:T}, z_{1:R+1}, t_{1:R+1}) & = \log \iota_{z_1} + \sum_{r=1}^{R} \left[\log \tau_{z_r,z_{r+1}} + \log f_{\eta_{z_r,z_{r+1}}}(d_r) \right] \\ & \quad + \sum_{r=1}^{R} \log f_{\varepsilon^{d_r}_{z_r,z_{r+1}}}(x_{t_r:t_{r+1}-1\wedge T}) \end{aligned}$$

In the SV sub-model we must account for the augmentation of the variance sequence $v_{1:R}$. Our SV sub-model CDLL is:

$$\begin{aligned}
&\log p(x_{1:T}, z_{1:R+1}, t_{1:R+1}, v_{1:R}) \\
&\quad = \quad \iota_{z_1} + \sum_{r=1}^{R} \left[\log \tau_{z_r,z_{r+1}} + \log f_{\eta_{z_r,z_{r+1}}}(d_r)\right] \\
&\qquad\qquad + \sum_{r=1}^{R} \log f_{\mathcal{N}}(x_{t_r:t_{r+1}-1\wedge T}; \mu_{z_r,z_{r+1}}, \sigma^2_{z_r,z_{r+1}}/v_r) \\
&\qquad\qquad + \sum_{r=1}^{R} \log f_{\mathcal{G}a}(v_r; \nu_{z_r,z_{r+1}}/2, \nu_{z_r,z_{r+1}}/2)
\end{aligned}$$

The "expectation" step in the EM algorithm is:

$$E_{\substack{z_{1:R+1} \\ t_{1:R+1} \\ v_{1:R}} \Big| x_{1:T}} \left[\log p(x_{1:T}, z_{1:R+1}, t_{1:R+1}, v_{1:R})\right]$$

Because they do not have any v_r terms, Q_ι, Q_τ, Q_η are all unchanged from (2.12). The Q_ε term becomes:

$$\begin{aligned}
&Q_\varepsilon(\varepsilon_{1:K,1:K}; \theta^{(n)}) \\
&\quad = \quad E_{\substack{z_{1:R+1} \\ t_{1:R+1} \\ v_{1:R}} \Big| x_{1:T}} \left[\sum_{r=1}^{R} \begin{matrix} \log f_{\mathcal{N}}(x_{t_r:t_{r+1}-1\wedge T}; \mu_{z_r,z_{r+1}}, \sigma^2_{z_r,z_{r+1}}/v_r) \\ + \log f_{\mathcal{G}a}(v_r; \nu_{z_r,z_{r+1}}/2, \nu_{z_r,z_{r+1}}/2) \end{matrix}\right] \\
&\quad \overset{(A.31)}{=} \quad \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} E_{v_r \Big| \substack{x_{t:t+d-1\wedge T} \\ z_r=j, z_{r+1}=k \\ t_r=t, t_{r+1}=t+d}} \begin{bmatrix} \log f_{\varepsilon^d_{j,k}}(x_{t:t+d-1\wedge T}) \\ + \log f_{\mathcal{G}a}(v; \nu_{j,k}/2, \nu_{j,k}/2) \end{bmatrix} S^{t,d}_{j,k}
\end{aligned} \tag{3.16}$$

Where the last expectation is taken under the posterior distribution of $v \sim \mathcal{G}a(\nu_{j,k}/2, \nu_{j,k}/2)$ after having observed $x_{t:(t+d-1)\wedge T}|v \sim$

$\prod_{\delta=0}^{(d-1)\wedge(T-t)} f_{\mathcal{N}}(x_{t+\delta};\mu_{j,k},\sigma_{j,k}^2/v)$. This posterior distribution is [3]:

$$v|x_{t:t+d-1\wedge T} \sim \mathcal{G}a\left(\frac{\nu_{j,k}+(d\wedge T-t+1)}{2},\frac{\nu_{j,k}}{2}+\frac{\sum_{\delta=0}^{d-1\wedge T-t}(x_{t+\delta}-\mu_{j,k})^2}{2\sigma_{j,k}^2}\right) \quad (3.17)$$

We abbreviate this expectation as $E_{j,k}^{t,d}$ for the remainder of this subsection. Each summand in (3.16) is:

$$\begin{aligned} E_{j,k}^{t,d}\left[\log f_{\varepsilon_{j,k}^d}(x_{t:(t+d-1\wedge T)}) \quad + \quad \log f_{\mathcal{G}a}(v;\nu_{j,k}/2,\nu_{j,k}/2)\right] \\ = \quad -\frac{d\wedge(T-t+1)}{2}\left(\log 2\pi - \log\sigma_{j,k}^2 + E_{j,k}^{t,d}[\log v]\right) \\ -\frac{E_{j,k}^{t,d}[v]}{2\sigma_{j,k}^2}\sum_{\delta=0}^{d-1\wedge T-t}(x_{t+\delta}-\mu_{j,k})^2 \\ +\frac{\nu_{j,k}}{2}\log\frac{\nu_{j,k}}{2}-\log\Gamma\left(\frac{\nu_{j,k}}{2}\right)+\left(\frac{\nu_{j,k}}{2}-1\right)E_{j,k}^{t,d}[\log v]-\frac{\nu_{j,k}}{2}E_{j,k}^{t,d}[v] \end{aligned}$$

The expectations of a $g\sim\mathcal{G}a(\alpha,\beta)$ random variable are $E[g]=\alpha/\beta$ and $E[\log g]=\psi(\alpha)-\log\beta$.[2] Thus

$$\begin{aligned} E_{j,k}^{t,d}[v] &\overset{(3.17)}{=} \frac{\nu_{j,k}+(d\wedge[T-t+1])}{\nu_{j,k}+\sum_{\delta=0}^{d-1\wedge T-t}(x_{t+\delta}-\mu_{j,k})^2/\sigma_{j,k}^2} \\ E_{j,k}^{t,d}[\log v] &\overset{(3.17)}{=} \psi\left(\frac{\nu_{j,k}+(d\wedge[T-t+1])}{2}\right) \\ &\quad -\log\left(\frac{\nu_{j,k}+\sum_{\delta=0}^{(d-1)\wedge(T-t)}(x_{t+\delta}-\mu_{j,k})^2/\sigma_{j,k}^2}{2}\right) \end{aligned}$$

[2]The function ψ is defined as $\psi(x)\triangleq\frac{d}{dx}[\log\Gamma(x)]$, and is called the *digamma* function.

First order optimality conditions yield the following update formulas:

$$\hat{\mu}_{j,k} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D} E_{j,k}^{t,d}[v]\, S_{j,k}^{t,d} \sum_{\delta=0}^{(d-1)\wedge(T-t)} x_{t+\delta}}{\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{\delta=0}^{d-1\wedge T-t} E_{j,k}^{t,d}[v]\, S_{j,k}^{t,d}\ (d \wedge T-t+1)} \tag{3.18a}$$

$$\hat{\sigma}_{j,k}^{2} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{D} E_{j,k}^{t,d}[v]\, S_{j,k}^{t,d} \sum_{\delta=0}^{(d-1)\wedge(T-t)} \left(x_{t+\delta}-\mu_{j,k}\right)^{2}}{\sum_{t=1}^{T}\sum_{d=1}^{D} S_{j,k}^{t,d}\,(d \wedge [T-t+1])} \tag{3.18b}$$

$$0 = \sum_{t=1}^{T}\sum_{d=1}^{D}\left[\log\frac{\hat{\nu}_{j,k}}{2} + 1 - \psi\left(\frac{\hat{\nu}_{j,k}}{2}\right) + E_{j,k}^{t,d}[\log v] - E_{j,k}^{t,d}[v]\right] S_{j,k}^{t,d} \tag{3.18c}$$

The expression on the RHS of (3.18c) is monotonic in $\nu_{j,k}$ on $(0,\infty)$, and so $\hat{\nu}_{j,k}$ can be found via a one-dimensional search [21], e.g. the bisection method.

3.3.3. Viterbi for the Augmented Renewal Sequence

We can rederive the Viterbi algorithm with $v_{1:R}$ as a part of the renewal sequence. We define $\mathscr{M}_k^t$ to be the maximum a posteriori probability of all partial augmented renewal sequences ending with a renewal in superstate k at time t:

$$\mathscr{M}_k^t \quad \triangleq \quad \max_{\substack{r=1,\ldots,t \\ z_{1:r-1},t_{1:r-1} \\ v_{1:r-1}}} p\left(\begin{matrix} z_r=k & z_{1:r-1} \\ t_r=t & , t_{1:r-1} \\ v_{r-1} & v_{1:r-2}\end{matrix}\middle| x_{1:T}\right) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T+D \\ k=1,\ldots,K\end{bmatrix} \tag{3.19}$$

Our record of best predecessors is now $\mathscr{Z}_k^t$, $\mathscr{T}_k^t$, $\mathscr{V}_k^t$, the values of $z_{r-1}, t_{r-1}, v_{r-1}$ in (3.19). Equation (3.20) shows how all these values can be computed using the previously computed values $\mathscr{M}_{1:K}^{t-1}$.

$$\mathscr{M}_k^t \quad \triangleq \quad \max_{\substack{r=1,\ldots,t \\ z_{1:r-1},t_{1:r-1} \\ v_{1:r-1}}} p\left(\begin{matrix} z_r=k & z_{1:r-1} \\ t_r=t & , t_{1:r-1} \\ v_{r-1} & v_{1:r-2}\end{matrix}\middle| x_{1:T}\right)$$

$$
\begin{aligned}
&= \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D \\ v_{r-1}}} \max_{\substack{r=1,\ldots,t-d \\ z_{1:r-2},t_{1:r-2} \\ v_{1:r-2}}} p\left(\begin{matrix} z_r=k & z_{r-1}=j & z_{1:r-1} \\ t_r=t & t_{r-1}=t-d, & t_{1:r-1} \\ v_{r-1} & v_{r-2} & v_{1:r-2} \end{matrix}\middle| x_{1:T}\right) \\
&\overset{(CRP)}{=} \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D \\ v_{r-1}}} p\left(v_{r-1}\middle|\begin{matrix} z_r=k & z_{r-1}=j \\ t_r=t & t_{r-1}=t-d, \\ & v_{r-2} \end{matrix}, x_{1:T}\right) \\
&\quad \times p\left(\begin{matrix} z_r=k \\ t_r=t \end{matrix}\middle|\begin{matrix} z_{r-1}=j \\ t_{r-1}=t-d, \\ v_{r-2} \end{matrix}, x_{1:T}\right) \\
&\quad \times \max_{\substack{r=1,\ldots,t-d \\ z_{1:r-2},t_{1:r-2} \\ v_{1:r-2}}} p\left(\begin{matrix} z_{r-1}=j & z_{1:r-1} \\ t_{r-1}=t-d, & t_{1:r-1} \\ v_{r-2} & v_{1:r-2} \end{matrix}\middle| x_{1:T}\right) \\
&\overset{\substack{(3.12)\\(2.3b)}}{=} \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D \\ v_{r-1}}} p\left(v_{r-1}\middle|\begin{matrix} z_r=k & z_{r-1}=j \\ t_r=t & t_{r-1}=t-d, \end{matrix}, x_{t-d:(T\wedge t-1)}\right) \\
&\quad \times p\left(\begin{matrix} z_r=k \\ t_r=t \end{matrix}\middle|\begin{matrix} z_{r-1}=j \\ t_{r-1}=t-d, \end{matrix}, x_{t-d:T}\right) \\
&\quad \times \max_{\substack{r=1,\ldots,t-d \\ z_{1:r-2},t_{1:r-2} \\ v_{1:r-2}}} p\left(\begin{matrix} z_{r-1}=j & z_{1:r-1} \\ t_{r-1}=t-d, & t_{1:r-1} \\ v_{r-2} & v_{1:r-2} \end{matrix}\middle| x_{1:T}\right) \\
&= \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,(t-1)\wedge D}} \max_{v_{r-1}} p\left(v_{r-1}\middle|\begin{matrix} z_r=k & z_{r-1}=j \\ t_r=t & t_{r-1}=t-d, \end{matrix}, x_{t-d:(T\wedge t-1)}\right) \\
&\qquad\qquad \times \frac{\beta_{j,k}^{t-d,d}}{B_j^{t-d}} \mathscr{M}_j^{t-d} \qquad (3.20) \\
(\mathscr{Z}_k^t,\ \mathscr{T}_k^t, \mathscr{V}_k^t) &\triangleq \operatorname*{arg\,max}_{\substack{j=1,\ldots,K \\ t'=1\vee t-D,\ldots,t-1\wedge T \\ v_{r-1}}} p\left(v_{r-1}\middle|\begin{matrix} z_r=k & z_{r-1}=j \\ t_r=t & t_{r-1}=t-d, \end{matrix}, x_{t-d:(T\wedge t-1)}\right) \\
&\qquad\qquad \times \frac{\beta_{j,k}^{t',t-t'}}{B_j^{t'}} \mathscr{M}_j^{t'}
\end{aligned}
$$

The mode of a $\mathcal{G}a(\alpha,\beta)$ distribution, when it exists, is $(\alpha-1)/\beta$, so for each j,k:

$$
\max_{v_{r-1}} p\left(v_{r-1}\middle|\begin{matrix} z_r=k & z_{r-1}=j \\ t_r=t & t_{r-1}=t-d, \end{matrix}, x_{t-d:t-1}\right) = \frac{\nu_{j,k}+d-2}{\nu_{j,k}+\sum_{\delta=1}^{d}\left(x_{t-\delta}-\mu_{j,k}\right)^2/\sigma_{j,k}^2} \qquad (3.21)
$$

The mode (3.21) only exists when $\nu_{j,k} + d - 2 > 0$, TODO d_r so the MAP augmented sequence only exists when $\nu_{j,k} \geq 1$ for each j, k. When this condition is not satisfied, a possible practical solution could be to compute the MAP of the *un*augmented sequence $(\hat{z}_{1:R+1}, \hat{t}_{1:R+1})$ as in (2.17). Then, the sequence $\hat{v}_{1:R}$ could be constructed as follows: For each r such that $\hat{d}_r \geq 2$, set $\hat{v}_r$ as in (3.21). For those r such that $\hat{d}_r = 1$ and $\nu_{z_r, z_{r+1}} < 1$, the posterior density of v_r is unbounded as $v_r \to 0$. So for those r, $\hat{v}_r$ could be set to some adhoc value.

When it exists, the MAP is recovered by the following algorithm:

$$\mathscr{M}^1_{1:K} \quad \leftarrow \quad \iota_{1:K}$$

Then for each $t = 2, \ldots, T + D$

$$\mathscr{M}^t_k \quad \leftarrow \quad \max_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,t-1\wedge D}} \frac{\nu_{j,k} + d - 2}{\nu_{j,k} + \sum_{\delta=1}^{d} \left(x_{t-\delta} - \mu_{j,k}\right)^2 / \sigma^2_{j,k}} \times \frac{\beta^{t-d,d}_{j,k}}{B^{t-d}_j} \mathscr{M}^{t-d}_j \qquad \text{for } [k=1,\ldots,K]$$

$$(\mathscr{Z}^t_k,\, d) \quad \leftarrow \quad \operatorname*{arg\,max}_{\substack{j=1,\ldots,K \\ d=1\vee t,\ldots,t-1\wedge D}} \frac{\nu_{j,k} + d - 2}{\nu_{j,k} + \sum_{\delta=1}^{d} \left(x_{t-\delta} - \mu_{j,k}\right)^2 / \sigma^2_{j,k}} \times \frac{\beta^{t-d,d}_{j,k}}{B^{t-d}_j} \mathscr{M}^{t-d}_j \qquad \text{for } [k=1,\ldots,K]$$

$$\mathscr{T}^t_k \quad \leftarrow \quad t - d_k \qquad \text{for } [k=1,\ldots,K]$$

$$\mathscr{V}_k^t \quad \leftarrow \quad \frac{\nu_{\mathscr{Z}_k^t,k} + d - 2}{\nu_{\mathscr{Z}_k^t,k} + \sum_{\delta=1}^{d} \left(x_{t-\delta} - \mu_{\mathscr{Z}_k^t,k}\right)^2 / \sigma^2_{\mathscr{Z}_k^t,k}}$$

Once $\mathscr{M}_{1:K}^{T+D}$ are computed, the MAP state sequence $\hat{z}_{1:R+1}, \hat{t}_{1:R+1}$ is constructed in reverse order, starting with the last renewal which must occur after time T:

$$(\hat{z}_{R+1},\, \hat{t}_{R+1}) \quad \leftarrow \quad \underset{\substack{k=1,\ldots,K \\ t=T+1,\ldots,T+D}}{\arg\max} \quad \mathscr{M}_k^t$$

And then for each r:

$$(\hat{z}_r,\, \hat{t}_r,\, \hat{v}_r) \quad \leftarrow \quad (\mathscr{Z}_{\hat{z}_{r+1}}^{\hat{t}_{r+1}},\, \mathscr{T}_{\hat{z}_{r+1}}^{\hat{t}_{r+1}},\, \mathscr{V}_{\hat{z}_{r+1}}^{\hat{t}_{r+1}})$$

The algorithm stops when $\hat{t}_r = 1$.

3.3.4. Simulated Data

We first draw a renewal sequence from a renewal process with Poisson holding-time distribution with parameter $\lambda = 20$. This sequence is shown in Fig. 3.7 on this page.

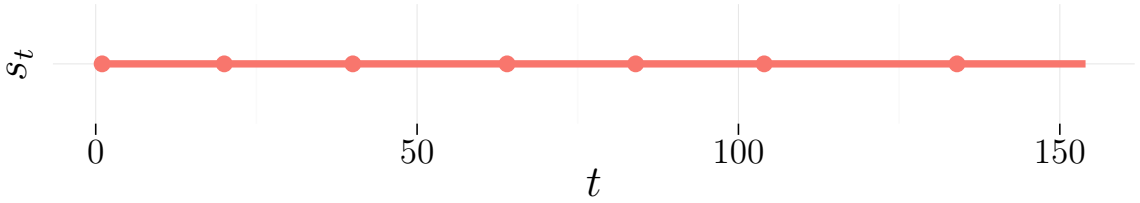


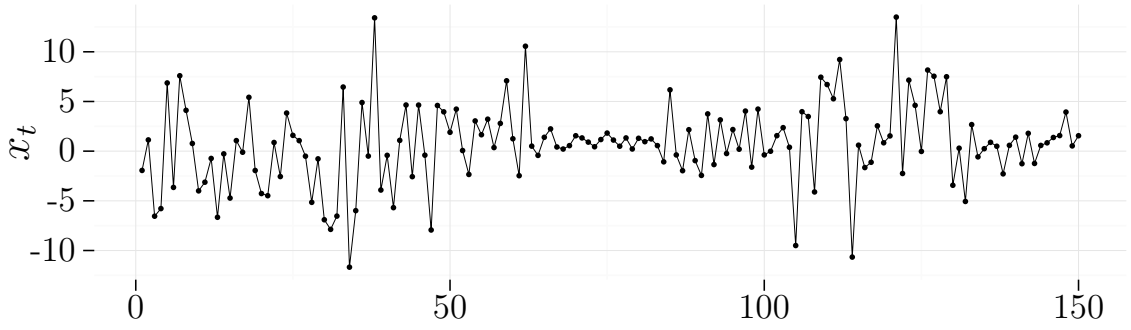
Figure 3.7.: A renewal process with Poisson holding-time distribution with parameter $\lambda = 20$.

There are 7 renewals in this sequence. The corresponding variances, whose inverses

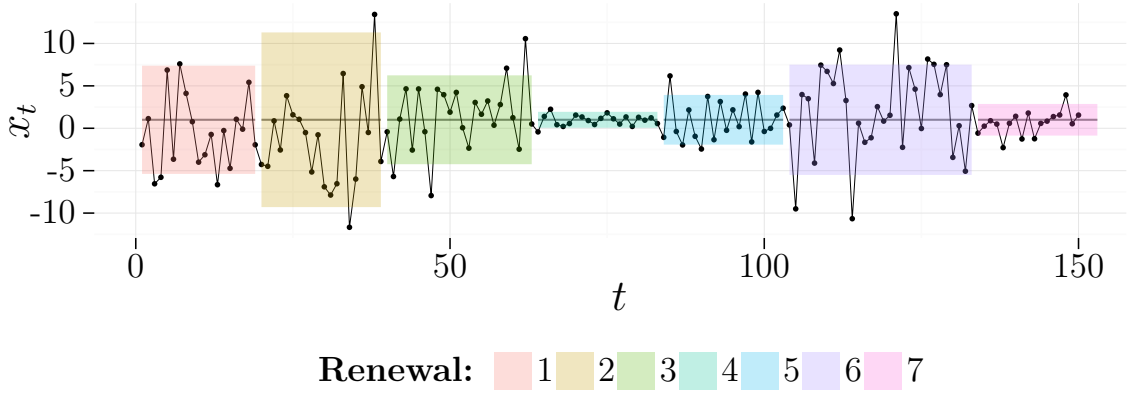
are drawn iid from a $\mathcal{G}a(1.5/2, 1.5/2)$, are:

$$1/v_{1:7} \quad = \quad (6.19, 16.14, 4.18, 0.13, 1.31, 6.43, 0.53)$$

Each observation is drawn from a $\mathcal{N}(1, 2^2/v_r)$ distribution, where v_r is used for the r^{th} observation subsequence. The observation sequence is shown in Fig. 3.8a. The bottom and top of the overlay in Fig. 3.8b are the .1 and .9 quantiles for each observation given the augmented renewal sequence.



(a) Observation sequence simulated from our stochastic volatility sub-model with $K = 1$, $\eta = \mathcal{P}ois(20)$, $\mu = 1$, $\sigma = 2$, and $\nu = 1.5$.



(b) The .1 and .9 quantiles of each observation in (a) overlayed.

Figure 3.8.: (a) An observation sequence realized from our stochastic volatility sub-model. (b) With augmented sojourns overlayed.

3.3.5. Inference

We used 1000 sets of random starting parameters. The starting parameters for λ were drawn from a uniform distribution on $[0, T]$. The starting parameters for μ were drawn from a normal distribution with mean the same as the sample mean of $x_{1:T}$ and variance equal to the sample variance of $\{\min x_{1:T}, \max x_{1:T}\}$. The starting parame-

ters for σ were drawn from a uniform distribution on $[0, sd(\{\min x_{1:T}, \max x_{1:T}\})]$.

3.3.5.1. Parameters

The MLE parameters are:

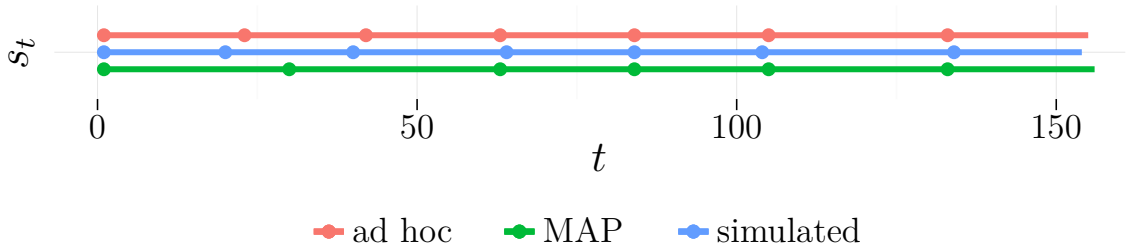
$$\hat{\lambda} = 21.19, \qquad \hat{\mu} = 0.88, \qquad \hat{\sigma} = 1.52, \qquad \hat{\nu} = 1.03 \tag{3.22}$$

3.3.5.2. Renewal Sequence

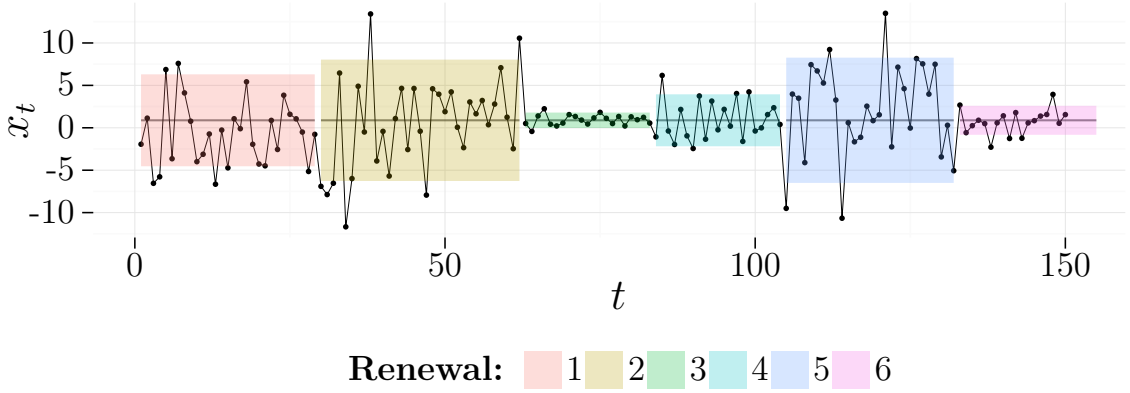
The results of the Viterbi algorithm, using both the MAP and ad hoc methods described in Section 3.3.3, are shown in Fig. 3.9a on the next page. The MAP augmented renewal sequence is overlayed in Fig. 3.9b.

3.3.6. Discussion

We have shown how the statespace of an MRP can be augmented. In this case, we augmented each superstate with a variance drawn from a $\mathcal{G}a(\nu/2, \nu/2)$ distribution. The sub-model presented provides another example of associating an event other than a state change with a renewal. In this case, this event is a change in volatility.



(a) Simulated and inferred renewal sequences for the stochastic volatility sub-model. Both the MAP and ad hoc methods described in Section 3.3.3 are shown.



(b) The MAP renewal and variance sequence for the data in Fig. 3.8a. The .1 and .9 quantiles, determined by $\hat{\mu} = 0.88$, $\hat{\sigma} = 1.52$, are overlayed.

Figure 3.9.: Inference on the observations in Fig. 3.8a

4. Applications

This chapter contains two instances of hidden Markov renewal sub-models applied to real datasets. With each application, we compare the HMRM's performance to that of previously published models.

4.1. Modeling Autocorrelations of Squared Returns

Rydén et al. [32] found that hidden Markov models were able to reproduce several stylized facts of financial time series. A notable exception was the slowly decaying ACF of squared returns. Bulla and Bulla [6] developed a model, based on Guédon's HSMM, with negative binomial holding-time distributions and normal emission distributions. We denote this model as "Bu" in the figures and tables of this section. They showed that this model was able to reproduce this stylized fact in most cases. They noted however that the model was not able to reproduce the slowly decaying ACF of four sector indices considered. Of these four sectors, three were from the financial industry.

We fit our bridging-means model of Section 3.1, denoted "BM", to these four sectors[1]. Our model contains a more elaborate renewal independence structure, allowing for holding-time and emission dependence on both the coinciding and next superstates. So we investigate the possibility that the "bridging" of means in our model may not be the source of any improved performance. We do this by also fitting another HMRM, but with emissions distributions that assume conditional independence of observations within a subsequence. That is, for each pair of superstates j, k the emission density of a sojourn of length d is:

$$f_{\varepsilon_{j,k}^{d}}\left(x_{t:(t+d-1)\wedge T}\right) \quad = \quad \prod_{\delta=0}^{(d-1)\wedge(T-t)} f_{\mathcal{N}}\left(x_{t+\delta}; \mu_j, \sigma_{j,k}^2\right) \tag{4.1}$$

We denote this model "IID". Like the bridging-means model, this model has 17 free parameters. But it does not transition gradually between superstates. Both the IID and BM models used negative binomial holding-time distributions.

Fig. 4.1 shows the empirical ACFs of these datasets with the ACFs of the fitted models. It is clear from visual inspection that both the BM and IID model's perform significantly better than Bulla's model. Tab. 4.1, which contains the sums of squared differences between each model's ACF and the empirical ACF, confirms this assertion. The difference between our BM and IID model is negligible. With the exception of the parameters fit to the Financials dataset (see Section 4.1.1.2), the difference in the parameters for these models is also relatively small.

[1]We thank Professor Jan Bulla for graciously providing the data used in this section.

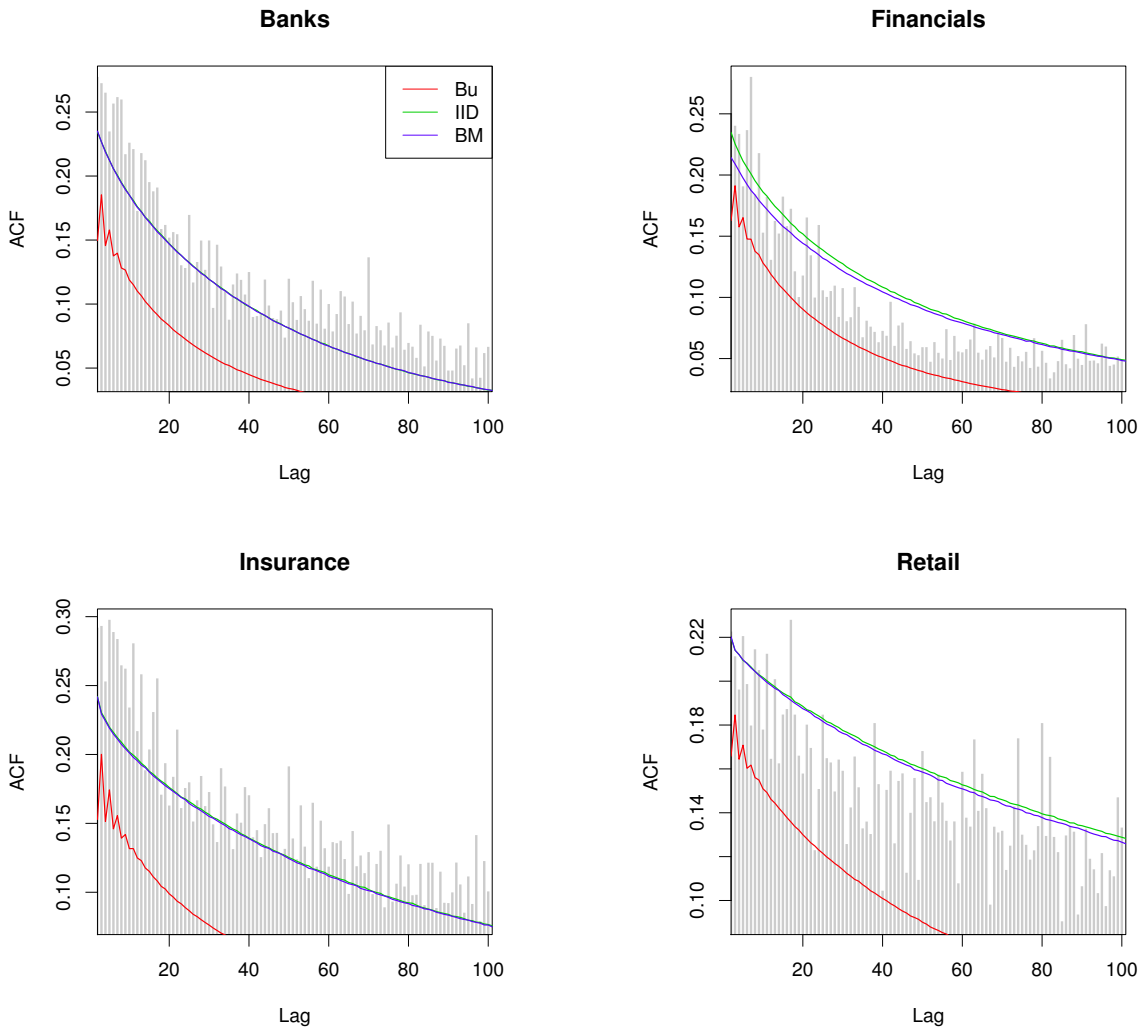


Figure 4.1.: ACFs of squared log returns for 4 sector indices. "Bu" is Bulla's model, "IID" is our HMRM with emissions as in (4.1), and "BM" is the bridging-means model of Section 3.1.

	Bu	IID	BM
Banks	0.53	0.09	0.09
Financials	0.17	0.07	0.07
Insurance	0.92	0.10	0.11
Retail	0.32	0.06	0.06

Table 4.1.: The sum of squared differences between the empirical ACF and those of the models in Fig. 4.1 at lags $1, \ldots, 100$. Like Bulla and Bulla [6], the ACFs were generated from simulation. We used an observation sequence of length 5×10^7.

4.1.1. MLE Parameters

4.1.1.1. Banks Dataset

Bridging-Means

$$\hat{\mu} = \begin{bmatrix} -0.94 \\ 8.15 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.77 & 1.22 \\ 0.67 & 0.49 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.13 & 0.12 \\ 0.05 & 10.74 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.01 & 0.03 \\ 0.01 & 0.11 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.13 & 0.87 \\ 0.97 & 0.03 \end{bmatrix}$$

IID HMRM

$$\hat{\mu} = \begin{bmatrix} 2.57 \\ 5.59 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.78 & 1.22 \\ 0.68 & 0.50 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.12 & 0.11 \\ 0.04 & 10.28 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.01 & 0.03 \\ 0.01 & 0.11 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.13 & 0.87 \\ 0.96 & 0.04 \end{bmatrix}$$

Bulla

$$\hat{\mu} = \begin{bmatrix} -11.02 \\ 6.25 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.16 \\ 0.68 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.08 & 0.08 \\ 0.08 & 0.08 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.03 & 0.03 \\ 0.01 & 0.01 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix}$$

4.1.1.2. Financials Dataset

Bridging-Means

$$\hat{\mu} = \begin{bmatrix} -15.40 \\ 15.73 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 1.14 & 2.49 \\ 0.54 & 0.73 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.05 & 0.04 \\ 0.08 & 0.18 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.05 & 0.01 \\ 0.00 & 0.02 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.83 & 0.17 \\ 0.42 & 0.58 \end{bmatrix}$$

IID HMRM

$$\hat{\mu} = \begin{bmatrix} -10.01 \\ 6.43 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.79 & 1.49 \\ 0.86 & 0.59 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.09 & 0.05 \\ 0.19 & 0.09 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.01 & 0.03 \\ 0.04 & 0.00 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.16 & 0.84 \\ 0.72 & 0.28 \end{bmatrix}$$

Bulla

$$\hat{\mu} = \begin{bmatrix} -14.14 \\ 6.31 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.16 \\ 0.67 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.09 & 0.09 \\ 0.09 & 0.08 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.03 & 0.03 \\ 0.01 & 0.01 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix}$$

4.1.1.3. Insurance Dataset

Bridging-Means

$$\hat{\mu} = \begin{bmatrix} -1.81 \\ 6.03 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 3.65 & 1.62 \\ 0.94 & 0.65 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.03 & 0.10 \\ 0.20 & 9.87 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.00 & 0.01 \\ 0.01 & 0.09 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.39 & 0.61 \\ 0.85 & 0.15 \end{bmatrix}$$

IID HMRM

$$\hat{\mu} = \begin{bmatrix} 2.05 \\ 4.67 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 3.66 & 1.61 \\ 0.93 & 0.65 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.03 & 0.10 \\ 0.19 & 13.22 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.00 & 0.01 \\ 0.01 & 0.12 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.38 & 0.62 \\ 0.86 & 0.14 \end{bmatrix}$$

Bulla

$$\hat{\mu} = \begin{bmatrix} -12.33 \\ 4.70 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.76 \\ 0.84 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.06 & 0.06 \\ 0.06 & 0.06 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.02 & 0.02 \\ 0.00 & 0.01 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix}$$

4.1.1.4. Retail Dataset

Bridging-Means

$$\hat{\mu} = \begin{bmatrix} 1.70 \\ 8.23 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 3.39 & 1.73 \\ 1.00 & 0.71 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.02 & 0.06 \\ 0.18 & 3.45 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.00 & 0.01 \\ 0.02 & 0.05 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.30 & 0.70 \\ 0.88 & 0.12 \end{bmatrix}$$

IID HMRM

$$\hat{\mu} = \begin{bmatrix} 1.41 \\ 7.66 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 3.39 & 1.75 \\ 1.01 & 0.71 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.02 & 0.06 \\ 0.18 & 3.30 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.00 & 0.01 \\ 0.02 & 0.05 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.30 & 0.70 \\ 0.87 & 0.13 \end{bmatrix}$$

Bulla

$$\hat{\mu} = \begin{bmatrix} -13.16 \\ 9.60 \end{bmatrix} \times 10^{-4} \qquad \hat{\sigma} = \begin{bmatrix} 2.92 \\ 0.88 \end{bmatrix} \times 10^{-2}$$

$$\hat{r} = \begin{bmatrix} 0.03 & 0.03 \\ 0.12 & 0.12 \end{bmatrix} \qquad \hat{p} = \begin{bmatrix} 0.00 & 0.00 \\ 0.01 & 0.01 \end{bmatrix}$$

$$\hat{\tau} = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix}$$

4.2. Forecasting Volatility

Hamilton and Susmel [16] used autoregressive conditional changing heteroskedasticity (ARCH) models with regime-switching to forecast the volatility of stock returns. The data used was the value-weighted portfolio of stocks traded on the NYSE contained in the CRISP data tapes from the week ended July 3, 1962 to the week ended December 29, 1987. There are $T = 1331$ observations in this dataset.

We compare Hamilton and Susmel's model to an HMRM using this dataset[2]. The specific HMRM we use is a combination of the Jump sub-model from Section 3.2, and the stochastic volatility sub-model from Section 3.3. We also add a first order auto-regressive component to this HMRM's observation subsequence model. We call this model a AR(1)-SV HMRM.

[2]This data was obtained from Hamilton's website:
http://econweb.ucsd.edu/∼jhamilto/SWARCH.ZIP.

4.2.1. The AR(1)-SV HMRM

We describe our model and present the EM updates to find the MLE. We then present the result of the EM and Viterbi algorithms applied to weekly returns from the NYSE.

4.2.1.1. Model Description

We use Negative Binomial holding-time distributions. The distribution of the observation subsequences can be described:

$$
\begin{aligned}
& p\left(x_{t:(t+d-1\wedge T)} \middle| \substack{z_r=j,\ z_{r+1}=k \\ t_r=t,\ t_{r+1}=t+d}, v_r\right) \\
&= f_{\mathcal{N}}\overbrace{\left(x_t; \mu_{g(j,k)}, \sigma^2_{g(j,k)}\right)}^{\text{initial distribution}} \\
&\quad \times \prod_{\delta=1}^{(d-1)\wedge(T-t)} f_{\mathcal{N}}(\underbrace{x_{t+\delta}; \phi_{j,k}x_{t+\delta-1} + \mu_{l(j,k)}}_{\text{autoregression}}, \sigma^2_{l(j,k)}/v_r)
\end{aligned} \tag{4.2}
$$

where, as in our stochastic volatility sub-model, an (inverse) variance v_r is drawn for each observation subsequence:

$$
v_r \middle| \substack{z_r=j,\ z_{r+1}=k \\ t_r=t-d,\ t_{r+1}=t} \sim \mathcal{G}a\left(\nu_{j,k}/2, \nu_{j,k}/2\right) \tag{4.3}
$$

So the first observation in any subsequence is drawn from an initial distribution $\mathcal{N}\left(\mu_{g(j,k)}, \sigma^2_{g(j,k)}\right)$. Subsequent observations in the subsequence satisfy $x_{t+1} =$

$\mu_{l(j,k)} + \phi_{j,k} x_t + \epsilon_t$ where $\epsilon_t | v_r \sim \mathcal{N}(0, \sigma^2_{l(j,k)}/v_r)$. Integrating out v_r from (4.2) using (4.3) yields the emission densities that are readily employed in the forward-backward algorithm.

$$\begin{aligned} f_{\varepsilon_{j,k}}(x_{t:(t+d-1)\wedge T}) &= f_{\mathcal{N}}(x_t; \mu_{g(j,k)}, \sigma^2_{g(j,k)}) \times (\pi \nu_{j,k} \sigma^2_{l(j,k)})^{-\frac{d-1}{2}} \frac{\Gamma\left(\frac{d-1+\nu_{j,k}}{2}\right)}{\Gamma\left(\frac{\nu_{j,k}}{2}\right)} \\ &\quad \times \left(\frac{\sum_{\delta=1}^{d-1\wedge T-t} (x_{t+\delta} - \phi_{j,k} x_{t+\delta-1} - \mu_{l(j,k)})^2}{\nu_{j,k} \sigma^2_{l(j,k)}} + 1 \right)^{-\frac{d-1+\nu_{j,k}}{2}} \end{aligned}$$

4.2.1.2. EM Updates

The EM updates to the parameters $\mu_{g(j,k)}, \sigma_{g(j,k)}$, which describe the distribution of the initial observation in each subsequence, are the same as in the Jump sub-model, see Section 3.2.2 on page 76. We present these updates again here:

$$\begin{aligned} \hat{\mu}_{g(j,k)} &= \frac{\sum_{t=1}^{T} x_t \sum_{d=1}^{D} S^{t,d}_{j,k}}{\sum_{t=1}^{T} \sum_{d=1}^{D} S^{t,d}_{j,k}} \\ \hat{\sigma}_{g(j,k)} &= \frac{\sum_{t=1}^{T} \left(x_t - \mu_{g(j,k)}\right)^2 \sum_{d=1}^{D} S^{t,d}_{j,k}}{\sum_{t=1}^{T} \sum_{d=1}^{D} S^{t,d}_{j,k}} \end{aligned}$$

The updates of the local parameters are similar to the stochastic volatility model of Section 3.3. Because we assume that the first observation in each observation subsequence is drawn from the initial distribution, the sums over δ start at 1, rather than 0 as in (3.18b). The expected values of an (inverse) variance drawn during a sojourn from superstate j to superstate k starting at time t with duration d are:

$$E_{j,k}^{t,d}\left[v\right] \quad = \quad \frac{\nu_{j,k} + (d - 1 \wedge [T-t])}{\nu_{j,k} + \sum_{\delta=1}^{d-1\wedge T-t} \left(x_{t+\delta} - \phi_{j,k} x_{t+\delta-1} - \mu_{l(j,k)}\right)^2 / \sigma_{l(j,k)}^2} \quad \text{for } \left[{}_{d=2,\ldots,D}\right]$$

$$\begin{aligned} E_{j,k}^{t,d}\left[\log v\right] \quad &= \quad \psi\left(\frac{\nu_{j,k} + (d - 1 \wedge [T-t])}{2}\right) \\ &\quad - \log\left(\frac{\nu_{j,k} + \sum_{\delta=1}^{d-1\wedge T-t} \left(x_{t+\delta} - \phi_{j,k} x_{t+\delta-1} - \mu_{l(j,k)}\right)^2 / \sigma_{l(j,k)}^2}{2}\right) \end{aligned}$$

EM updates for the local parameters μ_l, ϕ_l satisfy the following matrix equation:

$$\begin{pmatrix} \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d}\left[d-1\wedge(T-t)\right] & \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} x_{t+\delta-1} \\ \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} x_{t+\delta-1} & \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} x_{t+\delta-1}^2 \end{pmatrix} \times \begin{pmatrix} \hat{\mu}_{l(j,k)} \\ \hat{\phi}_{j,k} \end{pmatrix} \quad = \quad \begin{pmatrix} \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} x_{t+\delta} \\ \sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} x_{t+\delta-1} x_{t+\delta} \end{pmatrix}$$

The update for $\hat{\sigma}_{l(j,k)}$ is:

$$\hat{\sigma}_{l(j,k)} = \frac{\sum_{t=1}^{T}\sum_{d=2}^{D} E_{j,k}^{t,d}\left[v\right] S_{j,k}^{t,d} \sum_{\delta=1}^{d-1\wedge(T-t)} \left(x_{t+\delta} - \phi_{j,k} x_{t+\delta-1} - \mu_{l(j,k)}\right)^2}{\sum_{t=1}^{T}\sum_{d=2}^{D} S_{j,k}^{t,d}\left(d - 1 \wedge T - t\right)}$$

Finally, $\hat{\nu}_{j,k}$ is the solution to:

$$0 = \sum_{t=1}^{T}\sum_{d=2}^{D}\left[\log\frac{\hat{\nu}_{j,k}}{2} + 1 - \psi\left(\frac{\hat{\nu}_{j,k}}{2}\right) + E_{j,k}^{t,d}[\log v] - E_{j,k}^{t,d}[v]\right] S_{j,k}^{t,d}$$

4.2.1.3. Fitting the Model

The model as we have described above, with $K = 2$ superstates, would lead to 34 parameters. To engender parsimony we impose some additional constraints. First, each holding-time and emission distribution can depend only on the coinciding superstate, and not the next. That is, $\eta_{j,k} = \eta_j$ and $\varepsilon_{j,k} = \varepsilon_j$ for each $j = 1, \ldots, K$, $k = 1, \ldots, K$. The manifestation of this constraint is that every matrix in (4.4) has constant rows. We additionally restricted $\sigma_{g(1:K,1:K)}$ and $\nu_{1:K,1:K}$ each to a single value, as can be seen from their matrices in (4.4). Finally we do not allow state self-transitions, i.e., the transition probability matrix τ is 0 along the diagonal. This yields a model with 14 free parameters, which is comparable to the models of Hamilton and Susmel (see Tab. 4.2).

$$\mu_g = \begin{bmatrix} 2.38 & 2.38 \\ -1.83 & -1.83 \end{bmatrix} \qquad \sigma_g^2 = \begin{bmatrix} 1.86 & 1.86 \\ 1.86 & 1.86 \end{bmatrix}$$
$$\mu_l = \begin{bmatrix} 0.40 & 0.40 \\ -2.09 & -2.09 \end{bmatrix} \qquad \sigma_l^2 = \begin{bmatrix} 1.43 & 1.43 \\ 2.60 & 2.60 \end{bmatrix}$$
$$\phi = \begin{bmatrix} 0.28 & 0.28 \\ 0.20 & 0.20 \end{bmatrix} \qquad \nu = \begin{bmatrix} 4.35 & 4.35 \\ 4.35 & 4.35 \end{bmatrix}$$
$$r = \begin{bmatrix} 0.32 & 0.32 \\ 89.66 & 89.66 \end{bmatrix} \qquad p = \begin{bmatrix} 0.02 & 0.02 \\ 0.98 & 0.98 \end{bmatrix}$$
$$\tau = \begin{bmatrix} 0.00 & 1.00 \\ 1.00 & 0.00 \end{bmatrix} \tag{4.4}$$

The MAP renewal sequence found for this model is displayed in Fig. 4.2 on page 113.

4.2.2. AR(1)-SV HMRM Forecast Formula

We wish to compute the 1-step ahead forecasts $E\left[x_{t+1}|x_{1:t}\right], Var\left[x_{t+1}|x_{1:t}\right]$ under our AR(1)-SV HMRM. Our derivations yielding the formulas for these quantities are somewhat involved, see Section A.6. There we have derived forecasts for arbitrary steps-ahead.

$$
\begin{aligned}
&E\left[x_{t+1}|x_{1:t}\right] \\
&\overset{\substack{(A.34)\\(A.36)\\(A.38)}}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\mu_{g(j,k)} \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^t \\
&\quad + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{d=1}^{(D-1)\wedge t}\left(\phi_{j,k}x_t + \mu_{l(j,k)}\right) \\
&\qquad \times f_{\varepsilon_{j,k}^d}(x_{t+1-d:t}) \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^{t-d}
\end{aligned}
$$

$$
\begin{aligned}
&Var\left[x_{t+1}|x_{1:t}\right] \\
&\overset{\substack{(A.34)\\(A.36)\\(A.38)}}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sigma_{g(j,k)}^2 \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^t \\
&\quad + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{d=1}^{(D\wedge t+1)-1}\frac{\nu_{j,k}+d-1}{\nu_{j,k}+d-3} \times \sigma_{l(j,k)}^2 \\
&\qquad \times \frac{\nu_{j,k} + \sum_{\delta=1}^{d-1}(x_{t+1-d+\delta} - \phi_{j,k}x_{t-d+\delta} - \mu_{l(j,k)})^2/\sigma_{l(j,k)}^2}{\nu_{j,k}+(d-1)} \\
&\qquad \times f_{\varepsilon_{j,k}^d}(x_{t+1-d:t}) \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^{t-d}
\end{aligned}
$$

4.2.3. Comparison with Hamilton & Susmel

Using the same dataset, we compare the forecasts of our AR(1)-SV HMRM to the regime-switching ARCH models of Hamilton and Susmel [16][3]. All of Hamilton and Susmel's results are conditioned on the first four observations in the dataset. For the 1 step-ahead forecast, the following loss functions were considered:

$$\begin{aligned}
MSE &= \frac{1}{T-4}\sum_{t=5}^{T}\left\{(x_t - E\left[x_t|x_{1:t-1}\right])^2 - Var\left[x_t|x_{1:t-1}\right]\right\}^2 \\
MAE &= \frac{1}{T-4}\sum_{t=5}^{T}|(x_t - E\left[x_t|x_{1:t-1}\right])^2 - Var\left[x_t|x_{1:t-1}\right]| \\
[LE]^2 &= \frac{1}{T-4}\sum_{t=5}^{T}\left\{\log\left((x_t - E\left[x_t|x_{1:t-1}\right])^2\right) - \log Var\left[x_t|x_{1:t-1}\right]\right\}^2 \\
|LE| &= \frac{1}{T-4}\sum_{t=5}^{T}|\log\left((x_t - E\left[x_t|x_{1:t-1}\right])^2\right) - \log Var\left[x_t|x_{1:t-1}\right]|
\end{aligned}$$

All these loss functions are compared against a baseline model that Hamilton and Susmel call a "constant variance" model. Under this constant variance model:

$$\begin{aligned}
E_{\binom{\text{constant}}{\text{variance}}}\left[x_t \mid x_{1:t-1}\right] &= \frac{1}{T-4}\sum_{t'=5}^{T} x_{t'} && \text{for } [t=5,\ldots,T] \\
Var_{\binom{\text{constant}}{\text{variance}}}\left[x_t \mid x_{1:t-1}\right] &= \frac{1}{T-4}\sum_{t'=5}^{T}\left(x_t - E\left[x_{t'} \mid x_{1:t'-1}\right]\right)^2 && \text{for } [t=5,\ldots,T]
\end{aligned}$$

We were able to reproduce Hamilton and Susmel's published loss functions for the

[3] Only the 1-week-ahead forecast is used because we were unable to reproduce their values for the 4 and 8-week-ahead forecasts. We used the `rugarch` R package to successfully reproduce the log likelihood and 1 step-ahead forecast loss functions for Student t GARCH-L(1,1) model, but were unable to reproduce the 4 and 8-step-ahead forecast loss functions for this model. So we are not confident that we understand exactly how the loss functions were computed for these forecasts, and thus not confident a comparison with our model would be equitable.

"Student t GARCH-L(1,1)" model. The last row in Tab. 4.2 on this page contains the percent improvement of these loss functions computed under our AR(1)-SV HMRM. The baseline for this improvement is the constant variance model. The other rows in Tab. 4.2 were obtained from Hamilton and Susmel's paper. Tab. 4.2 shows that our model's forecast improves those of Hamilton and Susmel [16] by approximately 50% for each loss function considered.

			Percent improvement in loss			
Model	# Param.	$\log \mathcal{L}$	MSE	MAE	$[LE]^2$	$\lvert LE\rvert$
Student t GARCH-L(1,1)	7	-2822.0	-8	3	**15**	**10**
Student t SWARCH-L(3,2)	13	-2802.7	-7	11	10	**10**
Student t SWARCH-L(4,2)	15	-2798.1	**6**	**13**	7	9
2-state AR(1)-SV HMRM	14	-2798.6	9	21	22	15

Table 4.2.: Metrics from the models of Hamilton and Susmel [16] and our model. For each loss function, the best performing model(s) of Hamilton and Susmel [16] are emboldened. Because Hamilton and Susmel treated the first four observations as given, for our model we used $\log \mathcal{L} = p(x_{5:T}|x_{1:4})$.

Figure 4.2.: The MAP renewal sequence and variances of our 2-state AR(1)-SV HMRM model (4.4) overlayed on NYSE weekly returns from July 31, 1962 to December 29, 1987.

5. Conclusion

Each model we presented is based on observed values being dependent on some unobserved sequence. We first introduced the finite-mixture model, a model where each observation is dependent on an unobserved state. These states are iid from a probability vector of length K; this probability vector is a parameter in the model. Next we presented the hidden Markov model, which replaces the iid sequence of states with a sequence drawn from a Markov chain. These two models, the mixture model and the hidden Markov model, are precursors of our model, the hidden Markov renewal process.

The HMRM replaces the HMM's Markov chain with a Markov renewal process. The Markov renewal process can be thought of as a Markov chain augmented with renewal-times. The observations in an HMRM consist of observation subsequences, rather than single observations. The start of each subsequence coincides with the renewal-times of the hidden Markov renewal sequence.

Our development of the HMRM drew a parallel with the HMM, presenting a theorem that shows how knowledge of a renewal segregates the hidden and observed sequences into two conditionally independent sets of random variables: those occurring before

the renewal and those occurring after (or at) the renewal. We also presented a graph showing how renewals are related in the forward and Viterbi algorithms. The development of the forward-backward algorithm for the HMRM distinguished between forward/backward *renewal* probabilities and forward/backward *sojourn* probabilities.

Our HMRM extends previous HSMMs. Whereas previous HSMMs allow each observation subsequence to depend on only the coinciding superstate, the HMRM allows each subsequence to depend on the next superstate also. This allows, for example, the observations to gradually transition between the superstates.

Our formulation of the HMRM was initially kept general, we did not specify the emission or holding-time distributions. To more concretely demonstrate our HMRM we presented three sub-models that, by specifying these distributions, can be employed practically. We called the first sub-model the bridging-means sub-model. It illustrated how our extension could be employed to model observation sequences that gradually transition between superstates.

The next sub-model demonstrated that by using a hidden Markov renewal process, rather than a semi-Markov process, phenomena such as jumps can be inferred from an observation sequence. This model set the number of possible superstates, K, to 1. So the superstate sequence was constant, containing no information. This elucidated the importance of the renewal-time sequence, which in our jump sub-model, contains the times of the jumps. This renewal-time sequence is absent in an SMP.

Our last sub-model augmented the state space of the hidden Markov renewal process to include (inverse) variances. Because these variances are random, we termed the model a "stochastic volatility" sub-model. Throughout our exposition of HMRM sub-

models, we presented plots with overlays of rectangles – each rectangle corresponding to an observation subsequence. The width of the rectangles in HMRM based model are randomly distributed according to the holding-time distributions. In our stochastic volatility model, the height of each rectangle is also random. As in our jump sub-model, we allowed for only one superstate value. This provided another example of associating an event other than a state change with a renewal.

We showed how to simulate from each sub-model. Then, using only the observation sequence, we showed how to estimate the sub-model's parameters and the unobserved state/renewal sequence. By doing this with the three sub-models presented, we demonstrate that our HMRM framework can be used to perform inference for models that incorporate three key areas of financial time series: regime-switching, jumps, and stochastic volatility. Further, the bridging-means model demonstrated how regime-switching in our model can be gradual, rather than abrupt. This is because we have allowed observation subsequences to depend on both current and future superstates.

The penultimate chapter applied HMRMs to two real-world datasets. The first was daily returns from four European sector indices, and was previously analyzed by Bulla and Bulla, using the HSMM of Guédon [15]. Our HMRM was able to better model the ACF of squared returns for the indices considered. The gradual regime-switching of the bridging-means model was not the source of the improved modeling however. A HMRM that assumed the observation subsequences to be conditionally iid performed just as well.

The second real-world dataset was weekly returns from a portfolio of stocks on the

NYSE. We used an HMRM that modeled autoregressions and stochastic volatility. Compared to the regime-switching ARCH models of Hamilton and Susmel [16], this HMRM improved volatility forecasts by approximately 50% for each metric considered.

A. Derivations and Proofs

A.1. Mixture Model

A.1.1. Complete Data Likelihood, Posterior State Probability, Posterior Expectation

An expression for the complete data likelihood of a state sequence $s_{1:T}$ and observation sequence $x_{1:T}$ in a mixture model with parameter set $\theta = \{\alpha_{1:K}, \varepsilon_{1:K}\}$ is given by:

$$\begin{aligned} p(x_{1:T}, s_{1:T}; \theta) &\overset{(CRP)}{=} p(x_{1:T}|s_{1:T}; \theta)p(s_{1:T}; \theta) \\ &\overset{\substack{(1.3b)\\(1.3d)}}{=} \prod_{t=1}^{T} p(x_t|s_t; \theta)p(s_t; \theta) \\ &= \prod_{t=1}^{T} \varepsilon_{s_t}(x_t)\alpha_{s_t} \end{aligned} \tag{A.1}$$

For a mixture model with parameter set $\theta^{(n)} = \{\alpha_{1:K}^{(n)}, \varepsilon_{1:K}^{(n)}\}$, the posterior state probabilities, $A_k^t = p(s_t = k|x_{1:T}; \theta^{(n)})$, can be computed:

$$\begin{aligned} p(s_t = k|x_{1:T}; \theta^{(n)}) &\overset{\substack{(1.3b)\\(CRP)}}{=} p(s_t = k|x_t; \theta^{(n)}) && \text{for } \begin{bmatrix} t=1,\ldots,T \\ k=1,\ldots,K \end{bmatrix} \\ &= \frac{p(x_t|s_t = k; \theta^{(n)})p(s_t = k; \theta^{(n)})}{\sum_{j=1}^K p(x_t|s_t = j; \theta^{(n)})p(s_t = j; \theta^{(n)})} \\ &= \frac{\varepsilon_k^{(n)}(x_t)\alpha_k^{(n)}}{\sum_{j=1}^K \varepsilon_j^{(n)}(x_t)\alpha_j^{(n)}} && \text{(A.2)} \end{aligned}$$

An expression for the posterior expectation of a function of a state subsequence is:

$$\begin{aligned} E_{s_{1:T}|x_{1:T};\theta^{(n)}}\left[f(s_{t:u})\right] &= \sum_{s_{1:T}} f(s_{t:u})p(s_{1:T}|x_{1:T}; \theta) && \text{for } \left[1 \le t \le u \le T\right] \\ &= \sum_{s_{t:u}} f(s_{t:u}) \sum_{s_{1:t-1}} \sum_{s_{u+1:T}} p(s_{1:T}|x_{1:T}; \theta) \\ &= \sum_{s_{t:u}} f(s_{t:u})p(s_{t:u}|x_{1:T}; \theta) && \text{(A.3)} \end{aligned}$$

We did not use any of the mixture model assumptions (1.3); we will use this formula again with the HMM.

A.1.2. Optimizing Q_α and Q_ε

The following results are used to computed the update formulas for the mixture model EM algorithm.

Lemma A.1. *The maximizing values* $\hat{\alpha}_{1:K}$ *for*

$$\sum_{k=1}^{K} \log \alpha_k A_k$$

subject to the constraint

$$\sum_{k=1}^{K} \alpha_k = 1$$

satisfy:

$$\hat{\alpha}_k = \frac{A_k}{\sum_{j=1}^{K} A_j} \tag{A.4}$$

Proof. The optimality conditions [4] require:

$$\begin{pmatrix} A_1/\hat{\alpha}_1 \\ \vdots \\ A_K/\hat{\alpha}_K \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Multiplying each equation by $\hat{\alpha}_k/\lambda$ and summing all the equations gives

$$\frac{\sum_{k=1}^{K} A_k}{\lambda} \quad = \quad \underbrace{\sum_{k=1}^{K} \hat{\alpha}_k}_{=1}$$

where the right-hand side sums to 1 by the constraint. Hence $\lambda = \sum_{k=1}^{K} A_k$ and $\hat{\alpha}_k = A_k / \left[\sum_{j=1}^{K} A_j\right]$. □

Lemma A.2. *The values* $\hat{\mu}_{1:K}, \hat{\sigma}_{1:K}$ *maximizing*

$$Q(\mu_{1:K}, \sigma_{1:K}) \triangleq \sum_{k=1}^{K} \sum_{t=1}^{T} \log f_{\mathcal{N}}(x_t; \mu_k, \sigma_k^2)\, A_k^t Q(\mu_{1:K}, \sigma_{1:K}) \triangleq \sum_{k=1}^{K} \sum_{t=1}^{T} \log f_{\mathcal{N}}(x_t; \mu_k, \sigma_k^2)\, A_k^t$$

are

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T} x_t A_k^t}{\sum_{t=1}^{T} A_k^t} \tag{A.5a}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{t=1}^{T} (x_t - \hat{\mu}_k)^2 A_k^t}{\sum_{t=1}^{T} A_k^t} \tag{A.5b}$$

Proof. The log of the normal density gives:

$$Q(\mu_{1:K}, \sigma_{1:K}) \propto \sum_{k=1}^{K} \sum_{t=1}^{T} \left[\log \sigma_k + \frac{1}{2} \frac{(x_t - \mu_k)^2}{\sigma_k^2} \right] A_k^t$$

The optimality conditions for μ_k imply that

$$\sum_{t=1}^{T} \hat{\mu}_k A_k^t = \sum_{t=1}^{T} x_t A_k^t$$

and solving for $\hat{\mu}_k$ yields the first equation in (A.5b). The optimality conditions for σ_k imply that

$$\sum_{t=1}^{T} \hat{\sigma}_k^2 A_k^t = \sum_{t=1}^{T} (x_t - \hat{\mu}_k)^2 A_k^t$$

□

A.2. Hidden Markov Model

A.2.1. Complete Data Likelihood, Posterior State/Transition Probabilities

An expression for the complete data likelihood of a state sequence $s_{1:T}$ and observation sequence $x_{1:T}$ in an HMM with parameter set $\theta = \{\iota_{1:K}, \tau_{1:K,1:K}, \varepsilon_{1:K}\}$ is given by:

$$\begin{aligned} p(x_{1:T}, s_{1:T}) &\overset{(CRP)}{=} p(x_{1:T}|s_{1:T})p(s_{1:T}) \\ &\overset{\substack{(1.14b)\\(1.11a)}}{=} \prod_{t=1}^{T} p(x_t|s_t)p(s_t|s_{t-1}) \end{aligned} \tag{A.6}$$

First we show how the posterior state probability can be computed as a product of forward and backward probabilities

$$\begin{aligned} p(s_t = k, x_{1:T}) &\overset{(CRP)}{=} p(x_{t:T}|s_t = k, x_{1:t-1})p(s_t = k, x_{1:t-1}) \qquad \text{for } \begin{bmatrix} t=1,\dots,T \\ k=1,\dots,K \end{bmatrix} \\ &\overset{(1.15)}{=} p(x_{t:T}|s_t = k, x_{1:t-1})p(s_t = k, x_{1:t-1}) \end{aligned} \tag{A.7}$$

and the posterior state probability can be computed from the joint probability as $p(s_t = k|x_{1:T}) = p(s_t = k, x_{1:T})/\sum_j p(s_t = j, x_{1:T})$. The posterior transition probabil-

ities can be similarly factored:

$$\begin{aligned} p(s_t = j, s_{t+1} = k, x_{1:T}) &\overset{(CRP)}{=} p(x_{t+1:T}|s_t = j, s_{t+1} = k, x_{1:t-1}) \qquad \text{for } \begin{bmatrix} t=1,\ldots,T-1 \\ k=1,\ldots,K \end{bmatrix} \\ &\quad \times p(s_{t+1} = k|s_t = j, x_{1:t}) \\ &\quad \times p(x_t|s_t = j, x_{1:t-1}) \\ &\quad \times p(s_t = j, x_{1:t-1}) \\ &\overset{(1.15)}{=} p(x_{t+1:T}|s_{t+1} = k) \\ &\quad \times p(s_{t+1} = k|s_t = j) \\ &\quad \times p(x_t|s_t = j) \\ &\quad \times p(s_t = j, x_{1:t-1}) \end{aligned} \tag{A.8}$$

A.2.2. Optimizing Q_τ

The following result is used to computed the update formula for the TPM in the hidden Markov model EM algorithm.

Lemma A.3. *The maximizer* $\hat{\tau}_{1:K,1:K}$ *for*

$$\sum_{j=1}^{K}\sum_{k=1}^{K} \log \tau_{j,k} \sum_{t=2}^{T} N_{j,k}^{t}$$

subject to the constraints

$$\sum_{k=1}^{K} \tau_{j,k} = 1 \text{ for } j = 1, \ldots, K$$

satisfies:

$$\hat{\tau}_{j,k} = \frac{\sum_{t=2}^{T} N_{j,k}^{t}}{\sum_{l=1}^{K} \sum_{t=2}^{T} N_{j,l}^{t}} \tag{A.9}$$

Proof. The optimality conditions [4] require, for each $j = 1, \ldots, K$:

$$\begin{pmatrix} \sum_{t=2}^{T} N_{1,1}^{t}/\hat{\tau}_{1,1} & \cdots & \sum_{t=2}^{T} N_{1,K}^{t}/\hat{\tau}_{1,K} \\ \vdots & \ddots & \vdots \\ \sum_{t=2}^{T} N_{K,1}^{t}/\hat{\tau}_{K,1} & \cdots & \sum_{t=2}^{T} N_{K,K}^{t}/\hat{\tau}_{K,K} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \cdots & \lambda_1 \\ \vdots & \ddots & \vdots \\ \lambda_K & \cdots & \lambda_K \end{pmatrix}$$

Here we have represented K^2 equations as a single $K \times K$ matrix equation. Dividing each equation by the corresponding λ term and multiplying by the τ term gives

$$\begin{pmatrix} \sum_{t=2}^{T} N_{1,1}^{t}/\lambda_1 & \cdots & \sum_{t=2}^{T} N_{1,K}^{t}/\lambda_1 \\ \vdots & \ddots & \vdots \\ \sum_{t=2}^{T} N_{K,1}^{t}/\lambda_K & \cdots & \sum_{t=2}^{T} N_{K,K}^{t}/\lambda_K \end{pmatrix} = \begin{pmatrix} \hat{\tau}_{1,1} & \cdots & \hat{\tau}_{1,K} \\ \vdots & \ddots & \vdots \\ \hat{\tau}_{K,1} & \cdots & \hat{\tau}_{K,K} \end{pmatrix}$$

Computing the sum of each row and applying the constraint that $\sum_{k=1}^{K} \hat{\tau}_{j,k} = 1$ gives

$$\begin{pmatrix} \sum_{k=1}^{K} \sum_{t=2}^{T} N_{1,k}^{t}/\lambda_1 \\ \vdots \\ \sum_{k=1}^{K} \sum_{t=2}^{T} N_{K,k}^{t}/\lambda_K \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

so $\lambda_j = \sum_{k=1}^{K} \sum_{t=2}^{T} N_{j,k}^{t}$ for each $j = 1, \ldots, K$. Hence $\hat{\tau}_{j,k}$ is as in (A.9). □

A.3. Hidden Markov Renewal Model

A.3.1. HMRM Homogeneity

We first establish that the HMRM is homogeneous – the probability distribution of any future renewals and observations is constant with respect to the number of renewals that have occurred previously. This result is perhaps to be expected given MRP homogeneity (2.2b) and the homogeneity of the observations implied by (2.3a).

Theorem A.1 (HMRM Homogeneity)**.** *The distribution* $p(x_{t_r:T}, t_{r+1:r+\rho}, z_{r+1:r+\rho}|t_r, z_r)$ *is constant with respect to* r*. That is assuming* $t_{r':r'+\rho} = t_{r:r+\rho}$ *and* $z_{r':r'+\rho} = z_{r:r+\rho}$ *then*

$$p(x_{t_r:T}, t_{r+1:r+\rho}, z_{r+1:r+\rho}|t_r, z_r) = p(x_{t:T}, t_{r'+1:r'+\rho}, z_{r'+1:r'+\rho}|t_{r'}, z_{r'}) \tag{A.10}$$

Proof. Factoring both sides of (A.10) and applying the homogeneity properties im-

plied by (2.3a), (2.2b), shows that the two quantities are equal.

$$
\begin{aligned}
&p(x_{t_r:T}, t_{r+1:r+\rho}, z_{r+1:r+\rho} | t_r, z_r) \\
&\quad \overset{(CRP)}{=} p(x_{t_r:T} | t_{r:r+\rho}, z_{r:r+\rho}) \times p(t_{r+1:r+\rho}, z_{r+1:r+\rho} | t_r, z_r) \\
&\quad \overset{(CRP)}{=} \prod_{\substack{u:u\geq r,\\ t_u \leq T}} p(x_{t_u:(T\wedge t_{u+1}-1)} | x_{t_r:t_u-1}, t_{r:r+\rho}, z_{r:r+\rho}) \\
&\qquad\qquad \times \prod_{u=r}^{r+\rho-1} p(t_{u+1}, z_{u+1} | t_{r:u}, z_{r:u}) \\
&\quad \overset{\substack{(2.3b)\\(2.3a)\\(2.2)}}{=} \prod_{\substack{u:u\geq r,\\ t_u \leq T}} \varepsilon^{d_u}_{z_u, z_{u+1}}(x_{t_u:(T\wedge t_{u+1}-1)}) \times \prod_{u=r}^{r+\rho-1} \eta_{z_u, z_{u+1}}(d_u) \tau_{z_u, z_{u+1}}
\end{aligned}
$$

In the exact same manner as above we can show that:

$$
\begin{aligned}
&p(x_{t:T}, t_{r'+1:r'+\rho}, z_{r'+1:r'+\rho} | t_{r'}, z_{r'}) \\
&\qquad = \prod_{\substack{u:u\geq r',\\ t_u \leq T}} \varepsilon^{d_u}_{z_u, z_{u+1}}(x_{t'_u:(T\wedge t'_{u+1}-1)}) \times \prod_{u=r'}^{r'+\rho-1} \eta_{z_u, z_{u+1}}(d_u) \tau_{z_u, z_{u+1}}
\end{aligned}
$$

It is then clear that if $t_{r':r'+\rho} = t_{r:r+\rho}$ and $z_{r':r'+\rho} = z_{r:r+\rho}$, the two sides of (A.10) are also equal. □

A.3.2. Forward/Backward Probabilities

Here we derive formulas for the forward/backward sojourn/renewal probabilities. First the forward sojourn probabilities:

$$
\begin{aligned}
\phi_{j,k}^{t,d} &\triangleq p(\exists r\, s.t.\, z_r = j,\, z_{r+1} = k,\, t_r = t+1-d,\, t_{r+1} = t+1,\, x_{1:t}) \\
&= \sum_r p(z_r = j,\, z_{r+1} = k,\, t_r = t+1-d,\, t_{r+1} = t+1,\, x_{1:t}) \\
&\overset{(CRP)}{=} \sum_r p(x_{t+1-d:t} | z_r = j,\, z_{r+1} = k,\, t_r = t+1-d,\, t_{r+1} = t+1,\, x_{1:t-d}) \\
&\quad \times p(z_{r+1} = k,\, t_{r+1} = t+1 | s.t.\, z_r = j,\, t_r = t+1-d,\, x_{1:t-d}) \\
&\quad \times p(z_r = j,\, t_r = t+1-d,\, x_{1:t-d}) \\
&\overset{\substack{(2.7)\\(A.10)}}{=} p(x_{t+1-d:t} | z_r = j,\, z_{r+1} = k,\, t_r = t+1-d,\, t_{r+1} = t+1) \\
&\quad \times p(z_{r+1} = k,\, t_{r+1} = t+1 | z_r = j,\, t_r = t+1-d) \\
&\quad \times \sum_r p(z_r = j,\, t_r = t+1-d,\, x_{1:t-d}) \\
&\overset{\substack{(2.3b)\\(2.2b)}}{=} \varepsilon_{j,k}^d(x_{t+1-d:t}) \times \eta_{j,k}(d)\, \tau_{j,k} \times F_j^{t-d}
\end{aligned} \tag{A.11}
$$

The forward renewal probabilities:

$$
\begin{aligned}
F_k^t &\triangleq p(\exists r\, s.t.\, z_r = k,\, t_r = t+1,\, x_{1:t}) \\
&= \sum_{j=1}^K \sum_{d=1}^{D \wedge t} p(\exists r\, s.t.\, z_{r-1} = j,\, z_r = k,\, t_{r-1} = t+1-d,\, t_r = t+1,\, x_{1:t}) \\
&\overset{(2.8b)}{=} \sum_{j=1}^K \sum_{d=1}^{D \wedge t} \phi_{j,k}^{t,d}
\end{aligned} \tag{A.12}
$$

The backward sojourn probabilities, for $t + d \leq T$

$$\begin{aligned}
\beta_{j,k}^{t,d} &\triangleq p(x_{t:T},\, z_{r+1} = k,\, t_{r+1} = t + d | z_r = j,\, t_r = t) \\
&\overset{(CRP)}{=} p(x_{t+d:T}, | x_{t:t+d-1},\, z_{r+1} = k,\, t_{r+1} = t + d\, z_r = j,\, t_r = t) \\
&\quad \times p(x_{t:t+d-1} | z_{r+1} = k,\, t_{r+1} = t + d,\, z_r = j,\, t_r = t) \\
&\quad \times p(z_{r+1} = k,\, t_{r+1} = t + d | z_r = j,\, t_r = t) \\
&\overset{\substack{(2.7)\\(2.3b)\\(2.2b)}}{=} B_j^{t+d} \times \varepsilon_{j,k}^{d}(x_{t:t+d-1}) \times \eta_{j,k}(d)\, \tau_{j,k}
\end{aligned} \tag{A.13a}$$

and when $t + d \geq T + 1$:

$$\begin{aligned}
\beta_{j,k}^{t,d} &\triangleq p(x_{t:T},\, z_{r+1} = k,\, t_{r+1} = t + d | z_r = j,\, t_r = t) \\
&\overset{(CRP)}{=} p(x_{t:T} | z_r = j,\, z_{r+1} = k,\, t_r = t,\, t_{r+1} = t + d) \\
&\quad \times p(z_{r+1} = k,\, t_{r+1} = t + d | z_r = j,\, t_r = t) \\
&\overset{\substack{(2.3b)\\(2.2b)}}{=} \varepsilon_{j,k}^{d}(x_{t:T}) \times \eta_{j,k}(d)\, \tau_{j,k}
\end{aligned} \tag{A.13b}$$

Finally the backward renewal probabilities can be computed:

$$\begin{aligned}
B_j^t &\triangleq p(x_{t:T} | z_r = k,\, t_r = t) \\
&= \sum_{k=1}^{K} \sum_{d=1}^{D} p(x_{t:T},\, z_{r+1} = k,\, t_{r+1} = t + d | z_r = j,\, t_r = t) \\
&\overset{(A.13)}{=} \sum_{k=1}^{K} \sum_{d=1}^{D} \beta_{j,k}^{t,d}
\end{aligned} \tag{A.14}$$

A.3.3. Complete Data Likelihood, Posterior Renewal/Sojourn Probabilities, Likelihood

An expression for the complete data likelihood of a renewal sequence $(z_{1:R+1}, t_{1:R+1})$ with $t_R \leq T < t_{R+1}$ and observation sequence $x_{1:T}$ in an HMRM with parameter set $\theta = \{\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, \varepsilon_{1:K,1:K}\}$ is given by:

$$\begin{aligned}
& p(z_{1:R+1}, t_{1:R+1}, x_{1:T}) \\
&\quad = \quad p(x_{1:T}|z_{1:R+1}, t_{1:R+1}) \times p(z_{1:R+1}, t_{1:R+1}) \\
&\quad \overset{\substack{(2.3b)\\(CRP)}}{=} \prod_{r=1}^{R-1} \varepsilon_{z_r,z_{r+1}}^{d_r}(x_{t_r:t_{r+1}-1})\, \varepsilon_{z_r,z_{r+1}}^{d_R}(x_{t_R:T}) \\
&\qquad \times p(z_1, t_1) \times \prod_{r=1}^{R} p(z_{r+1}, t_{r+1}|z_{1:r}, t_{1:r}) \\
&\quad \overset{(2.2b)}{=} \prod_{r=1}^{R-1} \varepsilon_{z_r,z_{r+1}}^{d_r}(x_{t_r:t_{r+1}-1})\, \varepsilon_{z_r,z_{r+1}}^{d_R}(x_{t_R:T}) \\
&\qquad \times p(z_1, t_1) \times \prod_{r=1}^{R} \tau_{z_r,z_{r+1}} \eta_{z_r,z_{r+1}}(d_r)
\end{aligned} \tag{A.15}$$

The posterior renewal probability can computed by first finding a formula for the

joint probability

$$
\begin{aligned}
&p(\exists r \; s.t. \; z_r = k, t_r = t, x_{1:T}) \\
&\quad = \quad \sum_r p(z_r = k,\, t_r = t,\, x_{1:T}) \\
&\quad \overset{(CRP)}{=} \quad p(x_{t:T}|x_{1:t-1}, z_r = k,\, t_r = t) p(z_r = k,\, t_r = t,\, x_{1:t-1}) \\
&\quad \overset{\substack{(2.1)\\(A.10)}}{=} \quad p(x_{t:T}|z_r = k,\, t_r = t) \sum_r p(z_r = k,\, t_r = t,\, x_{1:t-1}) \\
&\quad = \quad p(x_{t:T}|z_r = k,\, t_r = t) p(\exists r \; s.t. \; z_r = k,\, t_r = t,\, x_{1:t-1}) \\
&\quad = \quad B_k^t F_k^{t-1}
\end{aligned} \tag{A.16}
$$

and then, since $t_1 = 1$, the likelihood can be computed

$$
\begin{aligned}
L \triangleq p(x_{1:T}) \quad &= \quad \sum_{k=1}^{K} p(\exists r \; s.t. \; z_r = k, t_1 = 1, x_{1:T}) \\
&\overset{(A.16)}{=} \quad \sum_{k=1}^{K} B_k^t F_k^{t-1}
\end{aligned} \tag{A.17}
$$

and finally normalizing the joint by the likelihood:

$$
\begin{aligned}
p(\exists r \; s.t. \; z_r = k, t_r = t|x_{1:T}) \quad &= \quad p(\exists r \; s.t. \; z_r = k, t_r = t, x_{1:T}) p(x_{1:T}) \\
&\overset{(A.16)}{=} \quad B_k^t F_k^{t-1} / L
\end{aligned} \tag{A.18}
$$

To get the computational formula for the posterior sojourn probabilities we first

compute the joint

$$
\begin{aligned}
&p(\exists r \text{ s.t. } z_r = j, t_r = t, z_{r+1} = k, t_{r+1} = t + d, x_{1:T}) \\
&\quad = \sum_r p(z_r = j, t_r = t, z_{r+1} = k, t_{r+1} = t + d, x_{1:T}) \\
&\quad \overset{(CRP)}{=} \sum_r p(x_{t:T}, z_{r+1} = k, t_{r+1} = t + d | z_r = j, t_r = t, x_{1:t-1}) \\
&\qquad\qquad \times p(z_r = j, t_r = t, x_{1:t-1}) \\
&\quad \overset{(2.7)}{\underset{(A.10)}{=}} p(x_{t:T}, z_{r+1} = k, t_{r+1} = t + d | z_r = j, t_r = t) \\
&\qquad\qquad \times \sum_r p(z_r = j, t_r = t, x_{1:t-1}) \\
&\quad = \beta_{j,k}^{t,d} \times F_j^{t-1} \qquad\qquad (A.19)
\end{aligned}
$$

and normalize by the likelihood

$$
\begin{aligned}
&p(\exists r \, s.t. \, z_r = j, t_r = t, z_{r+1} = k, t_{r+1} = t + d | x_{1:T}) \\
&\quad = p(\exists r \, s.t. \, z_r = j, t_r = t, z_{r+1} = k, t_{r+1} = t + d | x_{1:T}) / p(x_{1:T}) \\
&\quad \overset{(A.19)}{=} \beta_{j,k}^{t,d} \times F_j^{t-1} / L
\end{aligned}
$$

A.3.4. EM Algorithm

Lemma A.4 shows how to compute the general form of an expectation that appears multiple times in the "E" step of the EM algorithm for the HMRM.

Lemma A.4. *If* $z_{1:R+1}, t_{1:R+1} \sim \mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, \varepsilon_{1:K,1:K}, T)$ *then for any*

$r = 1, \ldots, R$

$$E_{z_{1:R+1}, t_{1:R+1}|x_{1:T};\theta}\left[\sum_{r=1}^{R} f(z_r, z_{r+1}, t_r, t_{r+1})\right] = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} f(j,k,t,t+d)S_{j,k}^{t,d} \quad \text{(A.20)}$$

Proof. [1]Let $\mathcal{P} \triangleq \left\{ {z_{1:R+1}, \atop t_{1:R+1}} \; s.t.\, t_R \leq T < t_{R+1} \right\}$. For any $z_{1:R+1}, t_{1:R+1} \in \mathcal{P}$:

$$r \in \{1, \ldots, R\} \implies t_r \in \{1, \ldots, T\}$$

by definition of R. Further, by definition of D:

$$t_r = t \implies t_{r+1} \in \{t+1, \ldots, t+D\}$$

Finally, $z_r \in \{1, \ldots, K\}$, $z_{r+1} \in \{1, \ldots, K\}$ for any r.

$$\begin{aligned}
&\sum_{{z_{1:R+1}, \atop t_{1:R+1}} \in \mathcal{P}} \sum_{r=1}^{R} f(z_r, z_{r+1}, t_r, t_{r+1}) p(z_{1:R+1}, t_{1:R+1}|x_{1:T}) \\
&\quad = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \sum_{{z_{1:R+1}, \atop t_{1:R+1}} \in \mathcal{P}} \sum_{r=1}^{R} f(j,k,t,t+d) \\
&\qquad\qquad \times p(z_{1:R+1}, t_{1:R+1}|x_{1:T}) \mathbb{I}\{z = j, z_{r+1} = k, t_r = t, t_{r+1} = d\} \\
&\quad = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \sum_{\substack{{z_{1:R+1}, \atop t_{1:R+1}} \in \mathcal{P}\ s.t. \\ \exists r \text{ with} \\ t_r = t, t_{r+1} = t+d, \\ z_r = j, z_{r+1} = k}} f(j,k,t,t+d) p(z_{1:R+1}, t_{1:R+1}|x_{1:T})
\end{aligned}$$

[1]Our proof uses elements of both [12] and [1].

$$
\begin{aligned}
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} f(j,k,t,t+d) \sum_{\substack{z_{1:R+1},\\ t_{1:R+1}\in\mathcal{P}\ s.t.\\ \exists r \text{ with}\\ t_r=t, t_{r+1}=t+d,\\ z_r=j, z_{r+1}=k}} p(z_{1:R+1}, t_{1:R+1}|x_{1:T}) \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} f(j,k,t,t+d) S_{j,k}^{t,d}
\end{aligned}
$$

□

We additionally use the following result in the "E" step of the EM algorithm for the HMRM:

$$
\begin{aligned}
&E_{z_{1:R+1},t_{1:R+1}|x_{1:T};\theta^{(n)}}\left[f(z_1)\right] \\
&\quad= \sum_{\substack{z_{1:R+1},\\ t_{1:R+1}\in\mathcal{P}}} f(z_1) p(z_{1:R+1}, t_{1:R+1}|x_{1:T}) \\
&\quad= \sum_{k=1}^{K} f(k) \sum_{\substack{z_{1:R+1},\\ t_{1:R+1}\in\mathcal{P}\ s.t.\\ z_1=k}} p(z_{1:R+1}, t_{1:R+1}|x_{1:T}) \\
&\quad= \sum_{k=1}^{K} f(k)\, E_k^1 \qquad \text{(A.21)}
\end{aligned}
$$

A.4. Bridging-Means Sub-model

Each of (A.22)-(A.27) is defined to be a sum based on a contiguous subset $x_{t:(t+d)\wedge(T-t)}$ of $x_{1:T}$, e.g., $S_{t,d}(x_{1:T})$ is simply the sum of the elements in the $x_{t:(t+d)\wedge(T-t)}$. We given recursive expression for each of (A.22)-(A.27), enabling their computation in $\mathcal{O}(1)$ time, given the previously computed values.

Lemma A.5. *Given the following definitions ($\triangleq$) the following recursive equalities (=) hold:*

$$
\begin{aligned}
S_{t,d}(x_{1:T}) &\triangleq \sum_{\delta=0}^{d-1\wedge T-t} x_{t+\delta} \\
&= \begin{cases} S_{t,d-1}(x_{1:T}) + x_{t+d-1} & \text{if } t+d-1 \leq T \\ S_{t,d-1}(x_{1:T}) & \text{if } t+d-1 \geq T+1 \end{cases}
\end{aligned} \tag{A.22}
$$

$$
\begin{aligned}
E_{t,d}(x_{1:T}) &\triangleq \sum_{\delta=0}^{d-1\wedge T-t} \delta x_{t+\delta} \\
&= \begin{cases} E_{t,d-1}(x_{1:T}) + (d-1)\, x_{t+d-1} \\ E_{t,d-1}(x_{1:T}) \end{cases}
\end{aligned} \tag{A.23}
$$

$$
\begin{aligned}
D_{t,d}(x_{1:T}) &\triangleq \sum_{\delta=0}^{d-1\wedge T-t} (d-\delta)\, x_{t+\delta} \\
&= \begin{cases} D_{t,d-1}(x_{1:T}) + S_{t,d}(x_{1:T}) \\ D_{t,d-1}(x_{1:T}) + S_{t,d}(x_{1:T}) \end{cases}
\end{aligned} \tag{A.24}
$$

$$
\begin{aligned}
V_{t,d}(x_{1:T}) &\triangleq \sum_{\delta=0}^{d-1\wedge T-t} \delta^2 x_{t+\delta} \\
&= \begin{cases} V_{t,d-1}(x_{1:T}) + (d-1)^2\, x_{d-1} \\ V_{t,d-1}(x_{1:T}) \end{cases}
\end{aligned} \tag{A.25}
$$

$$
\begin{aligned}
U_{t,d}(x_{1:T}) &\triangleq \sum_{\delta=0}^{d-1\wedge T-t} (d-\delta)^2\, x_{t+\delta} \\
&= \begin{cases} U_{t,d-1}(x_{1:T}) + 2D_{t,d}(x_{1:T}) - S_{t,d}(x_{1:T}) \\ U_{t,d-1}(x_{1:T}) + 2D_{t,d}(x_{1:T}) - S_{t,d}(x_{1:T}) \end{cases}
\end{aligned} \tag{A.26}
$$

$$C_{t,d}(x_{1:T}) \triangleq \sum_{\delta=0}^{d-1\wedge T-t} \delta\,(d-\delta)\,x_{t+\delta}$$

$$= \begin{cases} C_{t,d-1}(x_{1:T}) + E_{t,d-1}(x_{1:T}) + (d-1)\,x_{t+d-1} \\ C_{t,d-1}(x_{1:T}) + E_{t,d-1}(x_{1:T}) \end{cases} \tag{A.27}$$

$$P_{t,d} \triangleq \sum_{\delta=0}^{d-1\wedge T-t} (d-\delta)^2$$

$$= \begin{cases} \frac{1}{6}(1+d)d(1+2d) \\ P_{t,d-1} - (T-t+1)(T-t-2d+1) \end{cases} \tag{A.28}$$

$$Q_{t,d} \triangleq \sum_{\delta=0}^{d-1\wedge T-t} \delta^2$$

$$= \begin{cases} Q_{t,d-1} + (d-1)^2 \\ Q_{t,d-1} \end{cases} \tag{A.29}$$

$$R_{t,d} \triangleq \sum_{\delta=0}^{d-1\wedge T-t} \delta(d-\delta)$$

$$= \begin{cases} \frac{1}{6}(d-1)d(d+1) \\ R_{t,d-1} + \frac{1}{2}(T-t+1)(T-t) \end{cases} \tag{A.30}$$

Proof. First we consider the case when $t+d-1 \leq T$. (A.22), (A.23), and (A.25) are all trivial. For (A.24) we note that:

$$D_{t,d}(x_{1:T}) - D_{t,d-1}(x_{1:T}) = \sum_{\delta=0}^{d-1} (d-\delta)\,x_{t+\delta} - \sum_{\delta=0}^{d-2} (d-1-\delta)\,x_{t+\delta} = x_{t+d-1} + \sum_{\delta=0}^{d-2} x_{t+\delta}$$

For (A.26), we use the fact that $z^2 - (z-1)^2 = 2z - 1$:

$$\begin{aligned} U_{t,d}(x_{1:T}) - U_{t,d-1}(x_{1:T}) =& \sum_{\delta=0}^{d-1} (d-\delta)^2 x_{t+\delta} - \sum_{\delta=0}^{d-2} (d-\delta-1)^2 x_{t+\delta} \\ =& x_{t+d-1} + 2\sum_{\delta=0}^{d-2} (d-\delta) x_{t+\delta} - \sum_{\delta=0}^{d-2} x_{t+\delta} \\ =& 2\sum_{\delta=0}^{d-1} (d-\delta) x_{t+\delta} - \sum_{\delta=0}^{d-1} x_{t+\delta} \end{aligned}$$

For (A.27):

$$\begin{aligned} C_{t,d}(x_{1:T}) - C_{t,d-1}(x_{1:T}) &= \sum_{\delta=0}^{d-1} \delta\,(d-\delta)\, x_{t+\delta} - \sum_{\delta=0}^{d-2} \delta\,(d-1-\delta)\, x_{t+\delta} \\ &= (d-1)\, x_{t+d-1} + \sum_{\delta=0}^{d-2} \delta x_{t+\delta} \end{aligned}$$

Now the case when $t + d - 1 \geq T + 1$. (A.22), (A.23), and (A.25) are all trivially 0. For (A.24) we note that:

$$D_{t,d}(x_{1:T}) - D_{t,d-1}(x_{1:T}) = \sum_{\delta=0}^{d-1\wedge T-t} (d-\delta)\, x_{t+\delta} - \sum_{\delta=0}^{d-2\wedge T-t} (d-1-\delta)\, x_{t+\delta} = \sum_{\delta=0}^{T-t} x_{t+\delta}$$

For (A.26):

$$\begin{aligned} U_{t,d}(x_{1:T}) - U_{t,d-1}(x_{1:T}) =& \sum_{\delta=0}^{d-1\wedge T-t} (d-\delta)^2 x_{t+\delta} - \sum_{\delta=0}^{d-2\wedge T-t} (d-1-\delta)^2 x_{t+\delta} \\ =& 2\sum_{\delta=0}^{T-t} (d-\delta) x_{t+\delta} - \sum_{\delta=0}^{T-t} x_{t+\delta} \end{aligned}$$

For (A.27):

$$\begin{aligned} C_{t,d}(x_{1:T}) - C_{t,d-1}(x_{1:T}) &= \sum_{\delta=0}^{d-1\wedge T-t} \delta\,(d-\delta)\,x_{t+\delta} - \sum_{\delta=0}^{d-2\wedge T-t} \delta\,(d-1-\delta)\,x_{t+\delta} \\ &= \sum_{\delta=0}^{T-t} \delta x_{t+\delta} \end{aligned}$$

□

A.5. Stochastic Volatility Sub-model

The following lemma is an analog to Lemma A.4 for a hidden *augmented* Markov renewal process.

Lemma A.6. *If* $(z_{1:R+1}, t_{1:R+1}, v_{1:R}) \sim \mathcal{AMRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, \varepsilon_{1:K,1:K}, T)$ *then*

$$\begin{aligned} &E_{\substack{z_{1:R+1}\\ t_{1:R+1}\\ v_{1:R}}\Big| x_{1:T}} \left[\sum_{r=1}^{R} f(z_r, z_{r+1}, t_r, t_{r+1}, v_r)\right] \\ &\quad = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} E_{v_r \Big| \substack{x_{t:t+d-1\wedge T}\\ z_r=j, z_{r+1}=k\\ t_r=t, t_{r+1}=t+d}} \left[f(j,k,t,t+d,v)\right] S_{j,k}^{t,d} \end{aligned} \tag{A.31}$$

Proof. Let $\mathcal{P} \triangleq \left\{ {z_{1:R+1},}\atop{t_{1:R+1}} s.t.\, t_R \le T < t_{R+1} \right\}$. Then

$$\begin{aligned} &E_{\substack{z_{1:R+1}\\ t_{1:R+1}\\ v_{1:R}}\Big| x_{1:T}} \left[\sum_{r=1}^{R+1} f(z_r, z_{r+1}, t_r, t_{r+1}, v_r)\right] \\ &\quad = \int_{v_{1:R}} \sum_{r=1}^{R} f(z_r, z_{r+1}, t_r, t_{r+1}, v_r) p(z_{1:R+1}, t_{1:R+1}, v_{1:R} | x_{1:T}) dv_{1:R} \end{aligned}$$

$$
\begin{aligned}
&\overset{\substack{(CRP)\\(3.12)}}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \sum_{\substack{z_{1:R+1}\\ t_{1:R+1}}\in \mathcal{P}} \int_{v_{1:R}} \sum_{r=1}^{R} f(j,k,t,t+d,v_r) \\
&\qquad \times p(v_r|z_r=j,t_r=t,t_{r+1}=t+d,x_{1:T}) \\
&\qquad \times p(v_{\setminus r}|z_{1:R+1},t_{1:R+1},x_{1:T})p(z_{1:R+1},t_{1:R+1}|x_{1:T}) \\
&\qquad \times \mathbb{I}\{z=j,z_{r+1}=k,t_r=t,t_{r+1}=d\}dv_{1:R} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \sum_{\substack{z_{1:R+1}\\ t_{1:R+1}}\in \mathcal{P}\\ \substack{s.t.\ \exists r \text{ with}\\ t_r=t,t_{r+1}=t+d,\\ z_r=j,z_{r+1}=k}} \int_{v_{1:R}} f(j,k,t,t+d,v_r) \\
&\qquad \times p(v_r|z_r=j,t_r=t,t_{r+1}=t+d,x_{1:T}) \\
&\qquad \times p(v_{\setminus r}|z_{1:R+1},t_{1:R+1},x_{1:T}) \\
&\qquad \times p(z_{1:R+1},t_{1:R+1}|x_{1:T})dv_{1:R} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \int_{v_r} f(j,k,t,t+d,v_r) \\
&\qquad \times p(v_r|z_r=j,t_r=t,t_{r+1}=t+d,x_{1:T})dv_r \\
&\qquad \times \sum_{\substack{z_{1:R+1}\\ t_{1:R+1}}\in \mathcal{P}\\ \substack{s.t.\ \exists r \text{ with}\\ t_r=t,t_{r+1}=t+d,\\ z_r=j,z_{r+1}=k}} \int_{v_{\setminus r}} p(v_{\setminus r}|z_{\setminus r},t_{1:R+1},x_{1:T})dv_{\setminus r}p(z_{1:R+1},t_{1:R+1}|x_{1:T}) \\
&\overset{(3.12)}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} \int_{v_r} f(j,k,t,t+d,v_r) \\
&\qquad \times p(v_r|z_r=j,t_r=t,t_{r+1}=t+d,x_{t:(t+d-1\wedge T)})dv_r S_{j,k}^{t,d} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{d=1}^{D} E_{v_r\left|\substack{x_{t:t+d-1\wedge T}\\ z_r=j,z_{r+1}=k\\ t_r=t,t_{r+1}=t+d}\right.} \left[f(j,k,t,t+d,v)\right] S_{j,k}^{t,d}
\end{aligned}
$$

□

A.6. Forecasting

We wish to find an expression for the f-step ahead forecast distribution: $p(x_{t+f}|x_{1:t})$. To this end, we first define *forecast renewal probabilities* as:

$$F_j^{t,f} \quad \triangleq \quad p\left(\exists r \; s.t. \; z_r = k, \; t_r = t+1+f, \; x_{1:t} \right) \text{for} \begin{bmatrix} t=0,\ldots,T \\ k=1,\ldots,K \\ f=0,1,\ldots \end{bmatrix} \tag{A.32}$$

Note that $F_j^{t,0} = F_j^t$, with F_j^t as defined in (2.8a). These forecast renewal probabilities can be computed in a manner akin to the forward algorithm.

Throughout this section we will make the assumption that the distribution of observations at the beginning of an observation subsequence is independent of the length of the subsequence:

$$p\left(x_{t:t+d} \mid \exists r_{t_r=t}^{z_r=j}, {}_{t_{r+1}=t+d}^{z_{r+1}=k}\right) \quad = \quad p\left(x_{t:t+d} \mid \exists r_{t_r=t}^{z_r=j}, {}_{t_{r+1}=t+d'}^{z_{r+1}=k}\right) \quad \text{for} \begin{bmatrix} d=1,\ldots,D \\ d' \geq d \end{bmatrix} \tag{A.33}$$

Note this assumption does *not* hold in the case of the bridging-means model, but it does hold for the AR(1)-SV HMRM.

A.6.1. Computing the Forecast Renewal Probabilities

Applying (2.5) we have:

$$\begin{aligned} F_k^{t,f} &\triangleq p\left(\exists r_{t_r=t+f+1}^{z_r=k}, x_{1:t}\right) \\ &= \sum_{j=1}^{K}\sum_{d=1}^{D\wedge t+f} p\left(\exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, {}_{t_r=t+f+1}^{z_r=k}, x_{1:t}\right) \end{aligned}$$

When $d \leq f$ the renewal $t_r = t+f+1$, $z_r = k$ is independent of $x_{1:t}$ and we have

$$\begin{aligned} & p\left(\exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, {}_{t_r=t+f+1}^{z_r=k}, x_{1:t}\right) \\ &\overset{(CRP)}{=} p\left({}_{t_r=t+f+1}^{z_r=k} \mid \exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, x_{1:t}\right) \times p\left(\exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, x_{1:t}\right) \\ &= \tau_{j,k} f_{\eta_{j,k}}(d) \times F_j^{t,f-d} \end{aligned}$$

When $f+1 \leq d$ the renewal $t_r = t+f+1$, $z_r = k$ is not independent of $x_{t+f+1-d:t}$ and we have:

$$\begin{aligned} & p\left(\exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, {}_{t_r=t+f+1}^{z_r=k}, x_{1:t}\right) \\ &\overset{(CRP)}{=} p\left(x_{t+f+1-d:t} \mid \exists r_{t_{r-1}=t+f+1-d}^{z_{r-1}=j}, {}_{t_r=t+f+1}^{z_r=k}, x_{1:t+f-d}\right) \\ &\quad \times p\left({}_{t_r=t+f+1}^{z_r=k} \mid \exists r_{t_r=t+f+1-d}^{z_r=j}, x_{1:t+f-d}\right) \times p\left(\exists r_{t_r=t+f+1-d}^{z_r=j}, x_{1:t+f-d}\right) \\ &\overset{(2.7)}{\underset{(A.33)}{=}} f_{\varepsilon_{j,k}^d}(x_{t+f+1-d:t}) \times \tau_{j,k} f_{\eta_{j,k}}(d) \times F_j^{t+f-d,0} \end{aligned}$$

So these probabilities can be computed recursively for $f = 1, 2, \ldots$.

$$F_k^{t,f} \leftarrow \sum_{j=1}^{K}\left[\sum_{d=1}^{f} \tau_{j,k} f_{\eta_{j,k}}(d) \times F_j^{t,f-d} + \sum_{d=f+1}^{D\wedge t+f} f_{\varepsilon_{j,k}^d}(x_{t+f+1-d:t}) \times \tau_{j,k} f_{\eta_{j,k}}(d) \times F_j^{t+f-d,0}\right]$$

A.6.2. Forecast Distribution

We first consider the joint distribution $p(x_{t+f}, x_{1:t})$ rather than the forecast distribution $p(x_{t+f}|x_{1:t})$, as the later is easily recovered from the former:

$$\begin{aligned}
&p(x_{t+f}, x_{1:t}) \\
&\quad = \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{d=0}^{(D\wedge t+f)-1} p\left(x_{t+f}, \exists r_{t_r=t+f-d}^{z_r=j}, {}_{t_{r+1}>t+f}^{z_{r+1}=k} \mid x_{1:t}\right) \\
&\quad \overset{(CRP)}{=} \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{d=0}^{(f\wedge D)-1} p\left(x_{t+f}|\exists r_{t_r=t+f-d}^{z_r=j}, {}_{t_{r+1}>t+f}^{z_{r+1}=k}\right) \\
&\qquad\qquad \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^{t,f-d-1} \\
&\qquad + \sum_{d=f}^{(D\wedge t+f)-1} p\left(x_{t+f} \mid x_{t+f-d:t}, \exists r_{t_r=t+f-d}^{z_r=j}, {}_{t_{r+1}>t+f}^{z_{r+1}=k}\right) \\
&\qquad\qquad \times f_{\varepsilon_{j,k}^{d-f+1}}(x_{t+f-d:t}) \times \tau_{j,k}\eta_{j,k}^{>d} \times F_j^{t+f-d-1,0}
\end{aligned} \tag{A.34}$$

Where we have utilized the following notation for the complimentary distribution function of $\eta_{j,k}$:

$$\eta_{j,k}^{>d} \quad \triangleq \quad p\left({}_{t_{r+1}>t+d}^{z_{r+1}=k} \mid {}_{t_r=t}^{z_r=j}\right) \quad = \quad \sum_{\delta=d+1}^{D} f_{\eta_{j,k}}(\delta)$$

The remaining terms needed for the forecast distribution

$$p\left(x_{t+f}|\exists r_{t_r=t+f-d}^{z_r=j}, {}_{t_{r+1}>t+f}^{z_{r+1}=k}\right) \text{ and } p\left(x_{t+f} \mid x_{t+f-d:t}, \exists r_{t_r=t+f-d}^{z_r=j}, {}_{t_{r+1}>t+f}^{z_{r+1}=k}\right) \tag{A.35}$$

are dependent on the choice of emission distributions. In the next subsection we find formulas for these in the case of the AR(1)-SV HMRM.

When $f = 1$, computation of the forecast distribution serves a practical purpose. From the CRP, we can compute the likelihood from our $f = 1$ forecasts as

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

which can be tested against (2.10a). This helps ensure, at least in the $f = 1$ case, our forecast derivations and code are correct. These forecasts can also be used to compute $p(x_{1:t})$, which is used in conjunction with(A.34) to yield the forecast distribution of $p(x_{t+f}|x_{1:t})$.

A.6.3. AR(1)-SV HMRM Specifics

We split the computation of (A.35) into two cases. For the rest of this section we assume $z_r = j$, $z_{r+1} = k$, $t_r = t + f - d$, $t_{r+1} > t + f$ is given.

A.6.3.1. case 1: $p\left(x_{t+f}|\exists r^{z_r=j}_{t_r=t+f-d}, {}^{z_{r+1}=k}_{t_{r+1}>t+f}\right)$

In the case of the AR(1)-SV HMRM, this case is further split into two sub-cases.

sub-case 1a: $d = 0$ This case implies that a renewal is coincident with $t + f$:

$$x_{t+f} \sim \mathcal{N}\left(\mu_{g(j,k)}, \sigma^2_{g(j,k)}\right)$$

sub-case 1b: $d \in \{1, \ldots, f-1\}$ This case means that a renewal occurs after x_t, the last given observation, and an unobserved jump has occurred.

$$x_{t+f}|x_{t+f-d}, v \sim \mathcal{N}\left(\phi_{j,k}^{f} x_u + \mu_{l(j,k)} \sum_{\delta=0}^{f-1} \phi_{l(j,k)}^{\delta}, \frac{\sigma_{l(j,k)}^2}{v} \sum_{\delta=0}^{f-1} \phi_{l(j,k)}^{2\delta}\right)$$

So

$$\frac{x_{t+f} - \phi_{j,k}^{f} x_{t+f-d} - \mu_{l(j,k)} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta}}{\sigma_{l(j,k)} \sqrt{\sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{2\delta}}} | x_{t+f-d} \sim \mathcal{T}(\nu_{j,k})$$

Note that

$$\frac{x-\mu}{\sigma}|\mu, \sigma \sim \mathcal{T}(\nu)$$

$$\Longrightarrow$$

$$p(x|\mu, \sigma) = \frac{1}{\sigma} f_{\mathcal{T}}\left(\frac{x-\mu}{\sigma}; \nu\right)$$

Thus

$$p(x_{t+f}) = \frac{1}{\sigma_{l(j,k)} \sqrt{\sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{2\delta}}} \int_{x_{t+f-d}} f_{\mathcal{T}}\left(x_{t+f} - \phi_{j,k}^{f} x_{t+f-d} - \mu_{l(j,k)} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta}; \nu_{j,k}\right) \times f_{\mathcal{N}}\left(x_{t+f-d}; \mu_{g(j,k)}, \sigma_{g(j,k)}^2\right) dx_{t+f-d}$$

This integral cannot be computed analytically. However it might be possible to evaluate it using numerical integration routines. Fortunately, because we are primarily interested in the moments of this distribution, we needn't compute this integral.

A.6.3.2. case 2: $p\left(x_{t+f} \mid x_{t+f-d:t}, \exists r_{t_r=t+f-d}^{z_r=j,\ z_{r+1}=k,\ t_{r+1}>t+f}\right)$

In this model, for each observation subsequence starting at time u, we have $x_{u+1}|x_u, v = \phi_{j,k}x_u + \mu_{l(j,k)} + \epsilon_{u+1}$, where $\epsilon_{u+1}|v \sim \mathcal{N}(0, \sigma^2_{l(j,k)}/v)$ and v is the (inverse) variance for the corresponding sojourn. This implies that $x_{u+w}|x_u, v = \phi^w_{j,k}x_u + \mu_{l(j,k)}\sum_{\delta=0}^{w-1}\phi^\delta_{l(j,k)} + \sum_{\delta=0}^{w-1}\phi^\delta_{l(j,k)}\epsilon_{u+\delta+1}$. Thus

$$
\begin{aligned}
&x_{t+f}|v, x_{t+f-d:t} \\
&\quad\sim \mathcal{N}\left(\phi^f_{j,k}x_t + \mu_{l(j,k)}\sum_{\delta=0}^{f-1}\phi^\delta_{l(j,k)}, \frac{\sigma^2_{l(j,k)}}{v}\sum_{\delta=0}^{f-1}\phi^{2\delta}_{l(j,k)}\right) \\
&v|x_{t+f-d:t} \\
&\quad\sim \mathcal{G}a\left(\frac{\nu_{j,k}+(d-f)}{2}, \frac{\nu_{j,k}}{2} + \frac{\sum_{\delta=1}^{d-f}(x_{t+f-d+\delta} - \phi_{j,k}x_{t+f-d+\delta-1} - \mu_{l(j,k)})^2}{2\sigma^2_{l(j,k)}}\right)
\end{aligned}
$$

and so

$$
\begin{aligned}
&\frac{x_{t+f} - \phi^f_{j,k}x_t - \mu_{l(j,k)}\sum_{\delta=0}^{f-1}\phi^\delta_{l(j,k)}}{\sigma_{l(j,k)}\sqrt{\sum_{\delta=0}^{f-1}\phi^{2\delta}_{l(j,k)}}} \\
&\qquad\times\sqrt{\frac{\nu_{j,k}+(d-f)}{\nu_{j,k}+\sum_{\delta=1}^{d-f}(x_{t+f-d+\delta} - \phi_{j,k}x_{t+f-d+\delta-1} - \mu_{l(j,k)})^2/\sigma^2_{l(j,k)}}}|x_{t+f-d:t} \\
&\qquad\qquad\sim \quad \mathcal{T}\left(\nu_{j,k}+d-f\right)
\end{aligned}
$$

A.6.3.3. Moments

Case 1a is trivial:

$$\begin{aligned} E\left[x_{t+f} \mid \exists r_{t_r=t+f-d,\, t_{r+1}>t+f}^{z_r=j,\, z_{r+1}=k}\right] &= \mu_{g(j,k)} \\ Var\left[x_{t+f} \mid \exists r_{t_r=t+f-d,\, t_{r+1}>t+f}^{z_r=j,\, z_{r+1}=k}\right] &= \sigma_{g(j,k)}^2 \end{aligned} \tag{A.36}$$

For case 1b, we note that $x_{t+f}|v = \phi_{j,k}^{f-d} x_{t+f-d} + \mu_{l(j,k)} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} + \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} \epsilon_\delta$ where each $\epsilon_\delta|v \sim \mathcal{N}(0, \sigma_{l(j,k)}^2/v)$. This implies that $\sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} \epsilon_\delta|v \sim \mathcal{N}\left(0, \frac{\sigma_{l(j,k)}^2}{v} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{2\delta}\right)$, so $\frac{1}{\sigma_{l(j,k)} \sqrt{\sum_{\delta=0}^{d-1} \phi_{l(j,k)}^2}} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} \epsilon_\delta \sim \mathcal{T}(\nu_{j,k})$, and finally $Var\left[\sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} \epsilon_\delta\right] = \sigma_{l(j,k)}^2 \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^2 \nu_{j,k} / (\nu_{j,k} - 2)$ for $\nu_{j,k} > 2$. It is clear from the previous equation that that we cannot meaningfully forecast the variance when there exists j and k such that $\nu_{j,k} \leq 2$.

$$\begin{aligned} E\left[x_{t+f} \exists r_{t_r=t+f-d,\, t_{r+1}>t+f}^{z_r=j,\, z_{r+1}=k}\right] &= \phi_{j,k}^{f-d} \mu_{g(j,k)} + \mu_{l(j,k)} \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^{\delta} \\ Var\left[x_{t+f} \exists r_{t_r=t+f-d,\, t_{r+1}>t+f}^{z_r=j,\, z_{r+1}=k}\right] &= \phi_{j,k}^{2(f-d)} \sigma_{g(j,k)}^2 + \sigma_{l(j,k)}^2 \sum_{\delta=0}^{d-1} \phi_{l(j,k)}^2 \nu_{j,k} / (\nu_{j,k} - 2) \end{aligned} \tag{A.37}$$

For case 2:

$$
\begin{aligned}
&E\left[x_{t+f}z \mid x_{t+f-d:t}, \exists r^{z_r=j}_{t_r=t+f-d}, {}^{z_{r+1}=k}_{t_{r+1}>t+f}\right] \\
&\qquad = \quad \phi^f_{j,k}x_t + \mu_{l(j,k)}\sum_{\delta=0}^{f-1}\phi^{\delta}_{l(j,k)} \\
&Var\left[x_{t+f} \mid x_{t+f-d:t}, \exists r^{z_r=j}_{t_r=t+f-d}, {}^{z_{r+1}=k}_{t_{r+1}>t+f}\right] \\
&\qquad = \quad \frac{\nu_{j,k}+d-f}{\nu_{j,k}+d-f-2}\times\sigma^2_{l(j,k)}\sum_{\delta=0}^{f-1}\phi^{2\delta}_{l(j,k)} \\
&\qquad\qquad \times\frac{\nu_{j,k}+\sum_{\delta=1}^{d-f}(x_{t+f-d+\delta}-\phi_{j,k}x_{t+f-d+\delta-1}-\mu_{l(j,k)})^2/\sigma^2_{l(j,k)}}{\nu_{j,k}+(d-f)}
\end{aligned}
\tag{A.38}
$$

Nomenclature

ACF	auto-correlation function
CDL	complete data likelihood; for the MM, HMM, this is $p(x_{1:T}, s_{1:T}; \theta)$, for the HMRM it is $p(x_{1:T}, s_{1:R+1}, t_{1:R+1}; \theta)$
CDLL	complete data log likelihood, the logarithm of the CDL
CRP	chain rule of probability, $p(v_{1:N}) = \prod_{n=1}^{N} p(v_n \mid v_{1:n-1})$
DGM	directed graphical model, also called a *Bayesian network*
EDHMM	explicit duration hidden markov model, one of the simplest and most popular HSMMs
EM	expectation maximization
HMM	hidden Markov model
HRM	hidden renewal Model; an HMRM that does not switch between states – equivalently an HMRM with $K = 1$

HSMM	hidden semi-Markov model
iid	independent and indentically distributed
MAP	maximum a posteriori probability (estimate)
MLE	maximum likelihood estimator/estimation
MM	(finite) mixture model
MRP	Markov renewal process
s.t.	such that
SMK	semi-Markov kernel
SMP	semi-Markov process
SV	stochastic volatility
TPM	transition probability matrix
$\mathcal{C}at$	the categorical distribution, $x \sim Cat(\iota_{1:K})$ if $p(x = k) = \iota_k$ for $k = 1 \dots K$
$\mathcal{G}eo$	the geometric distribution, $x \sim \mathcal{G}eo(\pi)$ if $p(x = d) = (1 - \pi)^{d-1}\pi$ for $d = 1, 2, \dots$
$\mathcal{MC}$	$\mathcal{MC}(\iota_{1:K}, \tau_{1:K,1:K})$ is Markov chain with initial distribution $\iota_{1:K}$ and transition probability matrix $\tau_{1:K,1:K}$

$\mathcal{MRP}$	$\mathcal{MRP}(\iota_{1:K}, \tau_{1:K,1:K}, \eta_{1:K,1:K}, \varepsilon_{1:K,1:K})$ is a Markov renewal process with initial distribution $\iota_{1:K}$, transition probability matrix $\tau_{1:K,1:K}$, holding-time distributions $\eta_{1:K,1:K}$, and emission distributions $\varepsilon_{1:K,1:K}$
$\mathcal{NB}$	the negative binomial distribution, $x \sim \mathcal{NB}(r,p)$ if $p(x=d) = \binom{d+r-2}{d-1}(1-p)^{d-1}p^r$ for $d = 1, 2, \ldots$
$\mathcal{N}$	the normal distribution, $x \sim \mathcal{N}(\mu, \sigma^2)$ if $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $x \in \mathbb{R}$
$\mathcal{P}ois$	the Poisson distribution, $x \sim \mathcal{P}ois(\lambda)$ if $p(x=d) = \lambda^{d-1}\frac{e^{-\lambda}}{(d-1)!}$, for $d = 1, 2, \ldots$
$f_{\mathcal{D}}(\cdot;\theta)$	the density or mass function for a distribution $\mathcal{D}$ with parameter(s) θ
emissions	observed values in a model, $x_{1:T}$
factorize	exploiting independencies to express a probability or density as a product (of factors)
holding-time	the length of an observation subsequence, $d_r \triangleq t_{r+1} - t_r$, is the r^{th} holding-time
Markov property	$p(s_{t+1}\|s_{1:t}) = p(s_{t+1}\|s_t)$
renewal	a (superstate,renewal-time) pair, (z_r, t_r) is called the r^{th} renewal

renewal time	the starting time for an observation subsequence, t_r is the starting time for the r^{th} observation subsequence
sequence	a finite, ordered, set; e.g. $s_{1:T} = \{s_1, s_2, \ldots, s_T\}$
sojourn	a pair of adjacent renewals, $(z_r, t_r), (z_{r+1}, t_{r+1})$ is the r^{th} sojourn
state	an element of the unobserved sequence, s_t the state at time t
sub-model	a specification of ε, η, and the corresponding maximizers for Q_ε and Q_η
superstate	an unobserved value, z_r, associated with times $t_r, t_r + 1, \ldots, t_{r+1} - 1$ in an MRP
update formulas	the formulas for the next iteration's parameters, $theta^{(n+1)}$, in the EM algorithm
$\beta_{j,k}^{t,d}$	backward sojourn probability, $\beta_{j,k}^{t,d} \triangleq p(x_{t:T},\, z_{r+1} = k,\, t_{r+1} = t + d \mid z_r = k,\, t_r = t)$
$\hat{s}_{1:T}$	the maximum posterior state sequence, $\hat{s}_{1:T} \triangleq \arg\max_{s_{1:T}} p(s_{1:T} \mid x_{1:T})$
$\hat{z}_{1:R+1}, \hat{t}_{1:R+1}$	the maximum posterior renewal sequence, $\hat{z}_{1:R+1}, \hat{t}_{1:R+1} \triangleq \arg\max_{z_{1:R+1}, t_{1:R+1}} p(z_{1:R+1}, t_{1:R+1} \mid x_{1:T})$
$\mathscr{M}_k^t$	in an HMM – the maximum posterior probability of all state sequences ending with $s_t = k$, $\max_{s_{1:t-1}} p(s_t = k, s_{1:t-1} \mid x_{1:T})$; in

	an HMRM – the maximum posterior probability of all state sequences ending with $(z_r = k, t_r = t)$, $\max_{\substack{r=1,\ldots,t \\ z_{1:r-1}, t_{1:r-1}}} p(z_r = k, t_r = t, z_{1:r-1}, t_{1:r-1} \vert x_{1:T})$
$\mathscr{S}_k^t$	the value of s_t in the maximum posterior probability state sequence ending with $s_{t+1} = k$
$\mathscr{T}_k^t$	the value of t_{r-1} in the maximum posterior probability renewal sequence ending with $(z_r = k, t_r = t)$
$\mathscr{Z}_k^t$	the value of z_{r-1} in the maximum posterior probability renewal sequence ending with $(z_r = k, t_r = t)$
$\phi_{j,k}^{t,d}$	forward sojourn probability, $\phi_{j,k}^{t,d} \triangleq p(\exists r \, s.t. \, z_r = j, \, z_{r+1} = k, \, t_r = t - 1 + d, \, t_{r+1} = t, \, x_{1:t})$
A_k^t	posterior state probability, $A_k^t \triangleq p(s_t = k \vert x_{1:T}; \theta)$
F_k^t	in an HMM – the forward probability $F_k^t \triangleq p(s_t = k, x_{1:t-1})$; in an HMRM – the forward renewal probability $F_k^t \triangleq p(\exists r \, s.t. \, z_r = k, t_r = t + 1, x_{1:t})$
$S_{j,k}^{t,d}$	posterior sojourn probability, $S_{j,k}^{t,d} \triangleq p(\exists r \, s.t. \, z_r = j, \, z_{r+1} = k, \, t_r = t, \, t_{r+1} = t + d \vert x_{1:T})$
$*$	element-wise product, e.g. $y_{t:u} * z_{t:u} = (y_t z_t, y_{t+1}, z_{t+1}, \ldots, y_u z_u)$
$\perp\!\!\!\perp$	conditional independence; $A \perp\!\!\!\perp B \vert C \iff p(A, B \vert C) = p(A \vert C) p(B \vert C) \iff p(A \vert B, C) = p(A \vert C) \iff p(B \vert A, C) =$

	$p(B\|C)$
$\leftarrow$	denotes that a variable is being set to some value, typically used in the context of an algorithm
$\hat{}$	denotes a maximizer, e.g. $\hat{\theta} = \arg\max_{\theta} p(x_{1:T}; \theta)$
$\mathbb{I}$	the indicator function, $\mathbb{I}\{expr\} = 1$ if $expr$ is true, and $\mathbb{I}\{expr\} = 0$ if $expr$ is false
$\mathbb{N}$	the natural numbers, $1, 2, 3, ...$
$\sim$	is distributed as, e.g. $x \sim \mathcal{N}(\mu, \sigma^2)$ if $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $x \in \mathbb{R}$
$\vee$	the maximum of two numbers, $x \vee y \triangleq \max(x, y)$
$\wedge$	the minimum of two numbers, $x \wedge y \triangleq \min(x, y)$
$pa(v)$	in a DGM this is the set of all nodes that are parents of v, that is all nodes u such that $u \rightarrow v$
$sd(\cdot)$	the sample standard deviation
$v_{\setminus t}$	specifies all but the t^{th} element in a sequence, $v_{1:T} \setminus \{v_t\}$
:	Specifies a contiguous sequence of values, e.g. $v_{t:u} \equiv (v_t, v_{t+1}, \ldots, v_u)$. Two subscripts can be used to express a matrix, e.g. $\tau_{1:K,1:K}$; $\tau_{k,1:K}$ represents the k^{th} row and $\tau_{k,1:K}$ the k^{th} column. If $t > u$, then the convention is that $v_{t:u} = \varnothing$.

$\text{diag}[v]$	the diagonal matrix with the vector v as the diagonal
α	the state distribution in an MM, $p(s_t = k) = \alpha_k$ for every $t = 1 \dots T$
η	the holding-time distribution $\eta^d_{j,k} \triangleq p(t_{r+1} - t_r = d \mid z_r = j, z_{r+1} = k)$
ι	the initial distribution; in an HMM, $\iota_k = p(s_1 = k)$ for $k = 1 \dots K$, in an HMRM $\iota_k = p(z_1 = k)$
τ	the transition distribution or transition probability matrix; $\tau_{j,k} \triangleq p(s_{t+1} = k \mid s_t = j)$ for an HMM, $\tau_{j,k} \triangleq p(z_{r+1} = k \mid z_r = j)$ for an HMRM
θ	all the model's parameters; $\theta \triangleq \{\iota, \epsilon\}$ for an MM, $\theta \triangleq \{\iota, \epsilon, \tau\}$ for an HMM, $\theta \triangleq \{\iota, \epsilon, \tau, \eta\}$ for an HMRM
ε	the emmission distribution $\varepsilon_k(x_t) \triangleq p(x_t \mid s_t = k)$ for an MM and HMM, $\varepsilon^d_{j,k}(x_{t:(t+d-1 \wedge T)}) \triangleq p(x_{t:(t+d-1 \wedge T)} \mid \exists r \, s.t \, z_r = j, z_{r+1} = k, t_r = t, t_{r+1} = t + d)$ for an HMRM
K	the number of hidden (super)states in a model; in an MM or HMM $s_t \in 1, \dots, K$ for each t, in an HMRM $z_r \in 1, \dots, K$ for each r
d_r	the r^{th} holding-time, $d_r \triangleq t_{r+1} - t_r$

R	the number of renewals occurring (strictly) before time $T+1$ in an HMRM
s_t	the state of the model corresponding to observation x_t
T	the number of observations
t_r	the r^{th} renewal-time, equivalently the time of the beginning of the r^{th} observation subsequence $x_{t_r:t_{r+1}-1}$
x_t	the observation at time t
z_r	the superstate coinciding with the r^{th} observation subsequence

Bibliography

[1] Vlad Stefan Barbu and Nikolaos Limnios. *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis*, volume 191. Springer Science & Business Media, 2009.

[2] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, pages 164–171, 1970.

[3] Christopher M Bishop et al. *Pattern recognition and machine learning*. Number 4. springer New York, 2006.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Jan Bulla. Application of hidden markov and hidden semi-markov models to financial time series. 2006.

[6] Jan Bulla and Ingo Bulla. Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics & Data Analysis*, 51(4):2192–2209, 2006.

[7] Jan Bulla, Ingo Bulla, and Oleg Nenadić. hsmm-An R package for analyzing hidden semi-Markov models. *Computational Statistics & Data Analysis*, 54(3): 611–619, 2010.

[8] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[9] E. Çınlar. *Introduction to stochastic processes*. Prentice-Hall, 1975. ISBN 9780134980898. URL `https://books.google.com/books?id=UNZQAAAAMAAJ`.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[11] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8):1–18, 2011. URL `http://www.jstatsoft.org/v40/i08/`.

[12] Jack D Ferguson. Variable duration models for speech. In *Proceedings of the Symposium on the Application of HMMs to Text and Speech*, pages 143–179, 1980.

[13] G David Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3): 268–278, 1973.

[14] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[15] Yann Guédon. Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3):604–639, 2003.

[16] James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1):307–333, 1994.

[17] R.A. Howard. *Dynamic probabilistic systems: volume ii: semi-markov and decision processes*. Series in decision and control. John Wiley And Sons, Incorporated, 1971. URL `https://books.google.com/books?id=KXSRoAEACAAJ`.

[18] Jacques Janssen and Raimondo Manca. *Applied semi-Markov processes*. Springer Science & Business Media, 2006.

[19] Edward P.C. Kao. *An Introduction to Stochastic Processes*. Duxbury Press, Boston, MA, 1996.

[20] R Langrock and W Zucchini. Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1):715–724, 2011.

[21] Chuanhai Liu and Donald B Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):19–39, 1995.

[22] Sonia Malefaki, Samis Trevezas, and Nikolaos Limnios. An EM and a stochastic version of the EM algorithm for nonparametric Hidden semi-Markov models. *Communications in Statistics–Simulation and Computation®*, 39(2):240–261, 2010.

[23] Kevin P Murphy. Hidden semi-Markov models (hsmms). `http://www.cs.ubc.ca/~murphyk/Papers/segment.pdf`, 2002. URL `http://www.cs.ubc.ca/~murphyk/Papers/segment.pdf`.

[24] Jared O'Connell, Søren Højsgaard, et al. Hidden semi markov models for mul-

tiple observation sequences: The mhsmm package for R. *Journal of Statistical Software*, 39(4), 2011.

[25] Ronald Pyke. Markov renewal processes: definitions and preliminary properties. *The Annals of Mathematical Statistics*, pages 1231–1242, 1961.

[26] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

[27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[28] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

[30] Jeffrey Seth Rosenthal. *A first look at rigorous probability theory*. World Scientific, 2006.

[31] Tobias Rydén. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

[32] Tobias Rydén, Timo Teräsvirta, Stefan Åsbrink, et al. Stylized facts of daily return series and the hidden Markov model. *Journal of applied econometrics*, 13 (3):217–244, 1998.

[33] Conrad Sanderson. Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. Technical report, NICTA, September 2010.

[34] Till Tantau. *The TikZ and PGF Packages*. URL `http://sourceforge.net/projects/pgf/`.

[35] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

[36] Shun-Zheng Yu. Hidden semi-Markov models. *Artificial Intelligence*, 174(2): 215–243, 2010.

[37] Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series: an introduction using R*. CRC Press, 2009.