Nonparametric Bayesian Models for Markov Jump Processes

by

Ardavan Saeedi

B.Sc., Sharif University of Technology, 2007

M.Sc., Sharif University of Technology, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES
(Statistics)

The University Of British Columbia
(Vancouver)

August 2012

Abstract

Markov jump processes (MJPs) have been used as models in various fields such as disease progression, phylogenetic trees, and communication networks. The main motivation behind this thesis is the application of MJPs to data modeled as having complex latent structure. In this thesis we propose a nonparametric prior, the gamma-exponential process (GEP), over MJPs. Nonparametric Bayesian models have recently attracted much attention in the statistics community, due to their flexibility, adaptability, and usefulness in analyzing complex real world datasets. The GEP is a prior over infinite rate matrices which characterize an MJP; this prior can be used in Bayesian models where an MJP is imposed on the data but the number of states of the MJP is unknown in advance. We show that the GEP model we propose has some attractive properties such as conjugacy and simple closed-form predictive distributions. We also introduce the hierarchical version of the GEP model; sharing statistical strength can be considered as the main motivation behind the hierarchical model. We show that our hierarchical model admits efficient inference algorithms. We introduce two inference algorithms: 1) a "basic" particle Markov chain Monte Carlo (PMCMC) algorithm which is an MCMC algorithm with sequences proposed by a sequential Monte Carlo (SMC) algorithm; 2) a modified version of this PMCPC algorithm with an "improved" SMC proposal. Finally, we demonstrate the algorithms on the problems of estimating disease progression in multiple sclerosis and RNA evolutionary modeling. In both domains, we found that our model outperformed the standard rate matrix estimation approach.

Table of Contents

List of Tables

List of Figures

Glossary

Acronyms

CRF	Chinese restaurant franchise
CRP	Chinese restaurant process
DDP	dependent Dirichlet process
DP	Dirichlet process
EDSS	expanded disability status scale
GEP	gamma exponential process
HDP	hierarchical Dirichlet process
HGEP	hierarchical gamma exponential process
HMM	hidden Markov model
IHMM	infinite hidden Markov model
MCMC	Markov chain Monte Carlo
MGP	Moran gamma process
MH	Metropolis - Hastings
MJP	Markov jump process
MS	multiple sclerosis

PIMH	particle independent Metropolis - Hastings
PMCMC	particle Markov chain Monte Carlo
SMC	sequential Monte Carlo

Acknowledgments

It is difficult to overstate my gratitude to my supervisors Drs. Alexandre Bouchard-Côté and John Petkau. Alex's enthusiasm, his patience, his guidance, and his support have helped me greatly and this thesis would not have been possible without him. He is an inspiration and I am very honored to call him my teacher. John's great advice, support, friendship, especially his great efforts to explain things clearly and simply has been invaluable on both an academic and a personal level, for which I am extremely grateful.

Many thanks go to the numerous people who have been my teachers and mentors along my academic journey so far. I would especially like to express my gratitude to Dr. Alireza Haji (B.Sc. supervisor) and Dr. Kourosh Eshghi (M.Sc. supervisor) at Sharif University of Technology. I am very grateful for their generous support, advise, and help with various projects.

I would like to thank my student colleagues and friends for great discussions and for providing a stimulating and fun environment in which to learn and grow. Special thanks go to Chao Xiong, Hongyang Zhang, Liangliang Wang, Hossein Bashashati, Reza Skandari, Shahram Pourazadi, Sajjad Fayazi, Abbas Javaherian, and Ehsan Esfahanian. The secretaries at the Statistics department deserve my gratitude for providing assistance and administrative support, with special mention to Peggy, Andrea, and Elaine.

Last but not least, I am forever grateful to my parents for their love and support, especially my mother who has always been there for me and helped me grow as a person. To her I dedicate this thesis.

Chapter 1

INTRODUCTION

Inside every non-Bayesian, there is a Bayesian struggling to get out.
— Dennis V. Lindley

Markov jump processes (MJPs), continuous-time Markov processes with piecewise constant paths, have been used as models in many different fields including disease progression [5], phylogenetic trees [41], and communication networks [44]. The main motivation behind this thesis is the application of MJPs to data modeled as having complex latent structure.

As a concrete example, disease progression in multiple sclerosis (MS) can be considered as an MJP [31]. MS is one of the most complex disabling diseases affecting the central nervous system. A typical evolution of the disease begins with periods of relapses and remissions. At this stage of the disease, patients recover from the neurological symptoms that appear during a relapse period in the following remission period. However, most of the patients eventually move to another stage of the disease in which disability accumulates. The course of the disease can be assessed using disability scales such as the **expanded disability status scale (EDSS)**, an ordinal score ranging from 0 to 10 in half point steps. Since the patient's state (measured by the EDSS score) usually remains stable for a relatively long period of time, an MJP model is a reasonable choice for modeling the EDSS data.

In MS studies, EDSS scores are analyzed with the aim of describing or predicting the disease progression. The EDSS data available are partially observed data; measurements are only taken at certain time points. It is not reasonable to assume the EDSS scores, collected in time, are independent of each other; thus, they can be thought of as **sequential data** *for each individual patient*. Several authors have proposed Markov models for analyzing EDSS scores; see, for example [1] and [32]. However, due to the complex nature of the disease and also the partial observability of the data, more sophisticated models may be required. A **hidden Markov model (HMM)** is proposed in [31] as a possible model for disease progression. In an HMM, the states are not directly observed, but are reflected by the observations (e.g., EDSS scores). That is, we only observe a sequence of outputs which are dependent on an underlying Markov process. The currently available methods for carrying out inference in an HMM with an underlying MJP assume the number of the hidden states is either known (see for example [43] or [5]) or can be estimated using a model selection method (e.g., AIC, BIC, or penalized minimum-distance [30] methods). To successfully apply these methods to a real-world problem, we need a domain expert or a model selection method to specify the number of hidden states.

In complex settings, such as disease progression in MS, it is seldom the case that one has a defensible parametric model. **Nonparametric Bayesian** models, Bayesian models with nonparametric priors, have recently attracted much attention in the statistics community, due to their flexibility, adaptability, usefulness in analyzing complex real world datasets, and their ability to sidestep the model selection. Nonparametric Bayesian models can automatically infer an adequate model complexity from the data, without needing explicit Bayesian model comparison. In addition, while being highly adaptable, these models maintain some advantages of a fully model-based probabilistic framework. Exact inference in these models is usually impossible; thus, statistical inference typically is done using **Markov chain Monte Carlo (MCMC)**, a general approach used in approximate inference algorithms for Bayesian models.

In this thesis we propose a nonparametric prior, the **gamma exponential process (GEP)**, over MJPs. In particular, we propose a prior over infinite rate matrices which characterize an MJP. These priors can be used in Bayesian models where an MJP is imposed on the data but the number of states of the MJP is unknown in advance. Thus, if we consider a Bayesian MJP model for analyzing the MS data then these priors are a reasonable choice. We will briefly provide some background on MJPs and their rate matrices in Chapter 2. We will also review some basic nonparametric priors such as the **gamma process** and the **Dirichlet process (DP)** in this chapter.

There are a number of nonparametric Bayesian models available for sequential data. One class of these models are developed for the discrete-time framework; examples of these models are the **infinite HMM (iHMM)** [3] and the **sticky hierarchical Dirichlet process-HMM** [13]. Another class of nonparametric Bayesian models are developed for continuous-time transient processes. Transient processes are processes where the effect of an observation at time point t_1, on a prediction at time point t_2 should decrease as $|t_1 - t_2|$ increases. Examples of these models are the **order-based dependent Dirichlet process (DDP)** [17] and the **stick-breaking autoregressive process** [18]. We will review these and related models in Chapter 3. Our GEP model fills the gap for nonparametric priors over continuous-time recurrent (i.e., non-transient) processes.

The GEP model that we propose has some attractive properties such as conjugacy and simple closed-form predictive distributions. The simple predictive distributions that arise with the DP is one of the factors behind its widespread adoption in nonparametric Bayesian statistics; moreover, a simple predictive distribution can simplify the process of making statistical inference. We discuss the properties of GEPs in Chapter 4.

We extend the basic model of Chapter 4 to a hierarchical version in Chapter 5. Sharing statistical strength can be considered as the main motivation behind a hierarchical model. Informally, we want the rows of the rate matrix to share information on what states are frequently visited. We introduce the **hierarchical gamma exponential process (HGEP)** and the motivation behind it in Chapter 5.

After building an appropriate model (i.e., GEP and HGEP) for a complex latent structure, we need to infer this structure from the observed data. Typically we are not able to observe the whole sequence of states; that is, we only have partially observed sequences of data. In Chapter 6, we show that the HGEP model admits efficient inference algorithms. We introduce two inference algorithms: 1) a **particle Markov chain Monte Carlo (PMCMC)** algorithm which is an MCMC algorithm with sequences proposed by a **sequential Monte Carlo (SMC)** algorithm; 2) a modified version of this PMCPC algorithm with an “improved” SMC proposal.

In order to link the observed data to the latent structure we need a likelihood model; we assume a simple multinomial-Dirac likelihood model. Using this likelihood model in several experiments, we observe that despite the simplicity of this likelihood model the results are reasonable. The multinomial-Dirac likelihood model is described in Chapter 7.

We apply the model to three different datasets in Chapter 8: 1) a synthetic dataset generated from a random rate matrix; 2) an MS dataset obtained from a phase III clinical trial of a drug; and 3) an RNA dataset containing aligned RNA of species from the three domains of life. We evaluate our model on these datasets based on a held-out task. For the held-out task we construct three new datasets where each observation is held-out with 10% probability; then we reconstruct the observations at the held-out times and measure the mean error. The results show that our model outperforms the current available methods. We also apply the model to estimating disease progression in MS. We conclude the thesis in Chapter 9 with some final remarks and directions for future research.

Chapter 2

BASIC NOTIONS

In this chapter, we introduce some basic notions and theories used in the following chapters of the thesis. We start by defining a Markov jump process (MJP); next, we present the definitions and properties of gamma and Dirichlet processes as the important elements of our model.

2.1 Markov Jump Process (MJP)

An MJP can be described as a continuous time process with the Markov property on a discrete state space; a more formal definition is as follows:

Definition 1. *An MJP is a stochastic process $\{S_t, t \in \mathscr{T}\}$ with a discrete state space Ω and a continuous index set $\mathscr{T} = \mathbb{R}^+$ that satisfies the **Markov property**:*

$$p(S_{t+s} = j|S_u; u \leq t) = p(S_{t+s} = j|S_t). \tag{2.1}$$

If, moreover, the right hand side of Equation 2.1 *only depends on the time increment s, but not on t, then the* MJP *is called homogeneous.*

Here, we only consider homogeneous MJPs; thus, the term MJP will always refer to a homogeneous MJP.

An MJP is often defined through its instantaneous transition rate matrix Q:

$$Q = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} & \cdots \\ q_{2,1} & q_{2,2} & q_{2,3} & \cdots \\ q_{3,1} & q_{3,2} & q_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The off-diagonal element $q_{i,j}$ corresponds to the instantaneous transition rate from state i to state j:

$$q_{i,j} = \lim_{h \downarrow 0} \frac{P(S_{t+h} = j | S_t = i)}{h};$$

hence, they are non-negative. As a consequence of the definition of $q_{i,j}$, the waiting rate at state i denoted by $q_{i,i}$ is determined from $q_{i,i} = -\sum_{j \neq i} q_{i,j}$. The convention is to place $q_{i,i}$ in the diagonals of Q.

One particular probability of interest in an MJP is the transition probability from one state to another given a time interval t. The transition probability matrix $P(t) = (p_{i,j}(t))$ where:

$$p_{i,j}(t) = p(S_t = j | S_0 = i)$$

is calculated by matrix exponentiation; that is, $P(t) = e^{Qt}$. Another probability of interest is the initial probability distribution π where $\pi_i = p(S_0 = i)$.

Given the transition rate matrix Q, the behavior of a sample path from an MJP is specified by the following theorem [27]:

Theorem 2. *Assume an MJP process with transition rate matrix $Q = (q_{i,j})$, is in state i at time point t. The sample path behavior of the MJP can be described as follows:*

1. *The waiting time in the current state i is distributed according to $Exp(-q_{i,i})$.*

2. *Given its past,if a transition occurs at a particular time, the probability that this transition is from the current state i to state j is given by $M_{i,j} = q_{i,j}/|q_{i,i}|$,*

Note that in the second statement of Theorem 2, M is the transition probability matrix for a Markov chain without self-transitions (i.e., $M_{i,i} = 0, \ \ \forall i \in \Omega$). This Markov chain is called the **embedded Markov chain** of the MJP.

A sample from an MJP can be described as a list of pairs $X = (\theta_n, J_n)_{n=1}^{N}$, where J_n is the waiting time at jump *n-1* ($J_1 = 0$) and θ_n is the state at jump n . We assume the initial state of the sequence (i.e., θ_1 or S_0) is distributed according to the initial distribution of the MJP. For *finite* sequences with length N, we can assume a special absorbing state for the end of a *finite* sequence θ_{end}. We would then condition on $(\theta_{N+1} = \theta_{\text{end}})$ and $(\theta_n \neq \theta_{\text{end}}, n \in \{1, \ldots, N\})$, and set the total rate for the row corresponding to θ_{end} to zero. In the following, we only consider distributions over infinite sequences.

Based on Theorem 2, we can use Doob-Gillespie algorithm [8] to sample from an MJP. Denoting the jump times by $T_1, T_2, \ldots$, this algorithm can be described as follows:

1. For $T_0 = 0$ sample θ_1 from π. Set $J_1 = 0$, $S_0 = \theta_1$, and $n = 1$.
2. Sample J_{n+1} from $Exp(-q_{\theta_n, \theta_n})$.
3. Update the jump time: $T_n = T_{n-1} + J_{n+1}$.
4. For $T_{n-1} \leq t < T_n$ set $S_t = \theta_n$.
5. Sample θ_{n+1} from the row corresponding to the θ_n row of the probability transition matrix M. Set $n = n + 1$ and go to step 2.

A sample path from an MJP is sketched in Figure 2.1.

2.2 Gamma Process

The nonparametric prior that we propose is based on the **gamma process**, a stochastic process with independent, identically gamma distributed increments. Recall that the gamma distribution with shape parameter $a > 0$ and rate parameter $b > 0$ has a density given by:

$$\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} I_{(0,\infty)}(x),$$

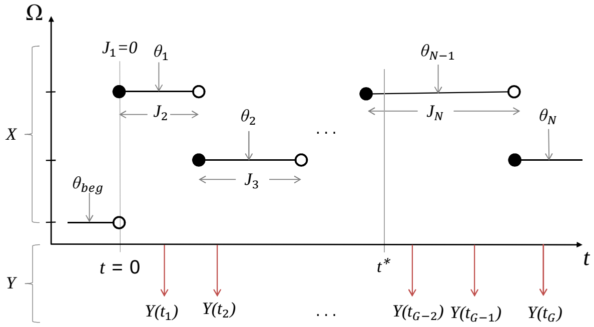


Figure 2.1: An illustration of notation for samples from MJPs. We assume the state space (Ω) is countable. The notation for the observations $Y(t_1),\dots,Y(t_G)$ is described in Chapter 6.

where $\Gamma(a) = \int_{z=0}^{\infty} z^{a-1}e^{-z}dz$ is the gamma function for $a > 0$, and $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise.

We use the terminology in [25] for defining the gamma process; thus we use the term **Moran gamma process (MGP)** and denote a gamma process by MGP. The MGP can be defined through three parameters:

1. concentration parameter also called the shape parameter $\alpha_0 > 0$
2. base probability distribution, $P_0 : \mathscr{F}_\Omega \to [0,1]$, where $\mathscr{F}_\Omega$ is the $\sigma - algebra$ over the sample space Ω
3. rate parameter $\beta_0 > 0$

We can combine the first two parameters into one new parameter called the **base measure parameter** H_0 where $H_0 = \alpha_0 P_0$ is a measure on the space $(\Omega, \mathscr{F}_\Omega)$. By this parameterization we can characterize an MGP just by two parameters H_0 and β_0.

Informally a sample from a gamma process is an atomic measure with random weighted point masses. The point masses $\omega \in \Omega$ and their corresponding weights $x > 0$ are points in a product space $\Omega \otimes [0,\infty)$. A formal definition of the MGP can be provided based on Poisson processes.

Consider a Poisson process with intensity $\lambda(d\omega dx) = H_0(d\omega)x^{-1}e^{-\beta_0 x}dx$ over the product space $\Omega \otimes [0,\infty)$. A realization of this Poisson process is an infinite set of atoms, $\{\omega_i, x_i\}_{i=1}^{\infty}$ where $\omega_i \in \Omega$ can be interpreted as the point mass of an atomic measure and x_i as its corresponding weight. A realization of an $MGP(H_0, \beta_0)$ is then defined as:

$$\mu = \sum_{i=1}^{\infty} x_i \delta_{\omega_i} \sim MGP(H_0, \beta_0). \tag{2.2}$$

Note that in contrast to the usual definition of the one-dimensional Poisson process where the process is defined over a line (usually time), here we define the process over a product space $\Omega \otimes [0,\infty)$. In such a spatial Poisson process the *counts of the number of events* inside each of a number of non-overlapping finite sub-regions of the space are Poisson distributed and independent of each other.

It can be shown [25] that the *total mass* $\mu(A) = \sum_{i=1}^{\infty} x_i \mathbf{1}(\omega_i \in A)$ of any measurable subset $A \subset \Omega$ is gamma distributed with shape parameter $H_0(A)$ and rate parameter β_0. In other words, a gamma process takes positive values independently distributed as gamma at each event generated by a Poisson process. Hence, we have an equivalent definition for the MGP based on its marginals.

Recall that by the **Kolmogorov consistency theorem**, in order to guarantee the existence of a stochastic process on a probability space $(\Omega', \mathscr{F}_{\Omega'})$, it is enough to provide a consistent definition of the marginals of this stochastic process. As already mentioned, in the case of an MGP, the marginals are gamma distributions:

Definition 3 (MGP)**.** *Let H_0, β_0 be of the types listed above. We say that $\mu : \mathscr{F}_{\Omega'} \to (\mathscr{F}_{\Omega} \to [0,\infty))$ is distributed according to the MGP distribution, denoted by $\mu \sim$ $\mathrm{MGP}(H_0, \beta_0)$, if for all measurable partitions of Ω, $(A_1, \ldots, A_K)$, we have:* [1]

$$(\mu(A_1), \mu(A_2), \ldots, \mu(A_K)) \sim \mathrm{Gamma}(H_0(A_1), \beta_0) \times \cdots \times \mathrm{Gamma}(H_0(A_k), \beta_0).$$

[1] We use the rate parameterization for the gamma density throughout.

2.2.1 MGP and Dirichlet process (DP)

There is a connection between the MGP and the DP which will be important in the following chapters.

As mentioned in Section 2.2 for any measurable subset $A \subset \Omega$ we have:

$$\mu(A) = \sum_{i=1}^{\infty} x_i 1(\omega_i \in A) \sim \text{Gamma}(H_0(A), \beta_0). \tag{2.3}$$

Dividing a sample from an MGP, μ by $\mu(\Omega)$ we get a random probability measure which is in fact a sample from a DP with concentration parameter $H_0(\Omega)$ and the base probability measure $H_0/H_0(\Omega)$. We will explain these parameters in the following section.

2.3 DP Definition and Properties

In the previous section, we mentioned that by normalizing the MGP we obtain a new stochastic process called the DP. In this section we describe the DP, some of its important properties, and different ways to represent it.

2.3.1 Definition

Intuitively the DP is a generalization of the **Dirichlet distribution** to an infinite number of dimensions. In fact, the Dirichlet distribution is itself the multivariate generalization of the beta distribution.

Recall that the beta distribution is a continuous probability distribution defined on the interval $(0,1)$ with two parameters $\alpha > 0$ and $\beta > 0$. This distribution is typically used in modelling proportions. The probability density function of the beta distribution is:

$$Beta(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x)$$

By extending a beta distribution to more than two dimensions we obtain a Dirichlet distribution. A K-dimensional Dirichlet distribution with positive parameters

$\alpha_1, \cdots, \alpha_K$ is defined by the following density function:

$$Dir(x_1, \cdots, x_{K-1}; \alpha_1, \cdots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

for all $x_1, \cdots, x_{K-1} > 0$ satisfying $\sum_{k=1}^{K-1} x_k < 1$ where $x_K = 1 - (\sum_{k=1}^{K-1} x_k)$.

This distribution is often used in the Bayesian framework as a prior for the multinomial distribution. That is due to the fact that the Dirichlet distribution is a conjugate prior over the multinomial distribution. That is, if a random variable has a multinomial distribution with a Dirichlet prior over the parameters of the distribution, then the posterior over the parameters is also Dirichlet. Denoting N independent observations from a multinomial distribution by $Y = (y_1, \cdots, y_N)$, we have

$$\begin{aligned} \mathbf{P} = (p_1, \cdots, p_K) &\sim Dir(\alpha_1, \cdots, \alpha_K) \\ Y|\mathbf{P} &\sim Mult(p_1, \cdots, p_K) \\ \Rightarrow \mathbf{P}|Y &\sim Dir(\alpha_1 + n_1, \cdots, \alpha_K + n_K) \end{aligned}$$

where n_k is the number of y_is in category k.

The DP is the result of extending the Dirichlet distribution to infinite dimensions. A DP, denoted by G, can be thought of as a probability distribution over probability measures (i.e., non-negative functions that integrate to one) and is defined by two parameters:

1. Base measure (G_0): a probability distribution $G_0 : \mathscr{F}_\Omega \rightarrow [0,1]$ which is the mean of the DP; that is, for any measurable set $A \subset \Omega$ we have $E[G(A)] = G_0(A)$. Here, $\mathscr{F}_\Omega$ is the $\sigma - algebra$ over the sample space Ω.

2. Concentration parameter (α_0): a positive real number that controls the variability around the mean G_0.

A realization from a $DP(\alpha_0, G_0)$ is a random distribution with values drawn from G_0. In other words, the support of the sample is the same as the base measure.

Similar to the definition of the MGP, by the Kolmogrov consistency theorem, we can define a DP on a probability space $(\Omega', \mathscr{F}_{\Omega'})$ by providing a consistent definition of its marginals. Following this approach we have the following definition for the DP:

Definition 4. $G : \mathscr{F}_{\Omega'} \to (\mathscr{F}_{\Omega} \to [0,1])$ *is DP distributed with base measure* G_0 *and concentration parameter* α_0*, denoted by* $G \sim DP(\alpha_0, G_0)$*, if for any finite measurable partition* $(A_1, \cdots, A_K)$ *of* Ω *we have:*

$$(G(A_1), G(A_2), \cdots, G(A_K)) \sim Dir(\alpha_0 G_0(A_1), \alpha_0 G_0(A_2), \cdots, \alpha_0 G_0(A_K)). \quad (2.4)$$

Note that, similarly to the gamma process, Definition 4 suggests that the two parameters α_0 and G_0 can be treated as a single parameter $H_0 = \alpha_0 G_0$; this will result in a notationally convenient parameterization $G \sim DP(H_0)$.

2.3.2 Conjugacy and posterior distribution

A conjugacy property holds for the DP. Let $G \sim DP(\alpha_0, G_0)$ and y be a single realization from G, $y \sim G$. Note that since G is a measure valued random variable, we can sample random variables from a realization of G. Roughly speaking G can be thought of as an infinite-dimensional multinomial distribution. Due to the Dirichlet-multinomial conjugacy, for a Dirichlet distribution induced by a fixed partition $(A_1, A_2, \cdots, A_K)$ of Ω (i.e., the support of G_0) and $y \in A_k$, we have:

$$\begin{aligned} G|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ y|G &\sim G \\ (G(A_1), G(A_2), \cdots, G(A_K))|y &\sim Dir(\alpha_0 G_0(A_1), \alpha_0 G_0(A_2), \cdots, \alpha_0 G_0(A_k) + 1, \cdots, \alpha_0 G_0(A_K)) \end{aligned}$$

That is, the single (data) point in the state space y only affects the single element of the partition A_k to which it belongs. Figure 2.2 illustrates an example where the point belongs to the element A_2 of the partition. Generally, for a sequence of N independent realizations $y_1, \cdots, y_N$ from G we have:

$$(G(A_1), G(A_2), \cdots, G(A_K))|y_1, \cdots, y_N \sim Dir(\alpha_0 G_0(A_1) + n_1, \alpha_0 G_0(A_2) + n_2, \cdots, \alpha_0 G_0(A_K) + n_K)$$

where n_k is the number of points in the kth element of the partition: $n_k = |\{i : y_i \in A_k\}|$. Thus, the posterior distribution over G must be a DP as well, since the above

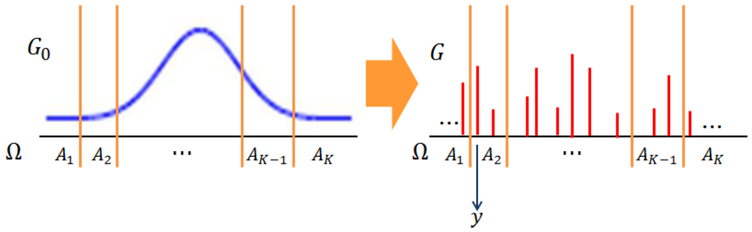


Figure 2.2: A realization G of a DP with base measure G_0 and support Ω. A partition $(A_1, A_2, \cdots, A_K)$ and a point y sampled from G are also illustrated.

is true for any finite partition. Formally, we have the following proposition for the posterior of a DP measure given a set of points $y_1, \cdots, y_N$.

Proposition 5. *Suppose $G \sim DP(\alpha_0, G_0)$. Given N independent observations $y_i \sim G$ the posterior distribution also follows a DP:*

$$G|y_1, \cdots, y_N, \alpha_0, G_0 \sim DP(\alpha_0 + N, \frac{\alpha_0 G_0 + \sum_{i=1}^N \delta_{y_i}}{\alpha_0 + N}).$$

where δ_{y_i} is a point mass located at y_i.

The proof provided in [12] follows directly from the Dirichlet-multinomial conjugacy.

2.3.3 Stick-breaking construction

Definition 4 is not constructive; that is, it does not provide a mechanism for sampling from a DP. Here we provide an alternative constructive definition for the DP.

Definition 6 (Stick-breaking construction [40])**.** *Let $\beta_c \sim Beta(1, \alpha_0)$ be independent. Define the stick lengths $\pi = \{\pi_c\}_{c=1}^{\infty}$, as follows:*

$$\pi_1 = \beta_1$$
$$\pi_c = \beta_c \prod_{1 \le c' < c} (1 - \beta_{c'}) = \beta_c (1 - \sum_{c'=1}^{c} \pi_{c'}) \quad c > 1.$$

We denote this construction by $\pi \sim GEM(\alpha_0)$ [2].

[2] GEM named after Griffiths, Engen and McCloskey; the abbreviation was first introduced in [36].

In words, we start from a stick with length 1; at step c a proportion β_c is broken off and used as the stick length. At each step, the rest of the stick is kept for the remaining steps; thus, if at some step c we have a stick with length L the new stick will have length $\beta_c L$. It is shown in [40] that π constructed by this mechanism is a probability distribution. The following theorem provides an alternative definition for the DP which is constructive.

Theorem 7. *Given a base measure G_0 consider the random measure*

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k},$$

where $\theta_k \overset{iid}{\sim} G_0$ and $\pi \sim GEM(\alpha_0)$. Then $G \sim DP(\alpha_0, G_0)$.

The proof is provided in [40]. That is, the two alternative definitions of DP provided in Definition 4 and Theorem 7 are equivalent. Theorem 7 shows that the random probability measures generated from a DP are discrete with probability one [12]. This leads to the predictive distribution, the distribution of a new point y_{N+1} given a set of N existing i.i.d. points $Y = \{y_1, \cdots, y_N\}$, for the DP known as the **Chinese restaurant process (CRP)**.

2.3.4 CRP definition

As mentioned in Section 2.3.3, the DP produces discrete random measures; thus, there is a positive probability of multiple points y_i taking identical values and hence forming clusters. Recall that the points $y_1, \cdots, y_N$ are i.i.d. draws from G, a realization of a DP with some concentration parameter α_0 and a base measure G_0; that is, $G \sim DP(\alpha_0, G_0)$. Since the values of the draws are repeated, let $y_1^*, \cdots, y_M^*$ be the unique values among $y_1, \cdots, y_N$ and n_k be the number of repeats of y_k^*. The predictive distribution for the $N+1$th point y_{N+1} given all the previous observations $y_1, \cdots, y_N$ is given by the following theorem:

Theorem 8. *Let $y_1, \cdots, y_N$ be an i.i.d. sample from $G \sim DP(\alpha_0, G_0)$, $y_1^*, \cdots, y_M^*$ be the unique values among $y_1, \cdots, y_N$, and n_k be the number of repeats of y_k^*. Then* *the predictive distribution for y_{N+1} is given by:*

$$y_{N+1}|y_1,\cdots,y_N,\alpha_0,G_0 \sim \sum_{i=1}^{M} \frac{n_i}{N+\alpha_0}\delta_{y_i^*} + \frac{\alpha_0}{N+\alpha_0}G_0. \tag{2.5}$$

For a proof of Theorem 8 see for example [23]. The CRP refers to the above distribution (i.e., predictive distribution of a DP). The interpretation of Equation 2.5 in terms of the CRP is as follows. Consider a restaurant with an infinite set of tables; tables correspond to clusters, the ith customer corresponds to the ith point y_i, and the kth dish corresponds to the kth unique value that points can attain y_k^*. The customers sequentially enter the restaurant and sit either at a previously occupied table or at a new table. Customer i enters and sits at an already occupied table k with probability proportional to the number of customers already at that table n_k; that is,

$$P(Y_i = y_k^*) = \tfrac{n_k \delta_{y_k^*}}{\alpha_0+N},$$

or sits at a new table with probability proportional to α_0; that is,

$$P(Y_i = y^*) = \tfrac{\alpha_0 G_0}{\alpha_0+N}.$$

When the table is occupied the new customer shares the same dish as the first customer at that table, and when the table is empty a new dish is sampled $y^* \sim G_0$.

2.3.5 DP mixture model

One of the most common applications of the DP is the DP mixture model which can be used for clustering data. In fact, the DP mixture model is a mixture model with countably infinite mixture components. Consider a set of points $y_1, y_2, \cdots, y_N$, a set of latent parameters $\theta_1, \cdots, \theta_n$, and a parametric distribution F. Then the DP mixture model can be written in the following form:

$$y_i|\theta_i \sim F(\theta_i)$$
$$\theta_i|G \sim G$$
$$G|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

Each observation y_i is sampled from $F(\theta_i)$ and each θ_i is sampled i.i.d. from G. As G is discrete, samples θ_i from G can have identical values; therefore, observations sampled from F with the shared parameter θ_i belong to the same mixture component (i.e., cluster).

2.4 Summary

In this chapter we reviewed some of the basic notions that will be useful in developing our model. In the subsequent chapters, we will show how our nonparametric prior over MJPs is built upon the MGP that we described in this chapter. Moreover, we will present the properties of our model that we believe can be better understood in connection with the properties of the DP explained in the current chapter. In the following chapter, we review some Bayesian nonparametric models and explain how our proposed model is different from these currently available models.

Chapter 3

RELATED MODELS

Our model can be considered as a prior over MJPs, which are a specific type of time series. In this chapter, we present the currently available Bayesian nonparametric models that are used in time series modelling; we describe their properties and group them into three main categories.

3.1 Hierarchical Dirichlet Process

In this section, we introduce the hierarchical Dirichlet process (HDP) [42], a non-parametric prior over a set of random probability measures over $(\Omega, \mathscr{F}_\Omega)$. We first explain the motivation behind the HDP from the perspective of mixture models; next, we describe different representations of the HDP which are in fact analogous to different representations of the DP.

3.1.1 Motivation

Consider a clustering problem where there are multiple groupings of data; we want to be able to share the clusters across the multiple groupings of data. As a concrete example, consider a document modeling problem where the goal is to model a corpus of documents. A common simplification in document modeling is "the bag of words" assumption by which the words are assumed to be exchangeable and the order of the words in the documents is ignored. Another common assumption in document modeling is that the words in each document are sampled from the distribution of a topic [4]; each topic has a multinomial distribution over a basic vocabulary. For example, in a document on the applications of machine learning in genetics, the words might be sampled from the topics “genetics” and “machine learning”. Now, considering a collection of documents (i.e., corpus) we would like to share the topics among the documents. We might also be interested in sharing among different corpus.

3.1.2 Definition

The HDP is defined by a set of hyperparameters: a baseline probability measure H, and concentration parameters γ and α. The HDP is defined as follows:

$$G_0|\gamma, H \sim DP(\gamma, H) \tag{3.1}$$

$$G_j|\alpha, G_0 \overset{iid}{\sim} DP(\alpha, G_0) \tag{3.2}$$

$$\theta_{ij}|G_j \overset{iid}{\sim} G_j \tag{3.3}$$

According to this definition a global measure G_0 is sampled from a DP with the baseline measure H and concentration parameter γ; G_j is the random probability measure for group j sampled from a DP with base measure G_0 and concentration parameter α. Given G_0, the G_js are independent from each other. In other words, through the global measure we are linking different groups together which will allow us to share statistical strength between them. The connection of the HDP with mixture models in general, and topic models in particular, is completed with the specification of a parametric distribution F for the observations:

$$x_{ij}|\theta_{ij} \sim F(\theta_{ij}). \tag{3.4}$$

Here we assume given θ_{ij}, the parameter of the ith cluster in the jth group, the observation x_{ij} is independent of all other observations. In comparison with the DP mixture model (Section 2.3.5), in the HDP mixture model we have a collection of mixture models that are connected through a global base measure G_0. Note that we are not limited to a single layer of hierarchy in the HDP. In fact, we can have as many layers as required for modeling through a recursive structure where the

Notation	Definition
θ_{ij}	ith customer in restaurant j
ϕ_k	kth sampled dish from the global menu H
ψ_{jt}	the dish served at table t in restaurant j
n_{jtk}	number of customers in restaurant j at table t eating dish k
$n_{jt.}$	number of customers in restaurant j at table t
$n_{j.k}$	number of customers in restaurant j eating dish k
m_{jk}	number of tables in restaurant j serving dish k
$m_{.k}$	number of tables serving dish k
$m_{j.}$	number of occupied tables in restaurant j
$m_{..}$	total number of occupied tables

Table 3.1: Variables used in the CRF

base measure for each layer is itself sampled from another DP. An example of a situation where we might need more than one layer would be sharing topics among different corpus.

3.1.3 Chinese Restaurant Franchise

In Section 2.3.4 we introduced the CRP: a representation for the predictive distribution of a DP where we marginalize out the random probability measure. There is an analogous representation for the HDP which is called a **Chinese restaurant franchise (CRF)**.

In a CRF we have multiple restaurants with a shared menu. For each restaurant, a customer entering the restaurant can sit at an occupied table or at a new table. A customer sitting at an occupied table orders the same dish as that for the first customer at that table; for a new table a dish is ordered from the shared menu across restaurants.

Each restaurant corresponds to a group and customers correspond to the θ_{ij}s in Equation 3.3 (i.e., θ_{ij} is the ith customer at the jth restaurant). Moreover, following the notation in [42], we denote K i.i.d. samples from the baseline measure H by $\phi_1, \cdots, \phi_K$; here H corresponds to the global menu of the dishes and ϕ_k is the kth sampled dish from the global menu. The notation needed for defining the predictive distribution of an HDP is provided in Table 3.1.

Similar to the DP, in order to obtain the predictive distribution for the HDP we have to marginalize out the random probability measures G_0 and the G_js. First we marginalize out G_j. From Theorem 8, the predictive distribution for θ_{ij} (ith latent parameter in group j) given $\theta_{1j}, \theta_{2j}, \cdots, \theta_{i-1,j}$ is:

$$\theta_{ij} | \theta_{1j}, \cdots, \theta_{i-1,j}, \alpha, G_0 \sim \sum_{t=1}^{m_{j.}} \frac{n_{jt.}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0. \tag{3.5}$$

Note that the similarity to Equation 2.5. That is due to the fact that in both we are marginalizing out realizations of a DP.

Marginalizing out G_0, using Theorem 8 again, we can obtain the predictive distribution for the dishes. Anytime we need to sample a new dish in Equation 3.5 (i.e., sampling from G_0), we sample from the predictive distribution of the dishes. The dishes are sampled in order of the tables being occupied in each restaurant; that is, we start from table 1 in restaurant j and assign a dish to it (ψ_{j1}), next we move to the second table (ψ_{j2}), and so on. The predictive distribution for ψ_{jt} given the dishes for all occupied tables in all restaurants, $\psi_{11}.\psi_{12}, \cdots, \psi_{21}, \cdots, \psi_{j,t-1}$, is:

$$\psi_{jt} | \psi_{11}, \psi_{12}, \cdots, \psi_{21}, \cdots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\gamma} \delta_{\phi_k} + \frac{\gamma}{m_{..}+\gamma} H. \tag{3.6}$$

Equation 3.5 is a mixture for the customers; a customer entering restaurant j: 1) sits at a preoccupied table with probability proportional to the number of customers sitting at that table or 2) sits at a new table with probability proportional to α_0. Equation 3.6 is a mixture for the dishes; whenever a new dish is needed for a new table: 1) it is chosen from the previously chosen dishes with probability proportional to the number of tables serving that dish or 2) a new dish is sampled with probability proportional to γ. In summary, to obtain a new sample θ_{ij} we first use Equation 3.5; if a new dish is needed from G_0 we proceed to Equation 3.6 and sample a new dish ψ_{jt} and set $\theta_{ij} = \psi_{jt}$.

We can combine Equations 3.5 and 3.6 as follows:

$$\begin{aligned}\theta_{ij}|\theta_{1j},\cdots,\theta_{i-1,j},\alpha \sim \sum_{k=1}^{K}\frac{n_{j.k}}{i-1+\alpha_0}\delta_{\phi_k}+\frac{\alpha_0}{i-1+\alpha_0}\mu \\ \mu=\sum_{k=1}^{K}\frac{m_{.k}}{m_{..}+\gamma}\delta_{\phi_k}+\frac{\gamma}{m_{..}+\gamma}H.\end{aligned} \tag{3.7}$$

This can be done due to the fact that

$$\sum_{t=1}^{m_{j.}} n_{jt.}\delta_{\psi_{jt}} = \sum_{k=1}^{K}\sum_{t=1}^{m_{jk}} n_{jtk}\delta_{\psi_{jt}} = \sum_{k=1}^{K} n_{j.k}\delta_{\phi_k}.$$

This representation will be useful in Chapter 5.

3.1.4 Stick-breaking representation for the HDP

Similar to the DP, there is a stick-breaking representation for the HDP. We have this representation for the global measure G_0 and also for the other random probability measures G_js, as all are sampled from a DP. Assume $\phi_k \overset{\text{iid}}{\sim} H$; it can be shown [42] the stick-breaking representation for G_0 and the G_js is as follows:

$$G_0 = \sum_{k=1}^{\infty}\beta_k\delta_{\phi_k} \tag{3.8}$$

$$G_j = \sum_{k=1}^{\infty}\pi_{jk}\delta_{\phi_k} \tag{3.9}$$

where $\beta = (\beta_k)_{k=1}^{\infty} \sim GEM(\gamma)$ are mutually independent, and $\pi_j = (\pi_{jk})_{k=1}^{\infty} \sim DP(\alpha_0,\beta)$. Note that the weights π_j are independent given β as the G_j are independent given G_0. In this representation, the support of G_0 is $\phi = (\phi_k)_{k=1}^{\infty}$; each G_j has the same support since each G_j is a realization of a DP with base measure G_0. In summary, the stick-breaking representation of the HDP can be written as follows:

$$\phi_k \overset{\text{iid}}{\sim} H \tag{3.10}$$

$$\beta|\gamma \sim GEM(\gamma) \tag{3.11}$$

$$\pi_j|\alpha_0,\beta \overset{\text{iid}}{\sim} DP(\alpha_0,\beta) \tag{3.12}$$

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \tag{3.13}$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \tag{3.14}$$

3.2 HDP-HMM

In this section we describe a closely related model to the HDP which is called the **hierarchical Dirichlet process hidden Markov model (HDP-HMM)** [42]. Before introducing the model we briefly review hidden Markov model (HMM).

3.2.1 Background on HMM

HMMs are doubly stochastic Markov chains. That is, there are finite number of observations at discrete time points; we denote the set of observations by $\mathscr{Y} = (Y_1, Y_2, \ldots, Y_N)$. Each of these observations is a $\mathscr{X}$-valued random variable; given its hidden state θ the observation is independent of all the other observations. In other words, there is a hidden state at each of the discrete time points; these hidden states follow a Markov chain on the space of hidden states Ω.

HMMs are defined by a transition probability matrix P that characterizes the dynamics of the hidden states, and a parametric family $\mathscr{P}$ indexed by the hidden states θ of the chain, $\mathscr{P} = \{L_\theta : \mathscr{F}_{\mathscr{X}} \to [0,1], \theta \in \Omega\}$. The parametric family determines the probability of "emitting" an observation by a hidden state. A graphical representation of an HMM is sketched in Figure 3.1.

An HMM can be viewed as a dynamic mixture model; the choice of the mixture component for each observation is not independent of the other observations but depends on choice of mixture component for the previous observation. That is, for each hidden state θ there is a corresponding random probability measure μ_θ which

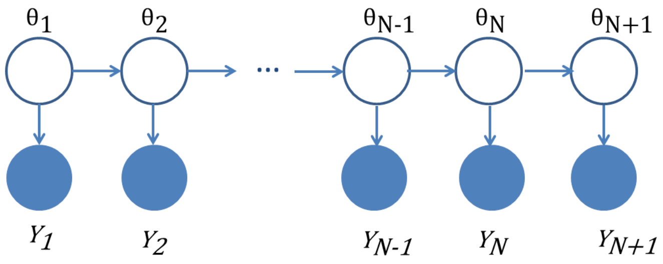


Figure 3.1: A graphical representation for an HMM

can be thought of as a multinomial distribution over mixture components.

More concretely, assume the current hidden state is θ_N where $\theta_N \in \Omega$. Since the hidden states follow a Markov chain, the next hidden state is a random variable distributed as a multinomial distribution that corresponds to the θ_Nth row of P, denoted $\theta_{N+1} \sim P_{\theta_N}$. The next observation would be a random variable sampled from a parametric family indexed by θ_{N+1}.

3.2.2 Definition

The dynamic mixture view of the HMM allows us to extend its definition to an infinite number of hidden states which yields the infinite mixture model. The result, an HDP-HMM [42], is a Bayesian model for the HMM with *a nonparametric prior over an infinite-dimension transition probability matrix*. We assume a DP prior over the random probability measure μ_θ (equivalent to G_j in 3.2) corresponding to each hidden state of the HMM; informally, we are assuming an infinite-dimension multinomial distribution for each hidden state. However, the DP priors for all the hidden states need to be linked as we want the chain to be irreducible.

One approach for linking the DP priors for the states could be using a shared base measure H among all of the priors, $\mu_\theta \overset{\text{iid}}{\sim} DP(\alpha, H) \quad \forall \theta \in \Omega$. This approach may work if the base measure has a countable support; otherwise, samples from the DPs will have different supports with probability one.

Another approach, that works even with uncountable support for the base measure, is using the HDP framework; we first sample a global measure μ_0 from a DP with base measure H and then use the global measure as a shared base measure for the DPs of the hidden states. Let θ_N be the current state of the chain, y_{N+1}

the observation emitted by the hidden state θ_{N+1}, and F a parametric distribution. Then, formally, the HDP-HMM model has the following form:

$$\mu_0 \sim DP(\gamma, H) \tag{3.15}$$

$$\mu_\theta \overset{\text{iid}}{\sim} DP(\alpha, \mu_0) \quad \forall \theta \in \Omega \tag{3.16}$$

$$\theta_{N+1} \mid \{\theta_n\}_{n=1}^{N}, \{\mu_\theta\} \sim \mu_{\theta_N} \tag{3.17}$$

$$y_{N+1} \mid \theta_{N+1} \sim F(\theta_{N+1}) \tag{3.18}$$

It is worth mentioning that a model was introduced in [3] that allows countably infinite hidden states for an HMM. This model, known as the **infinite hidden Markov model (IHMM)**, is in fact a CRF representation of an HDP-HMM. This model can be described as a two level hierarchy. Given the current state $\theta_N = i$, there could be:

1. a self transition with probability proportional to the number of self transitions that were observed before for state i, n_{ii} plus a parameter α (i.e., $n_{ii} + \alpha$),

2. a transition to an already visited state j with probability proportional to n_{ij}, the number of times this transition was observed before,

3. a move to a higher level of the hierarchy where an "oracle" process is invoked, with probability proportional to a parameter β. At this level of the hierarchy there are two possibilities:

   (a) a transition to an already visited state j, with probability proportional to n_j^0 the number of times state j was previously chosen by the oracle regardless of the previous state,

   (b) a transition to a novel state with probability proportional to a parameter γ.

The role of the oracle is similar to the role of μ_0 in the definition of the HDP-HMM (Equation 3.15); that is, it ties together the transition models to have destination states in common. Figure 3.2 illustrates the transition generative mechanism. The equivalence of this model and a CRF representation of HDP-HMM is shown in [42].

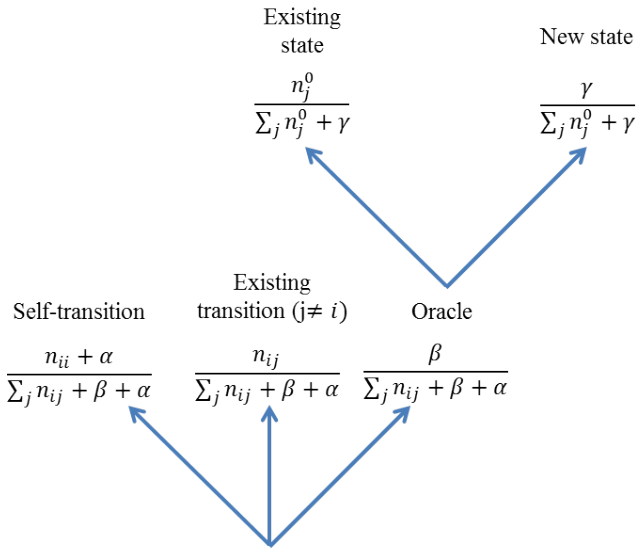

Existing state
$\frac{n_j^0}{\sum_j n_j^0 + \gamma}$
New state
$\frac{\gamma}{\sum_j n_j^0 + \gamma}$
Self-transition
$\frac{n_{ii} + \alpha}{\sum_j n_{ij} + \beta + \alpha}$
Existing transition (j≠ i)
$\frac{n_{ij}}{\sum_j n_{ij} + \beta + \alpha}$
Oracle
$\frac{\beta}{\sum_j n_{ij} + \beta + \alpha}$

Figure 3.2: Generative mechanism for the state transitions in the iHMM model

3.3 Sticky HDP-HMM

Although the HDP-HMM was successful in various applications (see for example [26] or [45]) it has limitations. One limitation of the HDP-HMM is inadequacy in modeling temporal persistence of states; the HDP-HMM creates unnecessarily rapid transitions between states in order to explain all the transitions. By using $\mu_\theta \overset{\text{iid}}{\sim} DP(\alpha, \mu_0) \;\; \forall \theta \in \Omega$ as in Equation 3.16, the HDP-HMM cannot differentiate between self-transitions and transitions to other states. Thus, although all states have similar transition distributions, there could be unnecessary transitions to other states. Particularly, in data sequences with persistent states, the flexibility of HDP-HMM will cause the sequence samples with rapid state switches to have high posterior probability. In other words, HDP-HMM will generate new redundant states instead of having larger probability of self transition.

The **sticky HDP-HMM** model introduced in [13] provides a solution to the problem of state-persistence in the HDP-HMM. The model simply includes a self-transition bias parameter. For presenting the sticky HDP-HMM model in this section, *we assume that all the hidden states are observed*; hence, by the data sequence we mean the sequence of the states.

The model proposed in [13] is a slight modification of the model in Section 3.1.4. Using the same notation as Sections 3.1.4 and 3.2 the model can be defined as follows:

$$\phi_k \overset{\text{iid}}{\sim} H \tag{3.19}$$

$$\beta \mid \gamma \sim GEM(\gamma) \tag{3.20}$$

$$\pi_\theta \mid \alpha_0, \beta, \eta \overset{\text{iid}}{\sim} DP(\alpha_0 + \eta, \frac{\alpha_0 \beta + \eta \delta_\theta}{\alpha_0 + \eta}) \tag{3.21}$$

$$\mu_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \tag{3.22}$$

$$\mu_\theta = \sum_{k=1}^{\infty} \pi_{\theta k} \delta_{\phi_k} \forall \theta \in \Omega \tag{3.23}$$

$$\theta_{N+1} \mid \{\theta_n\}_{n=1}^{N}, \{\mu_\theta\} \sim \mu_{\theta_N} \tag{3.24}$$

$$\tag{3.25}$$

In this definition η is the self-transition bias parameter introduced to the HDP; here $\alpha_0 \beta + \eta \delta_\theta$ means that $\eta > 0$ is added to the θth component of $\alpha_0 \beta$. This parameter will encourage self-transitions in the model; note that $\eta = 0$ reduces the model to the HDP-HMM.

3.4 Dependent Dirichlet Process (DDP)

DDPs are a general framework for modeling a collection of random probability measures $G_{\mathscr{T}} = \{G_t : t \in \mathscr{T} \subset \mathbb{R}^d\}$. Here G_t is a random probability measure on some space $(\Omega, \mathscr{F}_\Omega)$ which is indexed by elements $t \in \mathscr{T}$. We only consider time space in the following; hence, $\mathscr{T} = \mathbb{R}^+$. As in the previous section, for presenting the DDP model in this section, *we assume that all the states are observed.*

In the DDP we start from the stick-breaking construction of the DP (Section 2.3.3) and replace the weights and/or atoms with appropriate stochastic processes [29]. The general definition for a DDP provided in [29] is:

Definition 9. *A DDP is a collection of random probability measures* $G_{\mathscr{T}} = \{G_t :$

$t \in \mathscr{T}\}$; G_t is defined as

$$G_t = \sum_{k=1}^{\infty} \pi_k(t) \delta_{\theta_k(t)} \tag{3.26}$$

where $\{\theta_k(t) : t \in \mathscr{T}\}$, for $k = 1, 2, \cdots$ are independent realizations from a stochastic process G_0 defined on $\mathscr{T}$ and the $\pi_k(t)$s are distributed independently across t according to $GEM(\alpha_t)$.

Note that in the above definition both the set of atoms and the set of weights depend on t. The DP and HDP can be considered as special cases of the DDP; for any fixed t, the DDP yields a DP. Moreover, the HDP is an example of the DDP where we have dependence in the weights but not in the atoms (see Section 3.1.4).

The simplest construction for the DDP is obtained when the atoms are replaced with a stochastic process, but the weights remain constant at different values of t. The resulting model is called a **single-p DP** [29] which is simply a DP mixture of stochastic processes. The advantage of this model is that it is simple and fairly flexible; however, the disadvantage is that the model has a lack of "locality" due to the global sharing of π. More general DDP models discuss how π should vary across $\mathscr{T}$.

A model proposed in [17], the **order-based DDP**, uses a common collection of stick-breaking proportions $\beta = \{\beta_i\}_{i=1}^{\infty}$ for all G_t, $t \in \mathscr{T}$. Since the β_is are i.i.d. for any permutation σ, $\pi_c = \beta_{\sigma(c)} \prod_{1 \leq c' < c} (1 - \beta_{\sigma(c')})$ is valid. The model allows the permutation σ_t to vary with t. Note that at any time we have the usual stick breaking construction; thus, the marginal over t of the defined process is DP.

Another model which deals with a π varying across $\mathscr{T}$ is proposed in [10]; this model is in the discrete-time framework. The evolution of the random probability measure in the model is defined as $G_t = \nu G_{t-1} + (1 - \nu)\varepsilon_t$ where ν is a parameter and ε_t is a sample from a DP.

The continuous-time version of the previous model, called the **stick-breaking autoregressive process**, is introduced in [18]. The process is defined as $G_t = \tilde{G}_{N(t)}$ where $N(t)$ is a Poisson process with rate λ and $\tilde{G}_i$ is defined as follows:

$$\tilde{G}_i = \beta_i \tilde{G}_{i-1} + (1 - \beta_i) \delta_{\theta_i} \tag{3.27}$$

where $\beta_i \sim Beta(1, \alpha_0)$ and $\theta_i \sim G_0$. That is, there are jumps in the distributions at the arrival times of the Poisson process. At the i-th arrival time a new atom is introduced and the effect of the previous atoms decay with the rate β_i. It can be shown that both G_t and $\tilde{G}_i$ follow a DP with concentration parameter α_0 and base measure G_0. Moreover, for any set ω:

$$Corr(G_t(\omega), G_{t+s}(\omega)) = \rho^s \tag{3.28}$$

where $\rho = exp(-\frac{\lambda}{\alpha_0+1})$; thus, the dependence between the random measures decreases exponentially with their separation in time.

The common theme among all of the continuous-time DDP models presented in this section is that the correlation between the marginals of G_{t_1} and G_{t_2} should decay smoothly with $|t_1 - t_2|$. That is, if we have points at times t_1 and t_2 then these models are *transient* in the sense that the effect of a point at time t_1, on the point at time t_2 should decrease as $t_2 - t_1$ increases.

3.5 Summary

In this chapter we discussed some models that are closely related to our proposed model; we also mentioned that the DDP can be considered as a general framework that covers all of these models. There are other related models that we will not explain in detail as they have less importance in the following chapters of the thesis. However, for the sake of completeness we list these models here. Other models that can be considered as examples of the DDP are: the **Ornstein-Uhlenbeck DP** [16], a DDP model in the continuous-time framework with fixed atoms (i.e., the atoms do not depend on time); the **time series dependent DP** [35], a DDP in discrete time; the **nested DP** [38], a model which clusters similar distributions together, in contrast to the HDP which shares clusters among all distributions; and the **local DP** introduced in [7] that provides a distribution for a collection of random probability measures indexed by predictors.

All the models presented in this chapter can be grouped into four categories based on being 1) continuous/discrete-time and 2) transient/non-transient. All the models in Sections 3.1, 3.2, 3.3 are in the discrete-time framework, whereas the models in Section 3.4 are in the continuous-time framework (unless specified otherwise). In Sections 3.2 and 3.3 the models are non-transient in contrast to the models introduced in Section 3.4 which are transient. In this thesis we are going to propose a model for continuous-time non-transient processes

Chapter 4

GAMMA EXPONENTIAL PROCESS

In this chapter, we first present the formal definition of the gamma exponential process (GEP). Next, we show that this model has some attractive properties such as conjugacy and a closed form expression for the predictive distribution.

4.1 Definition

In a GEP, we first obtain the rows of the rate matrix Q by a transformation of i.i.d. samples from an MGP; then, we generate the states from Q with the Doob-Gillespie algorithm described in the Section 2.1. We will denote the normalization constant of a measure μ (i.e., $\mu(\Omega)$ where Ω is the support of μ) by $\|\mu\|$ and the normalized measure by $\bar{\mu} = \frac{\mu}{\|\mu\|}$. Moreover, *we assume that all the states are observed for now*, and treat the partially observed case in Chapter 6. That is, using the notation of Section 2.1, we assume $X = (\theta_n, J_n)_{n=1}^N$ is the list of observed events.

Definition 10 (GEP)**.** *Let H_0 be a base measure on a countable support Ω with $\|H_0\| < \infty$. The GEP is formally defined as follows:*

$$\begin{aligned} \mu_\theta &\overset{iid}{\sim} \mathrm{MGP}(H_0, \beta_0) \ \ \forall \theta \in \Omega \\ \theta_{N+1} \,\big|\, X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \bar{\mu}_{\theta_N} \\ J_{N+1} \,\big|\, X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \mathrm{Exp}\left(\|\mu_{\theta_N}\|\right) \end{aligned}$$

We will relax the countable base measure support assumption in the next section.

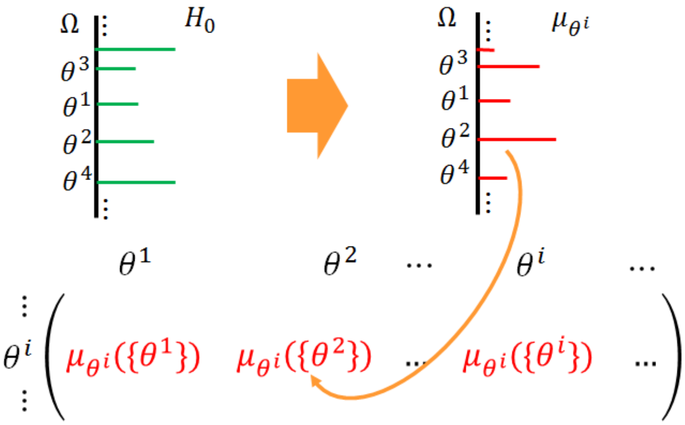


Figure 4.1: Generating samples of μ_θ from $\mathrm{MGP}(H_0, \beta_0)$. We use superscript i instead of subscript N to distinguish the ith atom from the state at the Nth jump. We present the row corresponding to state θ^i in the rate matrix. As an example, if $\theta_N = \theta^i$, then the probability of transition to θ^2 is proportional to $\mu_{\theta^i}(\{\theta^2\})$.

Figure 4.1 shows an example of generating samples of μ_θ from $\mathrm{MGP}(H_0, \beta_0)$.

Note that Definition 10 has positive self-transition rates as the diagonal elements so self-transitions are allowed. However, we will use this definition since for computing predictive distributions, it will be simpler to allow positive self-transition rates. The next proposition shows we can remove the self-transitions using a transformation and obtain an equivalent process; by equivalent we mean the two processes have the same distribution of waiting time in each state, and the same transition probability from a state to a different state.

Proposition 11. *Let S_t be the MJP obtained from a GEP. Then we can obtain an equivalent MJP (S_t^*) by arbitrarily ordering $\Omega = \theta^{(1)}, \theta^{(2)}, \ldots$, and setting:*

$$q^*_{\theta^{(i)},\theta^{(j)}} = \begin{cases} \mu_{\theta^{(i)}}(\{\theta^{(j)}\}) & \text{if } i \neq j \\ \|\mu_{\theta^{(i)}}\| \ \left(\bar{\mu}_{\theta^{(i)}}(\{\theta^{(i)}\}) - 1\right) & \text{o.w.} \end{cases}$$

S_t^* generated from Q^* has no self-transitions.

Proof. We can prove that the two processes are equivalent by showing given the current state, the distribution of the waiting time in that state and the transition probability to the next different state for the two processes are the same.

$P^*_{\theta^{(i)},\theta^{(j)}}$, the probability of transition from $\theta^{(i)}$ to a different state $\theta^{(j)}$ in the process S^*_t, is equal to $\frac{\mu_{\theta^{(i)}}(\{\theta^{(j)}\})}{\|\mu_{\theta^{(i)}}\| - \mu_{\theta^{(i)}}(\{\theta^{(i)}\})}$. Assuming the process is at the Nth jump, the same probability in S_t can be obtained as follows:

$$P_{\theta^{(i)},\theta^{(j)}} = P(\theta_{N+1} = \theta^{(j)} | \theta_N = \theta^{(i)}, i \neq j) = \frac{\frac{\mu_{\theta^{(i)}}(\{\theta^{(j)}\})}{\|\mu_{\theta^{(i)}}\|}}{1-\frac{\mu_{\theta^{(i)}}(\{\theta^{(i)}\})}{\|\mu_{\theta^{(i)}}\|}} = \frac{\mu_{\theta^{(i)}}(\{\theta^{(j)}\})}{\|\mu_{\theta^{(i)}}\| - \mu_{\theta^{(i)}}(\{\theta^{(i)}\})}.$$

Thus, the probability of transition from one state to another is the same for S_t and S^*_t.

Now we show the waiting time T in state $\theta^{(i)}$ for both processes is distributed as $Exp(\|\mu_{\theta^{(i)}}\| - \mu_{\theta^{(i)}}(\{\theta^{(i)}\}))$. By definition, this is true for S^*_t. For S_t we have:

$$\begin{aligned} P(T < t) =& P(\cup_{n=1}^{\infty}\{\text{leave state } \theta^{(i)} \text{at jump } n \text{ \& jump } n \text{ occurs before time } t\}) = \\ & \sum_{n=1}^{\infty} P\{\text{leave state } \theta^{(i)} \text{at jump } n\} \times \\ & \quad P\{\text{jump } n \text{ occurs before time } t | \text{leave state } \theta^{(i)} \text{at jump } n\} = \\ & \sum_{n=1}^{\infty} p_n P\{\text{jump } n \text{ occurs before time } t | \text{leave state } \theta^{(i)} \text{ at jump } n\} \end{aligned}$$

where $p_n = P\{\text{leave state } \theta^{(i)} \text{ at jump } n\} = (\frac{\mu_{\theta^{(i)}}(\{\theta^{(j)}\})}{\|\mu_{\theta^{(i)}}\|})^{n-1}(1 - \frac{\mu_{\theta^{(i)}}(\{\theta^{(i)}\})}{\|\mu_{\theta^{(i)}}\|})$ is the distribution for a geometric random variable with $p_{success} = 1 - \frac{\mu_{\theta^{(i)}}}{\|\mu_{\theta^{(i)}}\|}$.

Moreover, $P\{\text{jump } n \text{ occurs before time } t | \text{leave state } \theta^{(i)} \text{ at jump } n\}$ is the cumulative distribution function evaluated at time t for the sum of n i.i.d. exponential random variables. Note that, conditioned on the event that state $\theta^{(i)}$ has self-transitions up to jump n, the waiting time to each jump (up to jump n) has an $Exp(\|\mu_{\theta^{(i)}}\|)$ distribution. It is known that the geometric sum of i.i.d. exponential random variables with rate $\|\mu_{\theta^{(i)}}\|$ has an exponential distribution with rate $p_{success} \times \|\mu_{\theta^{(i)}}\|$ (see for example [34]). Hence, we have proved the waiting times have the same distribution in both processes. □

To understand the connection with the Doob-Gillespie algorithm presented in

Section 2.1 note that we can use Proposition 11 to obtain a rate matrix Q from the matrix generated from μ_θ in Definition 10.

As mentioned in Section 2.1, to completely define an MJP we also need the initial distribution. Instead of considering an initial distribution π we can equivalently assume that there is a special state θ_{beg} always present at the beginning of the sequence, and only at the beginning. That is, we always condition on $(\theta_0 = \theta_{\text{beg}}, J_0 = 0)$ and $(\theta_n \neq \theta_{\text{beg}}, n > 0)$, and drop these conditioning events from the notation. The initial distribution is then the set of transition probabilities out of θ_{beg}.

4.2 Conjugacy Property

In this section, we show that the posterior of each row, $\mu_\theta | X$, is also MGP distributed with updated parameters. The sufficient statistics for the parameters of $\mu_\theta | X$, the empirical transition measures F_θ and the empirical waiting times T_θ, are defined as follows:

$$F_\theta = \sum_{n=1}^{N} \mathbf{1}[\theta_{n-1} = \theta]\ \delta_{\theta_n}, \tag{4.1}$$

$$T_\theta = \sum_{n=1}^{N} \mathbf{1}[\theta_{n-1} = \theta]\ J_n. \tag{4.2}$$

The main result of this section is as follows:

Proposition 12. *The GEP is a conjugate family:* $\mu_\theta | X \sim \mathrm{MGP}\left(\mu'_\theta, \beta'_\theta\right)$, *where* $\mu'_\theta = F_\theta + H_0$ *and* $\beta'_\theta = T_\theta + \beta_0$.

For the proof of Proposition 12, we will need the following elementary lemma:

Lemma 13. *If* $V \sim \mathrm{Beta}(a,b)$ *and* $W \sim \mathrm{Gamma}(a+b,c)$ *are independent, then* $VW \sim \mathrm{Gamma}(a,c)$.

See for example [9] for a survey of standard beta-gamma results such as that stated in this lemma. We now prove the proposition:

Proof. For simplicity, fix an arbitrary state θ and drop the index (this is without loss of generality since the rows are iid); hence, $\mu = \mu_\theta \sim MGP(H_0, \beta_0)$, $\mu' = \mu'_\theta$, and $\beta' = \beta'_\theta$.

Let $(A_1, \ldots, A_K)$ be a measurable partition of Ω. By the Kolmogorov consistency theorem, it is enough to show that for all such partitions,

$$(\mu(A_1), \mu(A_2), \ldots, \mu(A_K)) \mid X \sim \text{Gamma}(\mu'(A_1), \beta') \times \cdots \times \text{Gamma}(\mu'(A_k), \beta'). \tag{4.3}$$

Assume for simplicity that $K = 2$ (the argument can be generalized to $K > 2$ without difficulty), and let $V = \mu(A_1)/\|\mu\|$, $W = \|\mu\|$. Then, by elementary properties of gamma distributed vectors:

$$V \sim \text{Beta}(H_0(A_1), H_0(A_2)), \tag{4.4}$$

$$W \sim \text{Gamma}(\alpha_0, \beta_0), \tag{4.5}$$

and V, W are independent (both conditionally on $X = (\theta_n, J_n)_{n=1}^N$ and unconditionally).

By beta-multinomial conjugacy, we have

$$V \mid X = V \mid (\theta_n, J_n)_{n=1}^N \overset{d}{=} V \mid \theta_1, \ldots, \theta_N \sim \text{Beta}(\mu'(A_1), \mu'(A_2)),$$

and by gamma-exponential conjugacy, we have

$$W \mid X \overset{d}{=} W \mid J_1, \cdots, J_N \sim \text{Gamma}(\|\mu'\|, \beta').$$

Using Lemma 13 with $a = \mu'(A_1), b = \mu'(A_2), c = \beta'$, we finally get that $\mu(A_1)|X = VW|X \sim \text{Gamma}(\mu'(A_1), \beta')$, which concludes the proof. □

Note that μ'_θ is the mean parameter of an unnormalized DP, as $\mu_\theta|X \sim \text{MGP}\left(\mu'_\theta, \beta'_\theta\right)$ which is an unnormalized DP.

4.3 Predictive Distribution

In this section we find an expression for the distribution of $(\theta_{N+1}, J_{N+1})|X$ which is the predictive distribution. We will need the following family of densities:

Definition 14 (Translated Pareto)**.** *Let $\alpha > 0, \beta > 0$. We say that a random variable T is* translated-Pareto*, denoted $T \sim \mathrm{TP}(\alpha, \beta)$, if it has density:*

$$f(t) = \frac{\mathbf{1}[t > 0]\alpha\beta^{\alpha}}{(t+\beta)^{\alpha+1}}. \tag{4.6}$$

Proposition 15. *The predictive distribution of the GEP is given by:*

$$(\theta_{N+1}, J_{N+1}) \mid X \sim \bar{\mu}'_{\theta_N} \times \mathrm{TP}(\|\mu'_{\theta_N}\|, \beta'_{\theta_N}). \tag{4.7}$$

Proof. By Proposition 12, it is enough to show that if $\mu \sim \mathrm{MGP}(H_0, \beta_0)$, $\theta|\mu \sim \bar{\mu}$, and $J|\mu \sim \mathrm{Exp}(\|\mu\|)$, then

$$(\theta, J) \sim \bar{\mu} \times \mathrm{TP}(\alpha_0, \beta_0),$$

where $\alpha_0 = \|H_0\|$.

Note first that because $(J|\theta) \overset{d}{=} J$ the minimum and argmin of independent exponential random variables are independent.

To get the distribution of J, we need to integrate over the rate parameter of the exponential waiting time distribution. Here we denote the distribution of the waiting time given the rate by $p(t|x)$ and the distribution of the rate parameter by $p(x)$. From Definition 10 and the properties of the gamma process described in Section 2.2, we know the rate parameter has a gamma distribution. Thus, to obtain the distribution of J we compute the following integral:

$$\begin{aligned} p(t) &= \int_{x>0} p(t|x)p(x)\,\mathrm{d}x \\ &= \int_{x>0} \mathrm{Exp}(t;x) \cdot \mathrm{Gamma}(x;\alpha_0,\beta_0)\,\mathrm{d}x \\ &= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \int_{x>0} x\exp(-xt) \cdot x^{\alpha_0-1}\exp(-\beta_0 x)\,\mathrm{d}x \\ &= \int_{x>0} x^{\alpha_0}\exp\left(-(\beta_0+t)x\right)\,\mathrm{d}x = \frac{\alpha_0\beta_0^{\alpha_0}}{(\beta_0+t)^{\alpha_0+1}} \end{aligned}$$

Hence $J \sim \mathrm{TP}(\alpha_0, \beta_0)$. □

A useful property of predictive distributions is exchangeability. We can show that these predictive distributions are indeed exchangeable:

Proposition 16. *Let $J_{j(\theta,1)}, J_{j(\theta,2)}, \ldots, J_{j(\theta,K)}$ be the subsequence of waiting times following state θ. Then the random variables $J_{j(\theta,1)}, J_{j(\theta,2)}, \ldots, J_{j(\theta,K)}$ are exchangeable. Moreover, the joint density of the sequence of waiting times $(J_{j(\theta,1)} = j_1, J_{j(\theta,2)} = j_2, \ldots, J_{j(\theta,K)} = j_K)$ is given by:*

$$p(j_1, j_2, \ldots, j_K) = \frac{\mathbf{1}[j_k > 0, k \in \{1, \ldots, K\}](\alpha_0)_K \beta_0^{\alpha_0}}{(\beta_0 + j_1 + \cdots + j_K)^{\alpha_0 + K}} \tag{4.8}$$

where the Pochhammer symbol $(x)_n$ is defined as $(x)_n = x(x+1)\cdots(x+n-1)$.

Proof. From Proposition 15, we have:

$$\begin{aligned}
p(j_1, j_2, \ldots, j_K) &= p(j_1)p(j_2|j_1) \times \cdots \times p(j_K|j_1, \ldots, j_K) \\
&= TP(\alpha_0, \beta_0) TP(\alpha_0 + 1, \beta_0 + j_1) \times \cdots \times TP(\alpha_0 + K - 1, \beta_0 + j_1 + \cdots + j_{K-1}) \\
&= \mathbf{1}[j_k > 0, k \in \{1, \ldots, K\}] \frac{\alpha_0 \beta_0^{\alpha_0}}{(\beta_0 + j_1)^{\alpha_0+1}} \frac{(\alpha_0 + 1)(\beta_0 + j_1)^{\alpha_0+1}}{(\beta_0 + j_1 + j_2)^{\alpha_0+2}} \\
&\times \cdots \times \frac{(\alpha_0 + K - 1)(\beta_0 + j_1 + \cdots + j_{K-1})^{\alpha_0+K-1}}{(\beta_0 + j_1 + \cdots + j_K)^{\alpha_0+K}} \\
&\propto \mathbf{1}[j_k > 0, k \in \{1, \ldots, K\}](\beta_0 + j_1 + \cdots + j_K)^{-\alpha_0 - K}.
\end{aligned}$$

The second line is obtained from the first by considering the fact that each jump to state θ updates the sufficient statistics F_θ and T_θ; consequently, the parameters of the TP (i.e., $\|\mu'_\theta\|$ and β'_θ) are updated. The normalization of the above expression is indeed equal to $1/((\alpha_0)_K \beta_0^{\alpha_0})$. □

Chapter 5

Hierarchical Gamma Exponential Process

5.1 Motivation

In this section, we present a hierarchical version of the GEP, where the rate matrix, instead of i.i.d. rows, has exchangeable rows. Informally, the motivation behind this construction is to have the rows share information on what states are frequently visited. This is similar to the motivation behind the HDP-HMM model presented in Section 3.2. Recall that in order to link the rows of the transition probability matrix in the HDP-HMM model, we used a random shared base probability measure distributed according to a DP.

As with HDPs, the hierarchical construction is especially important when Ω is uncountable. For such spaces, since each GEP sample has a random countable support, any two independent GEP samples will have disjoint supports with probability one. Therefore, a GEP alone cannot be used to construct recurrent Markov processes when Ω is uncountable. Fortunately, the hierarchical model introduced in this chapter addresses this issue: it yields a recurrent model for MJPs over both countable and uncountable spaces Ω.

5.2 Definition

The hierarchical gamma exponential process (HGEP) is constructed by making the base measure parameter of the rows shared and random. Formally, the model has the following form:

$$
\begin{aligned}
\mu_0 &\sim \mathrm{MGP}(H_0, \gamma_0) \\
\mu_\theta \mid \mu_0 &\overset{\text{iid}}{\sim} \mathrm{MGP}(\mu_0, \beta_0) \\
\theta_{N+1} \bigm| X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \bar{\mu}_{\theta_N} \\
J_{N+1} \bigm| X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \mathrm{Exp}(\|\mu_{\theta_N}\|).
\end{aligned}
\tag{5.1}
$$

In order to get a tractable predictive distribution, we introduce a set of auxiliary variables. As we will see shortly, these auxiliary variables are closely related to the variables used in the CRF metaphor presented in Section 3.1.3 to indicate when new *tables* are created in a given *restaurant* (i.e., m_{jk}).

5.3 Connection to CRF

From the connection between the DP and the gamma process described in Section 2.2.1, we know that if $X \sim DP(\alpha_0, H_0)$ and $Y \sim Gamma(\gamma_0, \beta_0)$ then $XY \sim MGP(\gamma_0 H_0, \beta_0)$. Hence, we can rewrite Definition 5.1 as:

$$
\begin{aligned}
\|\mu_0\| &\sim \mathrm{Gamma}(\|H_0\|, \gamma_0) && (5.2)\\
\bar{\mu}_0 &\sim DP(\|H_0\|, \bar{H}_0) && (5.3)\\
\|\mu_\theta\| \bigm| \|\mu_0\| &\overset{\text{iid}}{\sim} \mathrm{Gamma}(\|\mu_0\|, \beta_0) && (5.4)\\
\bar{\mu}_\theta \bigm| \bar{\mu}_0, \|\mu_0\| &\overset{\text{iid}}{\sim} DP(\|\mu_0\|, \bar{\mu}_0) && (5.5)\\
\theta_{N+1} \bigm| X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \bar{\mu}_{\theta_N} && (5.6)\\
J_{N+1} \bigm| X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \mathrm{Exp}(\|\mu_{\theta_N}\|). && (5.7)
\end{aligned}
$$

Note the similarity between Equations 5.3 and 3.1, and between Equations 5.5 and 3.2. The only difference is in Equation 5.5, where we condition on a random variable $\|\mu_0\|$, the normalization of the top level random measure. Now, as we have a similar structure to the HDP model, we can construct the predictive distribution for the states θ in the HGEP similarly to the CRF model of Section 3.1.3.

In the HGEP, a restaurant can be understood as a row in the rate matrix, the tables as groups of transitions to the same destination state, and a dish as a destination state. We can use an auxiliary variable A_n to indicate when the n-th transition creates a new table; $A_n = 1$ means the n-th transition is to a state not previously visited. The variable takes value $A_n = 0$ otherwise. We augment the sufficient statistics (i.e., F_θ and T_θ defined in Section 4.2) with empirical counts for the number of tables across all restaurants that share a given dish, $G = \sum_{n=1}^{N} A_n \delta_{\theta_n}$. In other words, $G(\{\theta\})$ indicates the number of transitions to the state θ from all the states.

We need to introduce one additional auxiliary variable, the normalization of the top level random measure, $\|\mu_0\|$. This auxiliary variable has no equivalent in CRFs.

Based on the construction of the CRF in Section 3.1.3 and the DP-based representation of the HGEP described above, we can write the predictive distribution for the state θ_{N+1} in the form of Equation 3.7:

$$\begin{aligned} \mu'' &= G + H_0 \\ \bar{\mu}'' &= \frac{G}{\|G+H_0\|} + \frac{\|H_0\|}{\|G+H_0\|}\bar{H}_0 \\ \theta_{N+1} \mid \theta_1, \cdots, \theta_N, \|\mu_0\|, \bar{\mu}'' &\sim \frac{F_{\theta_N}}{\|F_{\theta_N} + \mu_0\|} + \frac{\|\mu_0\|}{\|F_{\theta_N} + \mu_0\|}\bar{\mu}'' \end{aligned} \tag{5.8}$$

Note that $F_{\theta_N}(\{\theta_i\})$ is equivalent to $n_{N.i}$, the number of customers eating dish i in restaurant N, in Equation 3.7. In addition, $G(\{\theta_i\})$ is equivalent to $m_{.i}$, the number of tables serving dish i in the CRF model in Equation 3.7. We denote the predictive distribution for the next state presented in Equation 5.8 by $\bar{\mu}'^{(\mathrm{H})}_{\theta_N}$. We use the superscript (H) to distinguish from the non-hierarchical case.

Figure 5.1 illustrates an example of generating a state θ_{N+1} from the predictive distribution $\bar{\mu}'^{(\mathrm{H})}_{\theta_N}$. Assume that we are at the Nth jump and we have observed two distinct states a and b up to this time. The idea is to view the predictive distribution for the next state θ_{N+1} in the hierarchical model as a mixture of two possibilities. One of these two is selected with probability proportional to $(\|F_{\theta_N}\|, \|\mu_0\|)$. The two possibilities are:

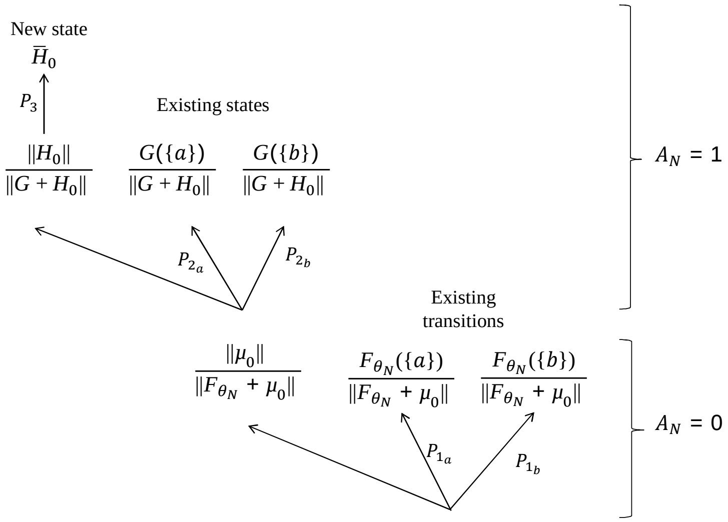

New state
$\overline{H}_0$
P_3
Existing states
$\frac{\|H_0\|}{\|G + H_0\|}$
$\frac{G(\{a\})}{\|G + H_0\|}$
$\frac{G(\{b\})}{\|G + H_0\|}$
P_{2_a}
P_{2_b}
$A_N = 1$
Existing
transitions
$\frac{\|\mu_0\|}{\|F_{\theta_N} + \mu_0\|}$
$\frac{F_{\theta_N}(\{a\})}{\|F_{\theta_N} + \mu_0\|}$
$\frac{F_{\theta_N}(\{b\})}{\|F_{\theta_N} + \mu_0\|}$
P_{1_a}
P_{1_b}
$A_N = 0$

Figure 5.1: An illustration of generating the next state given the current state θ_N in the HGEP model

1. to generate from $\frac{F_{\theta_N}}{\|F_{\theta_N}+\mu_0\|}$, the empirical distribution over the transitions starting at θ_N. This is "joining one of the existing tables in the current restaurant (i.e. the current state θ_N)" in the CRF analogy. In the example in Figure 5.1, this case corresponds to path P_{1_a} or P_{1_b}. When this alternative is selected, the successor state θ_{N+1} is determined with probability proportional to F_{θ_N} ("the new customer picks the dish of the selected existing table"). In the example, the next state is a or b with probability proportional to $F_{\theta_N}(\{a\})$ or $F_{\theta_N}(\{b\})$ correspondingly.

2. to generate recursively from a "back-off" distribution, $\bar{\mu}''$ ("creating a new table in the current restaurant"). When this alternative is selected, the new table picks a dish. The dish can be picked from G, the empirical distribution over the dishes picked by tables across all restaurants (corresponding to

paths P_{2_a} or P_{2_b}), or it can be picked from the "back-off" distribution, the normalized base measure, $\bar{H}_0$ (path P_3).

It remains to provide a formal definition for the table creation auxiliary variable A_n. By augmenting the sufficient statistics with an indicator over which of the two alternatives (1,2) is selected at each transition, the predictive distribution takes a tractable form. Formally, the definition of the table creation auxiliary variables is therefore as follows:

$$\mathbb{P}(A_{N+1} = a \mid \|\mu_0\|, X) \propto \|\mu_0\|^a \, \|F_{\theta_N}\|^{1-a} \, \mathbf{1}[a \in \{0,1\}]$$
$$\theta_{N+1} \bigm| A_{N+1}, X \sim (1 - A_{N+1})\bar{F}_{\theta_N} + A_{N+1}\bar{\mu}''.$$

5.4 Predictive Distribution

In the previous section we showed that the predictive distribution for the next state given the past states and $\|\mu_0\|$ has a CRF representation. In this section we present the predictive distribution for the next event given the auxiliary variables and the past events.

The main result of this section is as follows:

Proposition 17. *The predictive distribution of the HGEP is given by:*

$$(\theta_{N+1}, J_{N+1})\big|(X, \{A_n\}_{n=1}^N, \|\mu_0\|) \sim \bar{\mu}_{\theta_N}'^{(H)} \times \mathrm{TP}(\|\mu_{\theta_N}'^{(H)}\|, \beta_{\theta_N}').$$

Proof. Conditioning on $\|\mu_0\|$, we have that θ_{N+1} and J_{N+1} are independent. As mentioned $(\theta_{N+1}\big|X, \{A_n\}_{n=1}^N, \|\mu_0\|)$ can be viewed as an HDP with concentrations $\|H_0\|$ for the top level DP, and $\|\mu_0\|$ for the lower level DPs. Hence, the distribution $\bar{\mu}_{\theta_N}'^{(\mathrm{H})}$ is then the predictive distribution for θ_{N+1}.

For $J_{N+1}\big|X, \|\mu_0\|$, we know from Proposition 15 that the predictive distribution is a translated Pareto distribution with parameters equal to $\|F_{\theta_N} + \mu_0\|$, the normalization constant of the updated base measure, and $T_{\theta_N} + \beta_0$, the updated rate parameter. That is, $J_{N+1}\big|X, \|\mu_0\| \sim \mathrm{TP}(\|\mu_{\theta_N}'^{(\mathrm{H})}\|, \beta_{\theta_N}')$. □

The conditional distribution of the auxiliary variable $\|\mu_0\|$ given all the other random variables will be useful for the inference algorithm that we introduce in Chapter 6. It can be shown that, given all the other random variables, $\|\mu_0\|$ has a gamma distribution.

Proposition 18. *The conditional distribution of* $\|\mu_0\|$ *given the other variables is a gamma distribution:* $\|\mu_0\|\Big| X, \{A_n\}_{n=1}^N \sim \text{Gamma}(a,b)$, *where* $a = \|H_0 + G\|, b = \gamma_0 + \sum_{\theta\in\Omega}\log(\beta'_\theta/\beta_0)$.

Proof. The conditional distribution has density $p(x)$ proportional to:

$$\begin{aligned} p(\|\mu_0\| = x \mid (\theta_n, j_n)_{n=1}^N, \{A_n\}_{n=1}^N) \propto & p(\|\mu_0\| = x)\cdot p(j_1, j_2, \cdots, j_N|\|\mu_0\| = x) \\ & \cdot p(\theta_1, \cdots, \theta_N|\|\mu_0\| = x) \end{aligned}$$

For $p(\theta_1, \cdots, \theta_N|\|\mu_0\|)$, the joint density of the states given $\|\mu_0\|$, we must use the predictive distribution for the states given by Equation 5.8. Let i_ω be the set of indexes of the events for which we have $\theta_i = \omega$; $i_\omega = \{i : \theta_i = \omega\}$. We denote the size of the set $i_\omega = \{1_\omega, \cdots, N_\omega\}$ by $|i_\omega|$ and the number of times the transition from state ω results in a new table by G_ω. We have:

$$\begin{aligned} p(\theta_{1_\omega}, \cdots, \theta_{N_\omega}|\|\mu_0\| = x) &= p(\theta_{1_\omega}|\|\mu_0\| = x)p(\theta_{2_\omega}|\theta_{1_\omega}, \|\mu_0\| = x)\times \\ &\quad \cdots \times p(\theta_{N_\omega}|\theta_{1_\omega}, \cdots, \theta_{(N-1)_\omega}, \|\mu_0\| = x) \\ &\propto \frac{x^{G_\omega}}{(x)(x+1)\cdots(x+|i_\omega|-1)} \\ &\propto \frac{x^{G_\omega}}{(x)_{\|F_{\theta\omega}\|}} \end{aligned}$$

Therefore, for the sequence of states $\theta_1, \cdots, \theta_N$ we have:

$$\begin{aligned} p(\theta_1, \cdots, \theta_N|\|\mu_0\| = x) &\propto \prod_{\theta\in\Omega} \frac{x^{G_\theta}}{(x)_{\|F_\theta\|}} \\ &\propto x^{\|G\|} \prod_{\theta\in\Omega} \left((x)_{\|F_\theta\|}\right)^{-1} \end{aligned}$$

Finally, using Equations 5.2, 4.8, and 5.8, for $p(x)$ we have:

$$
\begin{aligned}
p(x) \propto \left(x^{\alpha_0 - 1}\exp(-\gamma_0 x)\right)\left(\prod_{\theta\in\Omega}\frac{(x)_{\|F_\theta\|}\beta_0^x}{(T_\theta+\beta_0)^{x+\|F_\theta\|}}\right) \\
\times\left(x^{\|G\|}\prod_{\theta\in\Omega}\left((x)_{\|F_\theta\|}\right)^{-1}\right) \\
\propto \left(x^{\alpha_0+\|G\|-1}\right)\exp\left(-x\left(\gamma_0+\sum_{\theta\in\Omega}\log\left(\beta'_\theta/\beta_0\right)\right)\right)
\end{aligned}
$$

□

Chapter 6

INFERENCE

In the previous two chapters we assumed the sequence of events is observed. Now we relax this assumption and use the models described in the last two chapters as a prior distribution over a sequence of hidden events. The goal of this chapter is to propose an inference algorithm to approximate the posterior of the hidden events X given the observations $\mathscr{Y}$. In particular, we are interested in approximating $\mathbb{E}[h(X)|\mathscr{Y}]$, the expectation of a general function of the events X under the posterior distribution of GEPs.

In most applications, the sequence of states is not directly nor fully observed. First, instead of observing the states θ, we observe $\mathscr{X}$-valued random variables Y_n distributed according to a parametric family $\mathscr{P}$ indexed by the states θ of the chain, $\mathscr{P} = \{L_\theta : \mathscr{F}_\mathscr{X} \to [0,1], \theta \in \Omega\}$. Second, the observations are generally available only for a finite set of times $\mathscr{T}$. For instance, in Figure 2.1 instead of observing the MJP denoted by X, we only have observations at time points $t_1, \cdots, t_G$ (we will discuss multiple sequences in the next section). To specify the random variables in question, we need a notation for the hidden event index at a given time t, $I(t) = \min\left\{N : \sum_{n=1}^{N+1} J_n > t\right\}$ (see Figure 2.1, where $I(t^*) = N-1$). Thus, assuming conditional independence, for an individual observation at time point t we have $Y(t)|X \overset{d}{=} Y(t)|\theta_{I(t)} \sim L_{\theta_{I(t)}}$. The set of all observed random variables is then defined as

$$\mathscr{Y} = (Y(t_1), Y(t_2), \ldots, Y(t_G) : t_g < t_{g+1}, \{t_i\} = \mathscr{T}).$$

We then denote the list of hidden events from time $t = 0$ to t_G by $X_{1:G} = (\theta_n, J_n)_{n=1}^{I(t_G)}$ where $t_G \in \mathscr{T}$.

We will describe the general inference framework for the model of Chapter 4. Extension to hierarchical models is direct (by keeping track of an additional sufficient statistic G, as well as the auxiliary variables $A_n, \|\mu_0\|$).

For simplicity, we assume in this section that $\mathscr{P}$ is a conjugate family with respect to H_0. Non-conjugate models can be handled by incorporating the auxiliary variables of Algorithm 8 in [33].

As mentioned our goal is to approximate expectations under the posterior distribution of GEPs. However, as the posterior distribution is intractable we cannot evaluate posterior expectations directly. Thus, we need to rely on the Monte Carlo methods and sample from approximations to the posterior distribution of the hidden events given the observations; that is, $p(X_{1:G}|\mathscr{Y})$.

Commonly used algorithms for sampling (approximately) from a posterior distribution in Bayesian statistics are based on **Markov chain Monte Carlo (MCMC)** methods. In order to *approximately* sample from the posterior distribution for our model we use a **particle Markov chain Monte Carlo (PMCMC)** algorithm [2], a special type of the MCMC algorithm well-suited for sampling sequential data.

6.1 MCMC Algorithms

In this section we briefly review the MCMC methods; for a detailed discussion and proofs of the results see for example [28]. MCMC is a general approach for generating samples from a target distribution $p(x)$ using a Markov chain mechanism. The samples $x(i)$ generated from an MCMC based algorithm form a Markov chain that converges to $p(x)$. We use MCMC when we cannot directly sample from $p(x)$ but we can evaluate it up to a normalizing constant.

Recall that the evolution of a Markov chain depends only on its current state and a probability transition matrix which we denote by T. As long as the Markov chain is *irreducible* and *aperiodic* there is a distribution $p(x)$ called the *invariant* distribution to which the Markov chain converges. Informally, an aperiodic irreducible Markov chain is a Markov chain that does not have a cyclic behavior and from any of its states there is a positive probability that all other states can be visited in finite number of steps[1].

The **detailed balance** condition, a condition that guaranties $p(x)$ is the invariant distribution for a Markov chain, is defined as:

$$p(x(i))\;T(x(i-1)|x(i)) = p(x(i-1))\;T(x(i)|x(i-1));$$

an MCMC sampler can be obtained by satisfying the detailed balance condition.

The most popular MCMC algorithm is the **Metropolis - Hastings (MH)** algorithm; most MCMC algorithms can be viewed as a special case or an extension of this algorithm. In every step of this algorithm, given the current state of the Markov chain x^c, a candidate state x^* is sampled from a proposal distribution $q(\cdot|x^c)$. Then x^* is accepted with the probability:

$$min\{1, \frac{p(x^*)q(x^c|X^*)}{p(x^c)q(x^*|x^c)}\};$$

if x^* is accepted the Markov chain moves to x^*, otherwise it remains at x^c. Note that we only need $p(x)$ up to a normalizing constant in order to compute the acceptance probability. Although the MH algorithm is simple, finding a good proposal distribution that has fast convergence rate can be a challenging task. The MH algorithm converges if we can show that the chain has no cycles and every state that has positive probability can be reached in a finite number of steps. Under these conditions the chain converges to the target distribution $p(x)$ after sufficient **burn-in period** which is the number of steps until the chain approaches $p(x)$.

In this definition every sample is dependent on the previous sample; as a result we have dependence between draws in the Markov chain. To reduce this dependence and get closer to i.i.d. samples, some have suggested **thinning**: keeping only the samples generated after every r-th iteration of the chain.

A variant of the MH algorithm is the **independent MH** algorithm for which the proposal is independent of the current state, $q(x^*) = q(x^*|x^c)$. Thus, the acceptance probability is $min\{1, \frac{p(x^*)q(x^c)}{p(x^c)q(x^*)}\}$.

Another MCMC algorithm that can be considered as a special case of the MH

[1] For a formal definition see for example [19].

is the **Gibbs sampler**. It can be shown that the Gibbs sampler is an MH algorithm with acceptance probability equal to 1. Suppose the goal is sampling from the n-dimensional joint probability distribution $p(x_1, x_2, \cdots, x_n)$ and sampling from this distribution is hard. If samples from the full conditionals $p(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$ can be easily obtained then we can produce samples from the joint distribution by sampling from the conditional distributions. We initialize the Gibbs sampler with an arbitrary sample $(x_1(0), x_2(0), \cdots, x_n(0))$. Then for iterations $i = 1$ to N the algorithm consists of n steps as follows:

$$\begin{aligned}
&\text{Sample } x_1(i) \sim p(x_1|x_2(i-1), x_3(i-1), \cdots, x_n(i-1))\\
&\text{Sample } x_2(i) \sim p(x_2|x_1(i-1), x_3(i-1), \cdots, x_n(i-1))\\
&\qquad\qquad \vdots\\
&\text{Sample } x_k(i) \sim p(x_k|x_1(i-1), x_2(i-1), \cdots, x_{k-1}(i-1), x_{k+1}(i-1), \cdots, x_n(i-1))\\
&\qquad\qquad \vdots\\
&\text{Sample } x_n(i) \sim p(x_n|x_1(i-1), x_2(i-1), \cdots, x_{n-1}(i-1))
\end{aligned}$$

With any MCMC sampler, the empirical distribution of the generated samples yields $\hat{p}(x)$, the approximation to the target distribution $p(x)$; in Bayesian contexts, the target distribution will be a posterior distribution. In addition to approximating the posterior distribution $p(x)$, the samples obtained from an MCMC sampler can be used to approximate the mode (or other characteristics) of the posterior distribution. For instance, to approximate the maximum a posteriori (MAP), if we have N samples, we find

$$\underset{x(i); i=1,\cdots,N}{\operatorname{argmax}} \; (\hat{p}(x(i))).$$

In applications we usually only consider the samples after the burn-in period.

The theory guarantees that the chain converges to the target distribution, but it does not say anything about how long we need to run the chain to obtain convergence. In particular, there are two practical issues that we need to address when using an MCMC sampler: how long should the burn-in period be and how long we need to run the chain after that. Different diagnostics are available for monitoring the chain's convergence. The easiest diagnostic is to simply visualize the state of the chain with respect to iterations; if there is no evident trend in the plot this suggests (approximate) convergence has been achieved. There are also some quantitative diagnostics such as **Geweke's** diagnostic [15], **Heidelberger and Welch's** diagnostic [21], and **Gelman and Rubin's** diagnostic [14]. Although these diagnostics are informative guides, they cannot definitely tell whether the chain has converged; judging the adequacy of convergence of MCMC samplers necessarily retains a subjective element.

If we are convinced that the chain has converged, then we need to know how long we need to run the chain to get a reasonable approximation. The answer depends on the objective of the study: the greater the precision needed for the study, the longer we need to run the chain. We can use **Raftery and Lewis's** diagnostic [37] to estimate how long our chain needs to run in order to estimate quantiles within a specified accuracy with some specified probability.

6.2 PMCMC Algorithm

As mentioned we use a special type of the MCMC algorithm, the PMCMC algorithm, to approximately sample from the posterior distribution in our model. There are different flavors of PMCMC algorithms; we use the particle independent Metropolis - Hastings (PIMH) algorithm [2], an MCMC algorithm with proposals that are **sequential Monte Carlo (SMC)** approximations of $p(X_{1:G}|\mathscr{Y})$. We will explain the SMC algorithm in detail in the following section; for now it can be thought of as an algorithm which provides an approximation for the posterior density $p(X_{1:G}|\mathscr{Y})$ and the marginal likelihood $p(\mathscr{Y})$.

As mentioned to sample from $p(X_{1:G}|\mathscr{Y})$ in a standard MCMC algorithm we sample from a proposal density $q(X_{1:G}|\mathscr{Y})$. Given a current state $X^c_{1:G}$, the proposed sample $X^*_{1:G}$ is then accepted with probability:

$$min\{1, \frac{p(X^*_{1:G}|\mathscr{Y})q(X^c_{1:G}|\mathscr{Y})}{p(X^c_{1:G}|\mathscr{Y})q(X^*_{1:G}|\mathscr{Y})}\}.$$

In the PMCMC algorithm we use an SMC approximation to $p(X_{1:G}|\mathscr{Y})$ which we denote by $\hat{p}(X_{1:G}|\mathscr{Y})$ as our proposal density. Note that *at each iteration of the PMCMC algorithm*, the sample from the $\hat{p}(X_{1:G}|\mathscr{Y})$ is a list of hidden events from

t_1 to t_g. It can be shown [2] that using an appropriate acceptance ratio we can make this proposal a valid MCMC move; the algorithm then converges to the desired posterior density $p(X_{1:G}|\mathscr{Y})$. The PMCMC algorithm follows the following simple form:

1. iteration $i = 0$:

   Run an SMC algorithm to obtain an approximation for $p(X_{1:G}|\mathscr{Y})$; sample $X_{1:G}(0) \sim \hat{p}(\cdot|\mathscr{Y})$ and denote the corresponding marginal likelihood estimate $\hat{p}(\mathscr{Y})$ by $L(0)$.

2. iteration $i \geq 1$:

   (a) run the SMC algorithm and sample $X^*_{1:G} \sim \hat{p}(\cdot|\mathscr{Y})$, and denote the corresponding marginal likelihood estimate by L_*

   (b) with probability $min\{1, \frac{L_*}{L(i-1)}\}$ accept $X^*_{1:G}$: set $X_{1:G}(i) = X^*_{1:G}$ and $L(i) = L_*$. Otherwise, set $X_{1:G}(i) = X_{1:G}(i-1)$ and $L(i) = L(i-1)$.

Denoting the total number of iterations by *iter* and the number of burn-in iterations by B, the collection of samples $\{X_{1:G}(i)\}_{i=B+1}^{\text{iter}}$ are approximately samples from the posterior density $p(X_{1:G}|\mathscr{Y})$ for large enough B (i.e., sufficient burn-in period). Table 6.1 summarizes the notation used in the PMCMC and the SMC algorithms.

In general, there may be several sequences of observations; for instance, we can have sequences of EDSS values from several MS patients. We can have different observation times for different sequences. We denote the number of time series by K, each of the form

$$\mathscr{Y}^{(k)} = \Big(Y^{(k)}(t_1^{(k)}), Y^{(k)}(t_2^{(k)}), \ldots, Y^{(k)}(t_{G^{(k)}}^{(k)}) : t_g^{(k)} < t_{g+1}^{(k)}, \{t_i^{(k)}\} = \mathscr{T}^{(k)}\Big), \quad k \in \{1, \ldots, K\};$$

moreover, we use the superscript $\backslash k$ to indicate all the sequences except k.

In case of multiple sequences, we resample the hidden events $X^{(k)}_{1:G^{(k)}}$ for one sequence k given the sufficient statistics of all the other sequences $(F_\theta^{(\backslash k)}, T_\theta^{(\backslash k)})$, and the sufficient statistics for the likelihood model denoted by $S_\theta^{(\backslash k)}$. The role of the sufficient statistics in the algorithm and their updating procedure will be explained in the next section where we describe the SMC algorithm. To simplify

Notation	Definition
$I(t)$	the hidden event index at a given time t: $I(t) = \min\left\{N : \sum_{n=1}^{N+1} J_n > t\right\}$
$X_{1:g}$	the list of hidden events from time $t = 0$ to t_g: $X_{1:g} = (\theta_n, J_n)_{n=1}^{I(t_g)}$
$Y(t)$	an observation at time t
$Y_{1:g}$	the observations from time t_1 to time t_g: $Y(t_1), Y(t_2), \ldots, Y(t_g)$
$X_{m,g}$	a particle, a list of hidden events indexed by n covering the first g observations: $X_{m,g} = (\theta_{m,n}, J_{m,n})_{n=0}^{N_{m,g}}$
$N_{m,g}$	the smallest number of events required to cover the first g observations in the m-th particle: $t_g \leq \sum_{n=0}^{N_{m,g}} J_{m,n}$
$X'_{m,g+1}$	a new particle extended from $X_{m,g}$
$\tilde{X}^{m}_{g+1}$	an event (a pair of state and waiting time) sampled from the proposal $q(\cdot\|Y(t_{g+1}), X'_{m,g+1})$
X^{m}_{g+1}	a list of events appended to $X_{m,g}$
$W_{m,g}$	the importance weight of the particle $X_{m,g}$
$F^{k}_{\theta,m,n}$	empirical transition measures for the m-th particle after the n-th jump in the k-th sequence
$T^{k}_{\theta,m,n}$	empirical waiting times for the m-th particle after the n-th jump in the k-th sequence
$G^{k}_{m,n}$	empirical counts for the number of transitions from all states to the same destination after the n-th jump for the m-th particle in the k-th sequence
$S^{k}_{\theta,m,n}$	sufficient statistics of the likelihood model for the m-th particle after the n-th jump in the k-th sequence
$\backslash k$	all the sequences except the k-th (used as superscript for the sufficient statistics)

Table 6.1: Notation used in the PMCMC algorithm

the notation we denote $X^{(k)}_{1:G^{(k)}}$ by $X^{(k)}$. The pseudocode for the PMCMC algorithm in the GEP model (with multiple sequences) is provided in Figure 6.1.

6.3 SMC Algorithm

SMC algorithms provide an approximation of $p(X_{1:G}|\mathscr{Y})$, the posterior density, and $p(\mathscr{Y})$, the marginal likelihood [2]. These algorithms approximate the pos-

Algorithm 1 PMCMC algorithm for inference in the GEP model

```
$K$: number of sequences
$k$: sequence number $k \in \{1,\dots,K\}$

For $i = 0$ to #iterations do
  For $k = 1$ to $K$ do
    - Run an SMC algorithm to obtain an approximation for $p(X^{(k)}|\mathscr{Y}^{(k)}, F_\theta^{(\backslash k)}, T_\theta^{(\backslash k)}, S_\theta^{(\backslash k)})$
    - Sample $X_*^{(k)} \sim \hat{p}(\cdot|\mathscr{Y}^{(k)}, F_\theta^{(\backslash k)}, T_\theta^{(\backslash k)}, S_\theta^{(\backslash k)})$, and estimate the corresponding marginal likelihood $L_*^{(k)}$
    - Sample $u \sim Unif[0,1]$
    If $u < \min\{1, \frac{L_*^{(k)}}{L^{(k)}(i-1)}\}$ or $i = 0$
      - Set $X^{(k)}(i) = X_*^{(k)}$, $L^{(k)}(i) = L_*^{(k)}$
      - Update $(F_\theta^{(k)}, T_\theta^{(k)}, S_\theta^{(k)})$
    else
      - Set $X^{(k)}(i) = X^{(k)}(i-1)$ and $L^{(k)}(i) = L^{(k)}(i-1)$
    End If
  End For
End For
```

Figure 6.1: Pseudocode for the PMCMC algorithm

terior density $p(X_{1:G}|\mathscr{Y})$ sequentially[2]. In other words, a SMC algorithm first approximates $p(X_{1:1}|Y(t_1))$, then $p(X_{1:2}|Y(t_1),Y(t_2))$ and so on. We denote the observations from time t_1 to time t_g (i.e., $Y(t_1),Y(t_2),\dots,Y(t_g)$) by $Y_{1:g}$.

Given the set of all observations $Y_{1:G}$, the approximation for the posterior is obtained from a set of M weighted random samples:

$$\hat{p}(dX_{1:G}|Y_{1:G}) := \sum_{m=1}^{M} W_{m,G}\delta_{X_{m,G}}(dX_{1:G}), \tag{6.1}$$

where $W_{m,G}$ is the importance weight of the random sample $X_{m,G}$. Each random sample $X_{m,g}$, $m \in \{1,\dots,M\}$ is called a **particle**; it consists of a list of hidden events indexed by n, containing both (hidden) states and waiting times: $X_{m,g} = (\theta_{m,n}, J_{m,n})_{n=0}^{N_{m,g}}$. We use $N_{m,g}$ to denote the smallest number of events required to cover the first g observations in the m-th particle; that is, $t_g \leq \sum_{n=0}^{N_{m,g}} J_{m,n}$. For $g = 0$ we let $N_{m,0} = 0$.

The algorithm propagates the particles $X_{m,g}$ and updates the weights $W_{m,g}$ from generation $g = 0$ up to generation G, where generation g corresponds to the obser-

[2] To simplify the description of the algorithm we begin by explaining it for the case where there is only one sequence; next, we generalize the algorithm for the case of multiple sequences

vation time t_g. A proposal density q is used to propagate the particles and evaluate the weights. Two different proposal densities will be introduced in Section 6.4; evaluation of the weights for each of the proposals will also be discussed.

Initially, we assume all the particles in generation $g = 0$ (i.e., time $= 0$) are set to contain only the beginning special state, $X_{m,0} = (\theta_{\text{beg}}, 0)$ for all $m \in \{1, \ldots, M\}^3$.

For generation $g = 1$, a proposal (i.e., importance) density $q(\cdot|Y(t_1))$ is used to approximate $p(X_{1:1}|Y_{1:1})$. We generate M particles from $q(\cdot|Y(t_1))$; each particle $X_{m,1}$ will be constructed such that the list of events contains enough events (i.e., smallest number of events) to cover the first observation time, t_1. In other words, $t_1 \leq \sum_{n=1}^{N_{m,1}} J_{m,n}$.

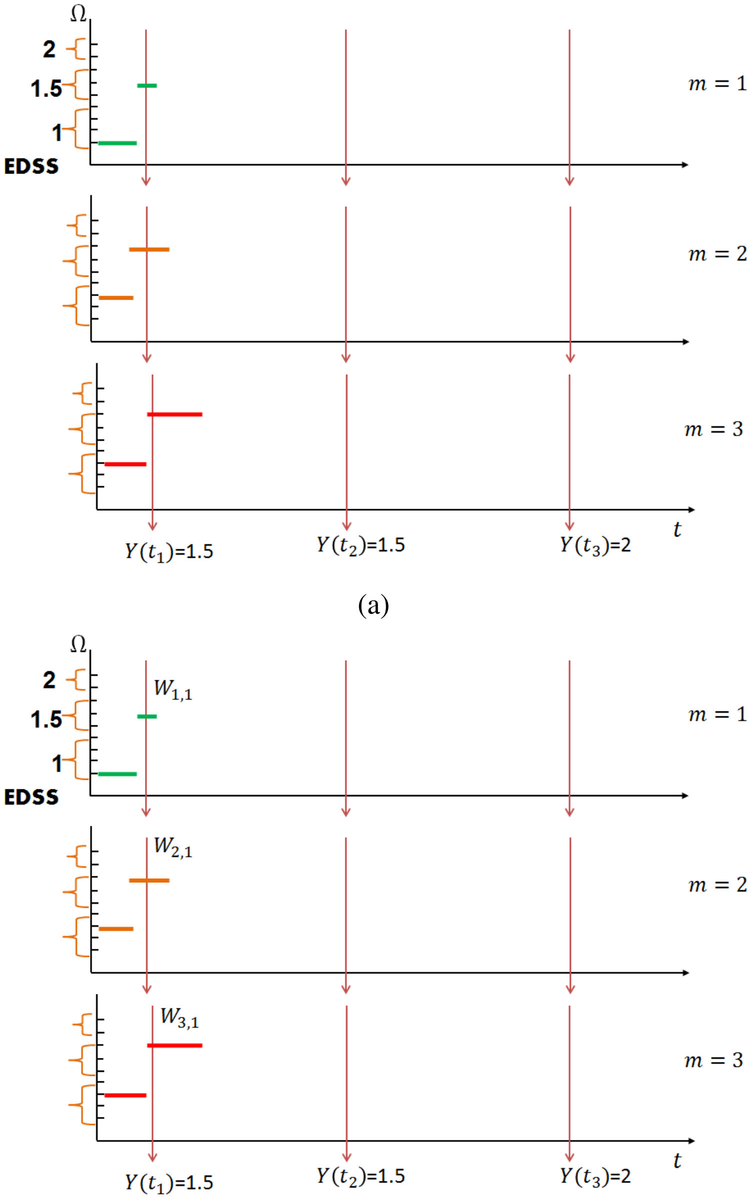
An illustration of the SMC algorithm is provided in Figures 6.2, 6.3, and 6.4 where we have three observations (i.e., EDSS$= 1.5, 1.5, 2$ at time points t_1, t_2, t_3). The first generation of three particles at observation time t_1 is shown in Figure 6.2(a).

In order to take into account the discrepancy between the two densities we assign importance weights $W_{m,1}$ to the particles (details to follow). This is shown in Figure 6.2(b) where we assign weights $W_{1,1}$, $W_{2,1}$, and $W_{3,1}$ to the particles. Next, in a resampling step, we sample M times from the weighted particle density $\hat{p}(dX_{1:1}|Y_{1:1})$ which is an approximation of $p(X_{1:1}|Y_{1:1})$. In Figure 6.3(a) we assumed we have sampled three times from the weighted particle density and obtained the same particles as before; we have three particles $X_{1,1}$, $X_{2,1}$, and $X_{3,1}$.

At time t_2, we extend each particle $X_{m,1}$ into a new particle $X'_{m,2}$ through the proposal density $q(\cdot|Y(t_2), X_{m,1})$ to obtain an approximation for $p(X_{1:2}|Y_{1:2})$. To extend a particle $X_{m,1}$ into a particle $X'_{m,2}$, we first copy the events of $X_{m,1}$ into $X'_{m,2}$, and then append new sampled events (from $q(\cdot|Y(t_2), X_{m,1})$) until the observation at time t_2 is covered. Figure 6.3(b) shows the extended particles $x'_{1,2}$, $x'_{2,2}$, and $x'_{3,2}$.

We need to compute the importance weights $W_{m,2}$ for the extended particles shown in Figure 6.3(b) as we are not directly sampling from $p(X_{1:2}|Y_{1:2})$. A new population of extended particles $(X_{m,2})_{m=1}^{M}$ is then obtained by resampling M times from $\hat{p}(dX_{1:2}|Y_{1:2})$, where

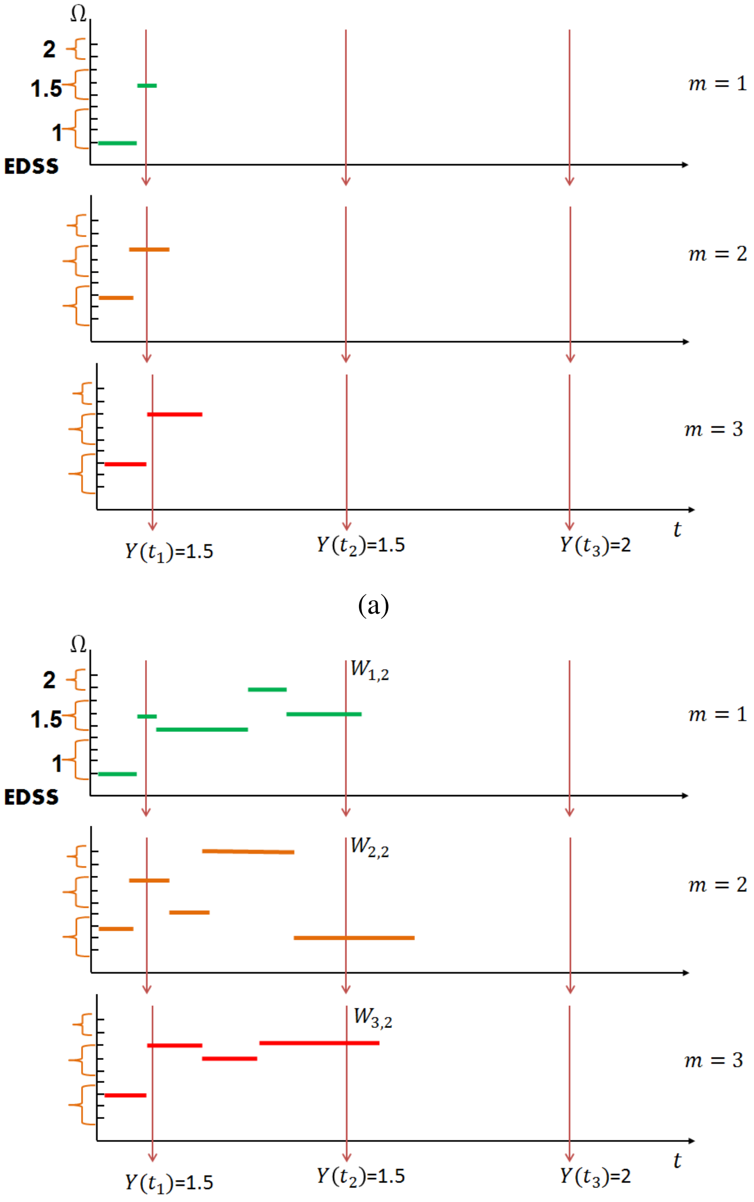
[3] As this event is present at the beginning of each particle, whenever $g > 0$ we initialize the sequence of hidden events for each particle from the first event $n = 1$. *That is, for $g > 0$ we define each particle as $X_{m,g} = (\theta_{m,n}, J_{m,n})_{n=1}^{N_{m,g}}$.*



(a)

(b)

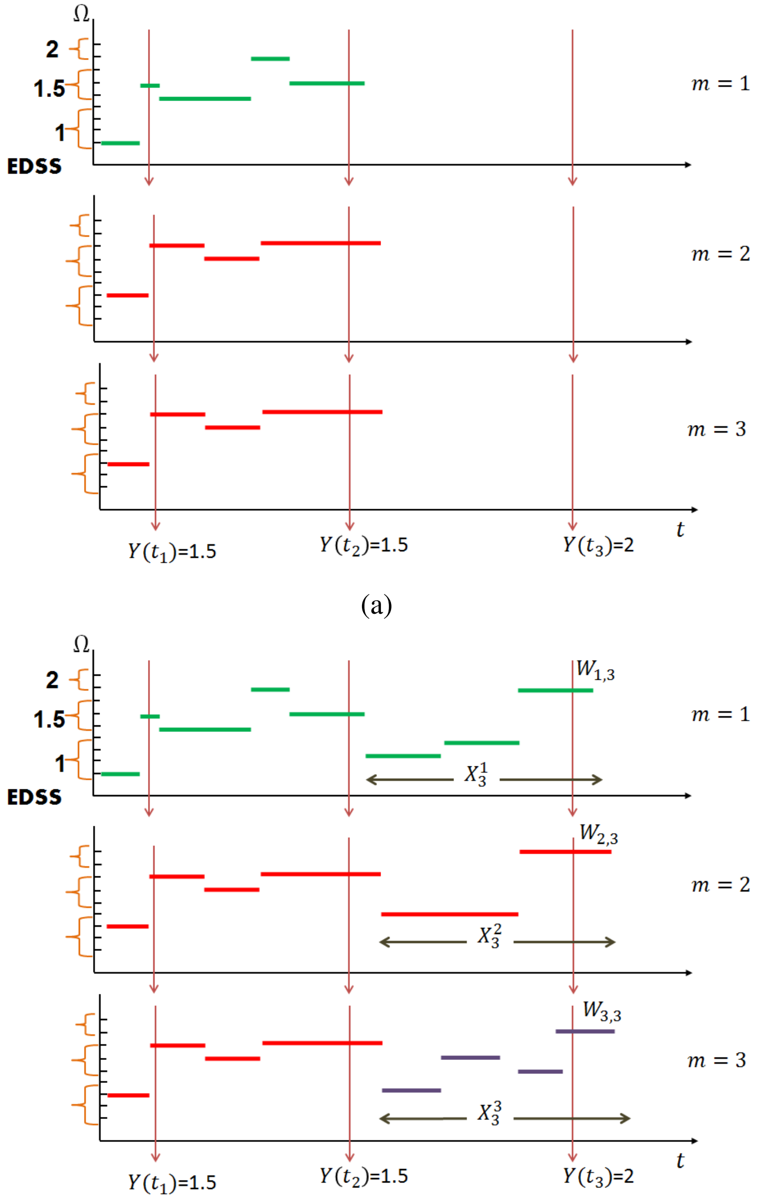
Figure 6.2: An illustration of the SMC algorithm with three particles and three observation times up to generation 1.



(a)

(b)

Figure 6.3: An illustration of the SMC algorithm with three particles and three observation times up to generation 2.



(a)

(b)

Figure 6.4: An illustration of the SMC algorithm with three particles and three observation times up to generation 3.

$$\hat{p}(dX_{1:2}|Y_{1:2}) = \sum_{m=1}^{M} W_{m,2}\delta_{X'_{m,2}}(dX_{1:2}).$$

The new population of random particles $X_{1,2}$, $X_{2,2}$, and $X_{3,2}$ is illustrated in Figure 6.4(a) where we have two samples from $X'_{3,2}$ and one sample from $X'_{1,2}$. This procedure is then repeated until time t_G is covered; for instance, in Figure 6.4(b) we stop the procedure as soon as the observation time t_3 is covered. This completes generating the particles; the approximation for the posterior distribution $p(X_{1:G}|Y_{1:G})$ is then obtained from $\hat{p}(dX_{1:G}|Y_{1:G})$ (e.g., $\sum_{m=1}^{M} W_{m,3}\delta_{X'_{m,3}}(dX_{1:3})$ for the example in Figure 6.4(b)).

The SMC algorithm can be summarized as below (for details see [2]). To simplify the notation, we omit writing $\forall m \in \{1,\ldots,M\}$; we also denote the sufficient statistics for the likelihood model $\mathscr{P}$ by S_θ.

1. at time $= 0$ set $X_{m,0} = (\theta_{\text{beg}}, 0)$ and $W_{m,0} = 1/M$.

2. for $g = 0$ to $G-1$

   (a) extend $X_{m,g}$ to a new particle $X'_{m,g+1}$:

      i. copy the contents of $X_{m,g}$ into $X'_{m,g+1}$ and create X^m_{g+1} a list of events that have been appended to $X_{m,g}$. Note that the list X^m_{g+1} contains no events at this stage as we have only copied the contents of $X_{m,g}$ into $X'_{m,g+1}$ (see Figure 6.4(b) for an example of the list of appended events to the particles $X_{1,2}$, $X_{2,2}$, and $X_{2,3}$: X^1_3, X^2_3 and X^3_3).

      ii. sample from $q(\cdot|Y(t_{g+1}), X'_{m,g+1})$, append the new sampled event $\tilde{X}^m_{g+1}$ to $X'_{m,g+1}$ and also to X^m_{g+1} (i.e., $X'_{m,g+1} := (X'_{m,g+1}, \tilde{X}^m_{g+1})$ and $X^m_{g+1} := (X^m_{g+1}, \tilde{X}^m_{g+1})$). Finally, update the sufficient statistics (e.g., F, T and S for the GEP model). Continue appending new events, and updating the sufficient statistics; stop as soon as the first $g+1$ observations are covered; that is, $t_{g+1} \leq \sum_{n=1}^{N_{m,g}} J_{m,n}$.

   (b) compute and normalize the weights for the extended particles; to do this, for an observation $\mathrm{d}y \in \mathscr{F}_{\mathscr{X}}$, let $L(\mathrm{d}y|S_\theta)$ denote the predictive likelihood given S_θ. It can be shown [2] that the appropriate importance weights are:

$$w_{m,g+1} = \frac{L(\,\mathrm{dy}|S_\theta)p(X^m_{g+1}|X_{m,g})}{q(X^m_{g+1}|Y(t_{g+1}),X_{m,g})}; \tag{6.2}$$

then the normalized weights are:

$$W_{m,g+1} := \frac{w_{m,g+1}}{\sum_{n=1}^{M} w_{n,g+1}}.$$

(c) after obtaining extended particles $X'_{m,g+1}$ and their weights, generate the new generation of particles $X_{m,g+1}$ by resampling M times from

$$\hat{p}(dX_{1:g+1}|Y_{1:g+1}) = \sum_{m=1}^{M} W_{m,g+1}\delta_{X'_{m,g+1}}(dX_{1:g+1}). \tag{6.3}$$

3. the approximation for the posterior density $p(X_{1:G}|\mathscr{Y})$ is then:

$$\hat{p}(dX_{1:G}|Y_{1:G}) := \sum_{m=1}^{M} W_{m,G}\delta_{X_{m,G}}(dX_{1:G}). \tag{6.4}$$

The SMC algorithm also provides an estimate of the marginal likelihood $p(\mathscr{Y})$ given by

$$\hat{p}(\mathscr{Y}) = \prod_{g=1}^{G} \frac{1}{M} \sum_{n=1}^{M} w_{n,g}. \tag{6.5}$$

As mentioned we use the marginal likelihood estimates in computing the acceptance probability in the PMCMC algorithm.

For the case of multiple sequences we need to keep the track of the sufficient statistics for each sequence. In the GEP model, for the k-th sequence the sufficient statistics are the empirical transition measures (Equation 4.1), the empirical waiting times (Equation 4.2), and the sufficient statistics of the likelihood model for the m-th particle after the n-th jump; we denote these sufficient statistics by $F^{(k)}_{\theta,m,n}$, $T^{(k)}_{\theta,m,n}$, and $S^{(k)}_{\theta,m,n}$ correspondingly. Note that the sufficient statistics for the likelihood

```
Algorithm 2 SMC algorithm to construct a proposal for the PMCMC

$G$: the number of observations in the current sequence $(k)$
$g$: particles generation $g \in \{0,1,\dots,G\}$
$M$: number of particles
$m$: particle number $m \in \{1,\dots,M\}$
$K$: the total number of sequences
$k$: sequence number $k \in \{1,\dots,K\}$
(We omit writing $\forall m \in \{1,\dots,M\}$ and superscript $(k)$ for every $X$ to avoid excessive notation)

For a given sequence $k$:
- Set $X_{m,0} = (\theta_{\text{beg}}, 0)$
For $g = 0$ to $g = G-1$ do
  - Extend $X_{m,g}$ to a new particle $X'_{m,g+1}$:
    Copy the events of $X_{m,g}$ into $X'_{m,g+1}$ and create $X^m_{g+1}$ a list of events appended to $X_{m,g}$
    Loop until $t^{(k)}_{g+1} \leq \sum_{n=1}^{N_{m,g+1}} J_{m,n}$ is satisfied
      sample a pair $(\theta, J)$ from $q(\cdot|Y(t_{g+1}), X'_{m,g+1})$
      append the new sampled event $\tilde{X}^m_{g+1}$ to $X'_{m,g+1}$; let $X'_{m,g+1} := (X'_{m,g+1}, \tilde{X}^m_{g+1})$
      append the new sampled event $\tilde{X}^m_{g+1}$ to $X^m_{g+1}$; let $X^m_{g+1} := (X^m_{g+1}, \tilde{X}^m_{g+1})$
      update the sufficient statistics $F^{(k)}_{\theta,m,n}$, $T^{(k)}_{\theta,m,n}$ and $S^{(k)}_{\theta,m,n}$ for events that have been generated so far
    End Loop
  - Compute the weight of the particles $w_{m,g+1} = \frac{L(Y(t_{g+1})|\theta_{I(t_{g+1})}, S^{(\backslash k)}_{\theta} + S^{(k)}_{\theta,m,n}) p(X^m_{g+1}|X_{m,g})}{q(X^m_{g+1}|Y(t_{g+1}), X_{m,g})}$
  - Generate the new population of particles $X_{m,g+1}$ by resampling $M$ times from $\hat{p}(dX_{1:g+1}|Y_{1:g+1})$ (Equation (6.3))
End For
- Approximate the posterior density $p(X_{1:G}|\mathscr{Y})$ and the marginal likelihood $p(\mathscr{Y})$ with Equations 6.4
and 6.5 correspondingly.
```

Figure 6.5: Pseudocode for the SMC step of the PMCMC algorithm

model depend on the chosen likelihood model; for instance, in the likelihood model that we will introduce in Chapter 7 the sufficient statistics are the frequencies of each observation. The pseudocode of the SMC algorithm in the GEP model (in case of multiple sequences) is presented in Figure 6.5.

6.4 SMC Algorithm Proposal Distributions

We describe two proposal distributions for the SMC algorithm in this section: the predictive distribution and a modified version of the predictive distribution.

6.4.1 The predictive distribution as the proposal distribution

If we use the predictive distribution as the proposal q in the SMC algorithm of Figure 6.5, then to extend the particle $X_{m,g}$ to $X'_{m,g+1}$ we need to sample the hidden event $\tilde{X}^m_{g+1}$ from Equation 4.7.

The sufficient statistics for this sampling step come from two sources: (1) the other sequences, which are held fixed (i.e., $(F_\theta^{(\backslash k)}, T_\theta^{(\backslash k)})$), and (2) the events that have been generated so far in the current particle, $X'_{m,g+1,n} := (\theta_{m,i}, J_{m,i})_{i=1}^{n}$. The corresponding sufficient statistics are denoted by $(F_{\theta,m,n}^{(k)}, T_{\theta,m,n}^{(k)})$.

Since these statistics are additive, we can write the sampling step as:

$$(\theta_{n+1}, J_{n+1}) | F_\theta^{(\backslash k)}, T_\theta^{(\backslash k)}, X'_{m,g+1,n} \sim \bar{\mu}'_{\theta_n} \times \mathrm{TP}(\|\mu'_{\theta_n}\|, \beta'_{\theta_n}),$$

where: $\mu'_\theta = F_\theta^{(\backslash k)} + F_{\theta,m,n}^{(k)} + H_0$ and $\beta'_\theta = T_\theta^{(\backslash k)} + T_{\theta,m,n}^{(k)} + \beta_0$. The weights of the particles have the simple form $w_{m,g+1} = L(\,\mathrm{d}y | S_\theta^{(\backslash k)} + S_{\theta,m,n}^{(k)})$ as the proposal $q(X_{g+1}^{m} | Y_{g+1}, X_{m,g})$ is the same as the predictive distribution $p(X_{g+1}^{m} | X_{m,g})$.

6.4.2 An “improved” proposal distribution based on a modified predictive distribution

For our experiments discussed in Chapter 8, we use a modified version of the hierarchical model of Chapter 5 and a relatively simple likelihood model. Therefore, in this section as opposed to the previous sections of this chapter we describe the proposal distribution for the HGEP model, which is used in our experiments. For the likelihood model, we assume each hidden state θ emits an observation deterministically; that is, given the hidden state the observation is known. However, multiple hidden states can emit the same observation.

More formally, we assume H_0 in the HGEP model of Section 5.2 is a product of uniform and multinomial distributions, (i.e., $H_0 = \mathrm{Unif} \times \mathrm{Mult}$). Moreover, for the likelihood model L_θ we use a Dirac delta. Each sample from H_0 is a hidden state θ in the form of a pair (u, y) where $y \in \Sigma$ the set of possible observations, and $u \in [0, 1]$. The uniform distribution can be thought as being responsible for generating unique identifiers for hidden states. Given a hidden state $\theta = (u, y)$ at time t the observation at that time step is a deterministic function of θ: $L_{(u,y)} = \delta_y$. Further details of the likelihood model are provided in Chapter 8.

Any probability distribution that we can easily sample from and for which we can compute the importance weights using Equation 6.2 is a candidate for the proposal distribution. In this section we define a proposal $q(X_{g+1}^{m} | Y(t_{g+1}), X_{m,g})$ which, in contrast to the previous proposal, takes the observation $Y(t_{g+1})$ into account. Based on the defined likelihood model, if we know $Y(t)$, the observation at time step t, then we know the set of possible hidden states for that time step. Thus, this new proposal is more suitable for our experiments where we use the described likelihood model. We call this the “improved” proposal distribution.

We denote the set of possible hidden states for observation $Y(t)$ at time step t by $\Omega_{Y(t)}$ where $\Omega_{Y(t)} := \{(u,y) \in \Omega : y = Y(t), L_{(u,y)} = 1\}$. We are interested in a proposal distribution which, given $X_{m,g} = (\theta_{m,i}, J_{m,i})_{i=1}^{n}$ and $Y(t_{g+1})$, proposes a sample of hidden events $(\theta_{m,i}, J_{m,i})_{i=n+1}^{N}$ with the property that $\theta_{I(t_{g+1})} \in \Omega_{Y(t_{g+1})}$. Informally, we are only interested in samples which end up in a hidden state that can emit $Y(t_{g+1})$.

To obtain the desired proposal, we first define a probability distribution which we call the “modified” predictive distribution. As the name suggests, we define this distribution by modifying the predictive distribution of the model in Section 5.4. This probability distribution is the key element in our new proposal. Next, we introduce the “improved” proposal distribution. Finally, we explain how the importance weights can be computed.

“Modified” predictive distribution

We change the predictive distribution for the hidden states of the model in Section 5.4 to generate only samples from $\theta \in \Omega_{Y(t_{g+1})}$. We denote the modified predictive distribution by variables with superscript M. In order to obtain the modified predictive distribution, we change the base measure H_0, defined above (i.e., $H_0 = \text{Unif} \times \text{Mult}$); and redefine the sufficient statistics F_θ and G.

We define the modified base measure as $H_0^M = \text{Unif} \times Y(t_{g+1})$; hence, every hidden state sampled from H_0^M is a pair (u,y) where y is always equal to $Y(t_{g+1})$. In other words all the hidden states sampled from H_0^M always emit the observation $Y(t_{g+1})$ at time t_{g+1}.

Next we define the modified sufficient statistics as:

$$F_\theta^M = \sum_{i=1}^{n} \mathbf{1}[\theta_{i-1} = \theta \;\; \& \;\; \theta_i \in \Omega_{Y(t_{g+1})}] \; \delta_{\theta_i} \tag{6.6}$$

$$G^M = \sum_{i=1}^{n} \mathbf{1}[\theta_i \in \Omega_{Y(t_{g+1})}] A_i \delta_{\theta_i}. \tag{6.7}$$

For F_θ^M, the empirical transition measures for state θ, we make the mass at all atoms where $\theta_i \notin \Omega_{Y(t_{g+1})}$ equal to zero and keep the mass for all the other atoms the same as their mass in F_θ. Similarly for the G^M, we only keep the mass at atoms where $\theta_i \in \Omega_{Y(t_{g+1})}$ and assign zero mass to all the other atoms.

Using the same normalization constant as in the original predictive distribution (i.e., $\|\mu_0\|$), we *define* the modified predictive distribution $\bar{\mu}_\theta^{\prime(M)}$ based on:

$$\mu^{\prime\prime(M)} = G^M + H_0^M \tag{6.8}$$

$$\mu_\theta^{\prime(M)} = F_\theta^M + \|\mu_0\|\bar{\mu}^{\prime\prime(M)} \tag{6.9}$$

This modified predictive distribution is used in our proposal. Note that any state sampled from this probability distribution will only emit $Y(t_{g+1})$, the observation at time t_{g+1} .

Probability density value for a sampled state

Before proceeding further, it is useful to briefly review computing the probability density value of a sampled state. This value is needed in obtaining the probability of a sampled path from the proposal which in turn is required for computing the importance weights.

For a state θ^* sampled from the predictive distribution $\bar{\mu}_{\theta_N}^{\prime(\mathrm{H})}$, we aim to compute $p(\theta_{N+1} = \theta^* \big| X, \{A_n\}_{n=1}^N, \|\mu_0\|)$. We denote this probability density value by $CRF(\theta^*; \bar{\mu}_{\theta_N}^{\prime(\mathrm{H})})$. Using Equation 5.8 (see Figure 5.1), we can compute this value from the following formula:

$$CRF(\theta^*; \bar{\mu}_{\theta_N}^{\prime(\mathrm{H})}) = \begin{cases} \frac{F_{\theta_N}(\{\theta^*\})}{\|F_{\theta_N}+\mu_0\|} & ;\ A_N = 0 \\ \frac{G(\{\theta^*\})}{\|G+H_0\|}\frac{\|\mu_0\|}{\|F_{\theta_N}+\mu_0\|} & ;\ A_N = 1\ \&\ \theta^* \in \{\theta_{1:N}\} \\ \frac{H_0(\theta^*)}{\|G+H_0\|}\frac{\|\mu_0\|}{\|F_{\theta_N}+\mu_0\|} & ;\ o.w. \end{cases} \tag{6.10}$$

The probability density value for a sampled state θ^* from the "modified" predictive distribution $\bar{\mu}'^{(M)}_{\theta}$ is defined similarly:

$$MCRF(\theta^*;\bar{\mu}'^{(\mathrm{M})}_{\theta_N}) = \begin{cases} \frac{F^M_{\theta_N}(\{\theta^*\})}{\|F^M_{\theta_N}+\mu_0\|} & ;\ A_N = 0 \\ \frac{G^M(\{\theta^*\})}{\|G^M+H^M_0\|} \frac{\|\mu_0\|}{\|F^M_{\theta_N}+\mu_0\|} & ;\ A_N = 1 \ \&\ \theta^* \in \{\theta_{1:N}\} \\ \frac{H^M_0(\theta^*)}{\|G^M+H^M_0\|} \frac{\|\mu_0\|}{\|F^M_{\theta_N}+\mu_0\|} & ;\ o.w. \end{cases} \qquad (6.11)$$

Sampling procedure for the “improved” proposal distribution

Now that we have defined the key element of our proposal distribution, we introduce the “improved” proposal distribution. The pseudocode for sampling from the proposal distribution, in case of multiple sequences, is provided in Figure 6.6. We define the proposal by defining its sampling procedure for the case of a single sequence.

We assume that we are at observation time t_g and the current hidden state for the m-th particle is given: $\theta_{m,I(t_g)}$. For this proposal we should emphasize that although we sample until we cover the next observation time, we do not assume knowledge of the time to the next transition after t_{g+1} nor of the state to which that transition is made. In other words, we only know there is a transition after t_{g+1} but the state and the time to the transition are unknown (see Figure 6.7 for an illustration). For all generations of the particles we always start sampling from an observation time. This point will become clear when we describe how to sample from the proposal and compute the probability of a sampled path.

To simplify the notation we omit the subscript m in describing the procedure and denote $I(t_g)$ by *num*. We use l to index the events occurring after time t_g; for instance, the index of the first jump after t_g is $l = num+1$. For initializing the procedure we let $l = num+1$. The sampling procedure is described below. Figure 6.8(a) illustrates the initial state of the $g+1$-th generation of a single particle.

Given the current state θ_{l-1}, we sample J_l, the time to the next jump, from the translated Pareto distribution TP($\|\mu'^{(\mathrm{H})}_{\theta_{l-1}}\|, \beta'_{\theta_{l-1}}$). Depending on the sampled value

Algorithm 3 Modified predictive distribution as the proposal for inference in the HGEP model

```
Given the current hidden state at time point $t_g$ (i.e., $\theta_{m,I(t_g)}$):
(From now on, we omit writing $\forall m \in \{1,\ldots,M\}$ to avoid excessive
notation and also drop particle index $m$ from the notation)
$K$: number of sequences
$k$: sequence number $k \in \{1,\ldots,K\}$
$num$: $I(t_g)$, the number of events up to time step $t_g$
$l = num + 1$
$J_l = 0$

**While** $\sum_{i=num+1}^{l} J_i < t_{g+1} - t_g$
    - Sample the time to next jump $J_l$ from TP($\|\mu'^{(\mathrm{H})}_{\theta_{l-1}}\|, \beta'_{\theta_{l-1}}$)
     - **If** $\sum_{i=num+1}^{l} J_i < t_{g+1} - t_g$ then
        - sample the next state $\theta_l$ from $\bar{\mu}'^{(\mathrm{H})}_{\theta_{l-1}}$
        - $l = l + 1$
        - update the sufficient statistics $F^k_{\theta,l}$, $T^k_{\theta,l}$, $S^k_{\theta,l}$, and $G^k_l$
     - **else if** $\theta_{l-1} \in \Omega_{Y(t_{g+1})}$
        - $J_l = t_{g+1} - \sum_{i=1}^{l-1} J_i$ and $\theta_l = \theta_{l-1}$
        - update the sufficient statistic $T^k_{\theta,l}$
     - **else**
        - sample $J_l$ from $Unif[0, t_{g+1} - \sum_{i=1}^{l-1} J_i]$
        - sample $\theta_l$ from $\bar{\mu}'^{(M)}_{\theta_{l-1}}$, the modified predictive distribution
        - update the sufficient statistics $F^k_{\theta,l}$, $T^k_{\theta,l}$, $S^k_{\theta,l}$, and $G^k_l$
        - $l = l + 1$
        - $J_l = t_{g+1} - \sum_{i=1}^{l-1} J_i$ and $\theta_l = \theta_{l-1}$
        - update the sufficient statistic $T^k_{\theta,l}$
     **End If**
**End While**
```

Figure 6.6: Pseudocode for sampling from the "improved" proposal distribution

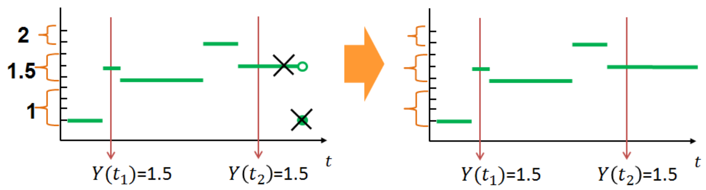
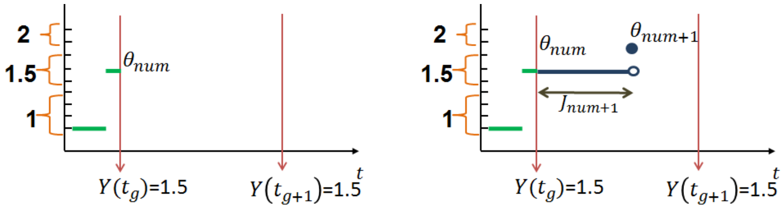


Figure 6.7: An illustration of sampling the last jump for a generation of particles in the SMC algorithm. For simplicity we show only one particle.



(a) (b)

Figure 6.8: Sampling procedure for the "modified" predictive distribution (case 1).

there are three different cases:

1. the sampled path does not pass the next observation time, $\sum_{i=num+1}^{l} J_i < t_{g+1} - t_g$. In this case, we sample the next state from $\bar{\mu}'^{(\mathrm{H})}_{\theta_{l-1}}$ defined in Equation 5.8; set $l := l+1$ and update the sufficient statistics $F_{\theta,l}$, $T_{\theta,l}$, $S_{\theta,l}$, and G_l.

   This case is depicted in Figure 6.8(b); in this example $J_{num+1} < t_{g+1} - t_g$ and $\theta_{num+1} \sim \bar{\mu}'^{(\mathrm{H})}_{\theta_{num}}$.

2. the sampled path passes the next observation time, and the observation emitted from the current state is consistent with $Y(t_{g+1})$ (i.e., the current state emits $Y(t_{g+1})$: $\theta_{l-1} \in \Omega_{Y(t_{g+1})}$). In this case, we discard the current J_l and instead set it equal to the remaining time to the observation time t_{g+1} (i.e., $J_l = t_{g+1} - \sum_{i=1}^{l-1} J_i$). Moreover, we set the next state equal to the current state: $\theta_l = \theta_{l-1}$ and update the sufficient statistic $T_{\theta,l}$. Note that we do not need to update the other sufficient statistics as the process remains in its current state without any transitions.

   An example of this case is given in Figure 6.9. In this example the observation emitted from θ_{num} (i.e., $EDSS = 1.5$) is consistent with $Y(t_{g+1})$; we have $J_{num+1} = t_{g+1} - t_g$ and $\theta_{num+1} = \theta_{num}$.

3. the sampled path passes the next observation time, and the current state does not emit $Y(t_{g+1})$ (i.e., $\theta_{l-1} \notin \Omega_{Y(t_{g+1})}$). In this case, we take the following two steps:

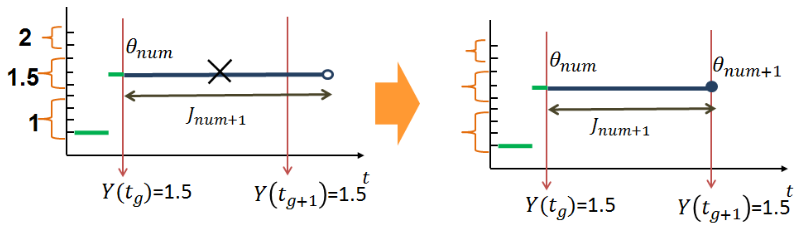


Figure 6.9: Sampling procedure for the “modified” predictive distribution (case 2).

(a) first we discard the current J_l and instead sample J_l from $Unif[0, t_{g+1} - \sum_{i=1}^{l-1} J_i]$; that is, we sample the time of the next jump uniformly between the current time, $\sum_{i=1}^{l-1} J_i$, and the next observation time t_{g+1}. Then, we sample the next state θ_l from $\bar{\mu}'^{(M)}_{\theta_{l-1}}$, the modified predictive distribution, and update the sufficient statistics $F_{\theta,l}$, $T_{\theta,l}$, $S_{\theta,l}$, and G_l.

(b) we set $l = l+1$, next we set $J_l = t_{g+1} - \sum_{i=1}^{l-1} J_i$ and $\theta_l = \theta_{l-1}$. That is, we do not sample the last jump; instead, we deterministically set the time to the last jump equal to $t_{g+1} - \sum_{i=1}^{l-1} J_i$ which is the remaining time till t_{g+1}. Finally, we only update $T_{\theta,l}$ as we do not assume knowledge of the state after t_{g+1} and for the next generation of the particles we are going to sample from time t_{g+1} forward.

For an example see Figure 6.10. In the example the sample for J_{num+1} is discarded and a new J_{num+1} is sampled from $Unif[0, t_{g+1} - t_g]$; the new jump time is denoted by t^*. For the next state we have $\theta_{num+1} \sim \bar{\mu}'^{(M)}_{\theta_{num}}$; hence, we have $\theta_{num+1} \in \Omega_{Y(t_{g+1})}$. Finally, the time to the last jump J_{num+2} is set equal to $t_{g+1} - (J_{num+1} + t_g)$.

The sampling procedure for this generation is stopped as soon as we cover the observation time t_{g+1}; that is, $\sum_{i=num+1}^{l} J_i < t_{g+1} - t_g$.

A sampled path from this proposal always ends up in a state $\theta_{I(t_{g+1})} \in \Omega_{Y(t_{g+1})}$. Whenever the path passes t_{g+1} we check whether the path is in the right group of states $\Omega_{Y(t_{g+1})}$; if not then we discard the sampled jump and sample the time of the jump from a uniform distribution and the state from the modified predictive distribution $\bar{\mu}'^{(M)}_{\theta}$. Therefore, the sampled path has the desired property.

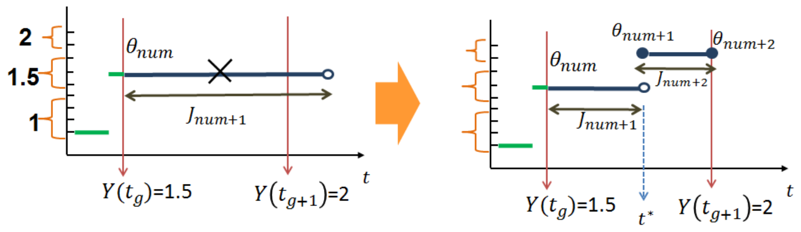


Figure 6.10: Sampling procedure for the "modified" predictive distribution (case 3).

Computing the importance weights

In order to compute the weights of the particles, we need the proposal and predictive probabilities for a sampled path. We denote the cumulative distribution of the translated Pareto distribution by $F_{TP}(t;\alpha,\beta)$, the updated MGP parameters after the ith transition by $\beta_\theta^{(i)}$ and $\mu_\theta^{(i)}$; and the predictive distribution of the ith event given events 1 to $i-1$ in the HGEP model by $p(\theta_i, J_i|\theta_{1:i-1}, J_{1:i-1})$. Setting $n = I(t_g)$, $N = I(t_{g+1})$, and $\Delta = t_{g+1} - \sum_{k=1}^{N-1} J_k$, the predictive probability for a sampled path $(\theta_i, J_i)_{i=n+1}^N$ (after dropping the particle index m and the superscript H) is given by:

$$
\begin{aligned}
&p((\theta_i,J_i)_{i=n+1}^N|X_g = (\theta_i,J_i)_{i=1}^n, \|\mu_0\|) \\
&\qquad = \prod_{i=n+1}^{N} p(\theta_i,J_i|\theta_{1:i-1},J_{1:i-1}) \cdot (1 - F_{TP}(\Delta; \|\mu_{\theta_N}^{(N)}\|, \beta_{\theta_N}^{(N)})) \\
&\qquad = \prod_{i=n+1}^{N} p(\theta_i|\theta_{1:i-1},J_{1:i-1}) \cdot p(J_i|\theta_{1:i-1},J_{1:i-1}) \cdot (1 - F_{TP}(\Delta; \|\mu_{\theta_N}^{(N)}\|, \beta_{\theta_N}^{(N)})) \\
&\qquad = \prod_{i=n+1}^{N} CRF(\theta_i; \bar{\mu}_{\theta_{i-1}}^{(i-1)}) \cdot TP(J_i; \|\mu_{\theta_{i-1}}^{(i-1)}\|, \beta_{\theta_{i-1}}^{(i-1)}) \cdot (1 - F_{TP}(\Delta; \|\mu_{\theta_N}^{(N)}\|, \beta_{\theta_N}^{(N)})).
\end{aligned}
$$

The third equality is due to the fact that given the current state the next state and jump time are independent. Figure 6.11 provides an example which clarifies the above computation. In the figure to keep the notation uncluttered we drop the parameters of the distributions. The figure is an example of a sampled path beginning from the n-th event; each state is sampled from the predictive distribution for

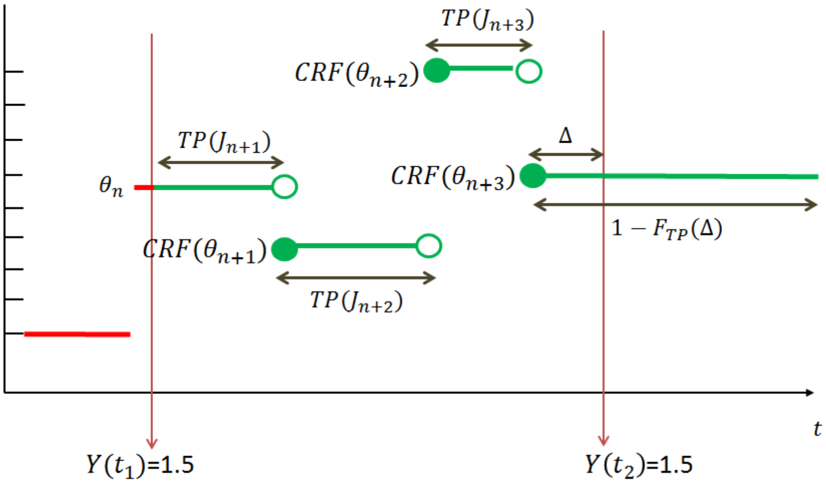


Figure 6.11: Computing the predictive probability for a sampled path

the states (denoted by CRF in the figure) given all the previous events. Moreover, each waiting time is sampled from the translated Pareto distribution. Finally, for the transition after t_{g+1} we know it happens at a time greater than Δ; hence, we use $1-F_{TP}(\Delta)$.

Based on the algorithm presented in Figure 6.6, assuming the proposed hidden events are $(\theta_{m,i},J_{m,i})_{i=n+1}^{N}$, the probability of a sampled path $q((\theta_i,J_i)_{i=n+1}^{N}|Y(t_{g+1}),X_g = (\theta_i,J_i)_{i=1}^{n},\|\mu_0\|)$ for the proposal can be obtained for two different scenarios:

1. If the $N-1$th (hidden) state, the state at the penultimate jump, belongs to $\Omega_{Y(t_{g+1})}$ (i.e., $\theta_{N-1}\in\Omega_{Y(t_{g+1})}$), then we are certain that the events at both jumps $N-1$ and N are sampled from the predictive distribution. In other words, the path is sampled from the predictive distribution at all the jumps.

2. If the $N-1$th state does not belong to $\Omega_{Y(t_{g+1})}$ then there could be two indistinguishable cases:

   (a) The Nth event is sampled from the predictive distribution and the Nth sampled state belongs to $\Omega_{Y(t_{g+1})}$; thus, similar to the previous case all the hidden events are sampled from the predictive distribution.

(b) The Nth jump is sampled from the uniform distribution and the Nth state is sampled from the modified CRF.

Denoting the modified predictive distribution by p^*, for the proposal probability we have:

$$\begin{aligned}
&q((\theta_i,J_i)_{i=n+1}^{N}|Y(t_{g+1}),X_g=(\theta_i,J_i)_{i=1}^{n},\|\mu_0\|)\\
&\quad=\mathbf{1}[\theta_{N-1}\in\Omega_{Y(t_{g+1})}]\prod_{i=n+1}^{N}p(\theta_i,J_i|\theta_{1:i-1},J_{1:i-1})\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{(N)}\|,\beta_{\theta_N}^{(N)}))\\
&\quad+(1-\mathbf{1}[\theta_{N-1}\in\Omega_{Y(t_{g+1})}])((\prod_{i=n+1}^{N}p(\theta_i,J_i|\theta_{1:i-1},J_{1:i-1})\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{(N)}\|,\beta_{\theta_N}^{(N)})))\\
&\qquad\qquad+\prod_{i=n+1}^{N-1}p(\theta_i,J_i|\theta_{1:i-1},J_{1:i-1})\cdot p^*(\theta_N,J_N|\theta_{N-1},J_{N-1})\\
&\qquad\qquad\qquad\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{\prime(\mathrm{H})}\|,\beta(N)_{\theta_N})))\\
&\quad=\mathbf{1}[\theta_{N-1}\in\Omega_{Y(t_{g+1})}](\prod_{i=n+1}^{N}CRF(\theta_i;\bar{\mu}_{\theta_{i-1}}^{(i-1)})\cdot TP(J_i;\|\mu_{\theta_{i-1}}^{(i-1)}\|,\beta_{\theta_{i-1}}^{(i-1)})\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{(N)}\|,\beta_{\theta_N}^{(N)})))\\
&\quad+(1-\mathbf{1}[\theta_{N-1}\in\Omega_{Y(t_{g+1})}])(\prod_{i=n+1}^{N}CRF(\theta_i;\bar{\mu}_{\theta_{i-1}}^{(i-1)})\cdot TP(J_i;\|\mu_{\theta_{i-1}}^{(i-1)}\|,\beta_{\theta_{i-1}}^{(i-1)})\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{(N)}\|,\beta_{\theta_N}^{(N)}))\\
&\qquad\qquad+\prod_{i=n+1}^{N-1}CRF(\theta_i;\bar{\mu}_{\theta_{i-1}}^{(i-1)})\cdot TP(J_i;\|\mu_{\theta_{i-1}}^{(i-1)}\|,\beta_{\theta_{i-1}}^{(i-1)})\\
&\qquad\qquad\qquad\cdot MCRF(\theta_N;\bar{\mu}_{\theta_{N-1}}^{(N-1,M)})\cdot\frac{1}{t_{g+1}-\sum_{i=1}^{N-1}J_i}\\
&\qquad\qquad\qquad\cdot(1-F_{TP}(\Delta;\|\mu_{\theta_N}^{(N)}\|,\beta_{\theta_N}^{(N)}))).
\end{aligned}$$

The weights of the particles now have the form:

$$w_{m,g+1}=\frac{p(X_{g+1}^{m}|X_{m,g})}{q(X_{g+1}^{m}|Y(t_{g+1}),X_{m,g})} \tag{6.12}$$

due to the fact that now we always have $L(Y(t_{g+1})|\theta_{I(t_{g+1})},S_{\theta}^{(\backslash k)}+S_{\theta,m,n}^{(k)})=1$.

Chapter 7

LIKELIHOOD MODEL

The likelihood model that we use in our experiments was introduced in Section 6.4.2. In this section we review the likelihood model and explain the changes that we make in the HGEP model in order to apply it in our experiments.

As mentioned, we assume that the observation y is a deterministic function of the hidden state and H_0 is a product of multinomial and uniform distributions, i.e. $H_0 = \text{Unif} \times \text{Mult}$. Thus, every hidden state sampled from H_0 is in the form of a pair $\theta = (u, y)$: u is a unique identifier for every hidden state $u \in [0, 1]$ which is sampled from the uniform distribution while the observations y take a value in the set of possible observations Σ. The set of possible observations consists of the finite number of symbols that we can observe.

Given $\theta = (u, y)$ at a time step t, we have a Dirac delta for L_θ: $L_{(u,y)} = \delta_y$. Moreover, we add a Dirichlet distribution over the parameters of the multinomial distribution; we denote these parameters $p_1, \cdots, p_{|\Sigma|}$ by $p_{1:|\Sigma|}$. This likelihood model can therefore be thought as a Dirichlet-Multinomial-Dirac model.

For our experiments we add another level of hierarchy to the HGEP. Informally, this level is responsible for counting the number of times a transition is made from hidden states emitting the same observation $y \in \Sigma$ to another hidden state. This allows the model to share statistical strength between states emitting the same observation. The model can be represented similarly to Equation 5.1:

$$\begin{aligned}
p_{1:|\Sigma|} &\sim Dir(\alpha_3) \\
H_0 &= \text{Mult}(p_{1:|\Sigma|}) \times \text{Unif} \\
\mu_0 &\sim \text{MGP}(H_0, \|H_0\|) \\
\mu_y|\mu_0 &\overset{\text{iid}}{\sim} \text{MGP}(\mu_0, \gamma_0) \\
\mu_\theta|\mu_y &\overset{\text{iid}}{\sim} \text{MGP}(\mu_y, \beta_0) \\
\theta_{N+1}\big|X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \bar{\mu}_{\theta_N} \\
J_{N+1}\big|X, \{\mu_\theta\}_{\theta\in\Omega} &\sim \text{Exp}(\|\mu_{\theta_N}\|).
\end{aligned} \tag{7.1}$$

The predictive distribution for the above model can be obtained recursively. However, we need to introduce a new additional auxiliary variable K_y which we will explain momentarily. Similar to the HGEP predictive distribution (Equation 5.8) of Section 5.3 we have the following equations for the predictive distribution of the hidden states:

$$\bar{\mu}''' = \frac{G}{\|G+H_0\|} + \frac{\|H_0\|}{\|G+H_0\|}\bar{H}_0$$

$$\bar{\mu}'' = \frac{K_y}{\|K_y+\mu_0\|} + \frac{\|\mu_0\|}{\|K_y+\mu_0\|}\bar{\mu}''' \tag{7.2}$$

$$\theta_{N+1}|\theta_1,\cdots,\theta_N,\|\mu_y\|,\bar{\mu}'' \sim \frac{F_{\theta_N}}{\|F_{\theta_N}+\mu_y\|} + \frac{\|\mu_y\|}{\|F_{\theta_N}+\mu_y\|}\bar{\mu}''$$

Figure 7.1 illustrates an example of sampling from this predictive distribution. Here $\alpha_2 = \|H_0\|$ and $\alpha_1 = \|\mu_0\|$. Note that in this model the hyperparameters are α_1, α_2, α_3, β_0, and γ_0.

The bottom level of the hierarchy is the same as in Figure 5.1 of Section 5.3. From the bottom level we can move to a higher level DP with probability proportional to $\|\mu_y\|$ which is the normalization of the middle level random measure.

For the middle level we define auxiliary variables B_n similar to A_n. That is, the

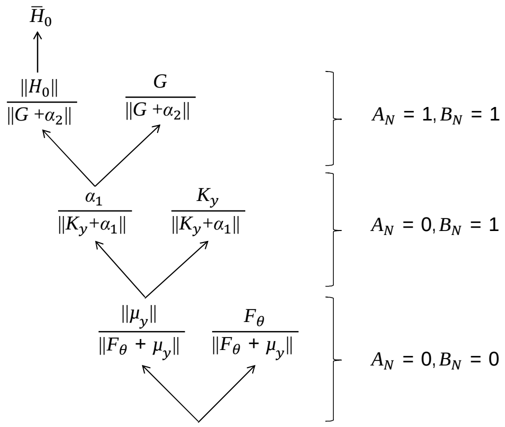

$\overline{H}_0$
$\frac{\|H_0\|}{\|G+\alpha_2\|}$
$\frac{G}{\|G+\alpha_2\|}$
$A_N = 1, B_N = 1$
$\frac{\alpha_1}{\|K_y+\alpha_1\|}$
$\frac{K_y}{\|K_y+\alpha_1\|}$
$A_N = 0, B_N = 1$
$\frac{\|\mu_y\|}{\|F_\theta + \mu_y\|}$
$\frac{F_\theta}{\|F_\theta + \mu_y\|}$
$A_N = 0, B_N = 0$

Figure 7.1: Sampling from the predictive distribution of the HGEP model used in the experiments

event $B_n = 1$ means we have moved from the bottom level to the middle level of the hierarchy at transition n. Furthermore, at this level of hierarchy we augment the sufficient statistics with the empirical counts for the number of transitions already made from a state that emits the observation y; that is, $K_y = \sum_{n=2}^{N} \mathbf{1}[L_{(\theta_{n-1},y)} = 1]\ B_n \delta_{\theta_n}$. As at the bottom level, we have two possibilities: to choose the state at this level with probability proportional to K_y or to default to another DP with probability proportional to α_1. As in Section 5.3, we use the auxiliary variable A_n to determine when we default to the top level DP (note that $A_n = 1$ implies $B_n = 1$).

Finally, the top level is the same as the top of the hierarchy in the model of Section 5.3. At this level, we can have a transition to an entirely new state with probability proportional to α_2 (i.e. $\|H_0\|$). We generate a unique identifier for this new state by sampling from $Unif[0,1]$. Furthermore, to decide on the observation for this new state, we use a multinomial distribution over the possible observations Σ.

Chapter 8

EXPERIMENTS

In this section we present the results of our experiments. First, we demonstrate the behavior of state trajectories and sojourn times sampled from the prior to give a qualitative idea of the range of time series of hidden states that can be captured by our model. Next, we evaluate our model quantitatively by applying it to held-out tasks for three different datasets: synthetic, multiple sclerosis (MS) patients, and RNA evolutionary. Each dataset consists of multiple time series with observations at different time points. Note that in each dataset we have finite number of possible observations which we call "symbols". A sample dataset with symbols A, C, T and G having the same structure as our three datasets is provided in Table 8.1. We will explain each of these datasets in detail later in this chapter.

Furthermore, we apply the model to the problem of comparing disease progression on the two arms of a clinical trial in MS; evaluating the effectiveness of a drug is a typical problem in MS clinical research. In order to compare the arms of a clinical trial we need a model for MS disease progression; however, as mentioned in Chapter 1, MS can be considered as a disease with complex latent structure. By using an HGEP as the model for the latent structure we compare the disease progression in the treatment and the placebo arms. Finally, we compare the performance of the two proposal distributions we introduced in Section 6.4.

Sequence 1					
Time points	0	1	1.5	2.3	4.1
Observations	A	A	T	C	C
Sequence 2					
Time points	1.1	1	3.1	4	
Observations	T	C	T	G	
			⋮		
Sequence M					
Time points	0	2	2.5	4.3	7
Observations	G	C	T	A	A

Table 8.1: A sample dataset with symbols A, T, C and G having the same structure as datasets used in the experiments

8.1 Qualitative Analysis of the Prior

The behavior of the prior can change as a function of the three parameters of the HGEP model: β_0, $||H_0||$ and γ_0. We sampled a sequence of length $T = 800$ and present the state-time plots. At least four types of behavior can be distinguished:

1. short sojourn times and high volatility of states (Figure 8.1(a))

2. long sojourn times with low volatility (Figure 8.1(b))

3. many new states with short sojourn times (Figure 8.1(c))

4. high tendency to create new states, long sojourn times (Figure 8.1(d))

The concentration parameter $||H_0||$ has the same interpretation as the concentration parameters of HDPs. In addition, in conjunction with the rate parameter β_0, it controls the waiting times. These behaviors show that the GEP process can describe various kinds of stochastic processes. For all of our experiments we chose (arbitrarily) values of 1 for all hyper-parameters.

8.2 Quantitative Analysis

In this section, we use the likelihood model introduced in Chapter 7 for discrete observations to evaluate our method on three held-out tasks. We considered three

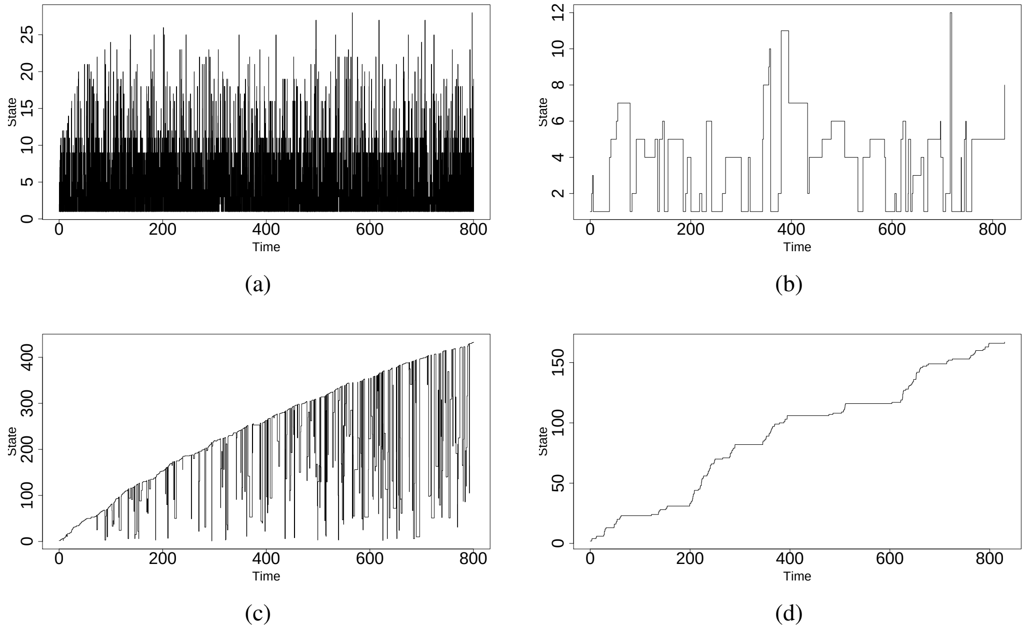


Figure 8.1: Qualitative behavior of the prior: (a) $\beta_0 = 10$, $\|H_0\| = 5$, $\gamma_0 = 100$, (b) $\beta_0 = 100$, $\|H_0\| = 5$, $\gamma_0 = 10$, (c) $\beta_0 = 10$, $\|H_0\| = 500$, $\gamma_0 = 10$, (d) $\beta_0 = 1$, $\|H_0\| = 5000$, $\gamma_0 = 1$

evaluation datasets obtained by holding out each observed datapoint with a 10% probability (see Table 8.2 for the properties of the datasets). We then reconstructed the observations at these held-out times, and measured the mean error. The time series were scanned for 1000 iterations of the PMCMC algorithm; that is, we carried the PMCMC algorithm for each dataset for 1000 iterations.

For HGEP, at each iteration of the PMCMC algorithm, we sample a latent path from an approximation to the posterior $p(X^{(k)}|\mathscr{Y}^{(k)})$ for every sequence k. Hence, for each held-out time of that sequence we have a sampled latent state. Based on our defined likelihood model, we know the observation the latent state emits. We reconstruct the held-out observation using this emitted observation for each iteration; we count the number of times each symbol is reconstructed up to that iteration and assign the symbol with the highest count to that held-out time. In other words, reconstruction is done by using the Bayes estimator approximated

Datasets					Results (mean error)		
Name	# sequences	# datapoints	# heldout	# characters	Baseline	EM	HGEP
Synthetic	1000	10000	878	4	0.703	**0.404**	0.446
MS	72	384	31	3	0.516	0.355	**0.277**
RNA	1000	6167	508	4	0.648	0.596	**0.426**

Table 8.2: Summary statistics and mean error results for the experiments. All experiments were repeated 5 times.

from 1000 posterior samples (one after each scan through all the sequences).

We repeated all experiments 5 times with different random seeds which control the randomness of sampling from the posterior. To compute the error for each iteration, we count the number of correct estimates and divide it by the total number of held-out observations; we only consider the error for the 1000th iteration as the error for that repetition of the experiment. The mean error is the average computed error for the 5 repetitions.

We compared against the standard maximum likelihood rate matrix estimator, estimated by the EM described in [22]. This method also estimates the transitions between two observations. As a result, we can estimate the whole trajectory for a sequence given only observations at finite time points. We reconstruct each held-out observation from this estimated trajectory, as we know the state of the sequence at every time point.

We also report in Table 8.2 the mean error for a simple estimate where we reconstruct each held-out observation using the most common symbol in the whole dataset; we call this estimate the baseline estimate. Figure 8.2 shows the mean error as a function of the number of iterations for all three datasets.

8.2.1 Synthetic data experiment

We used the Erdös-Rényi model [11], a model for generating random graphs, to generate a random rate matrix. The model is defined by two parameters: the number of nodes and the probability of having an edge between two nodes in the graph. We used the probability parameter $1/5$ and 10 nodes to generate a random graph. This random graph corresponds to a matrix of size 10×10 with elements equal to 1 where we have an edge between two nodes in the graph.

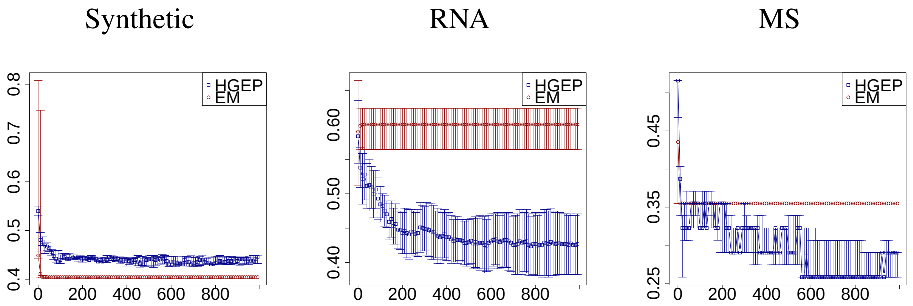


Figure 8.2: Mean reconstruction error on the held-out data as a function of the number of Gibbs scans for the experiments.

The non-diagonal zeros in this matrix correspond to entries with $\text{Unif}(0, 1/100)$ rate, and the non-diagonal ones in this matrix correspond to entries with $\text{Unif}(0, 1/2 + 1/100)$ rate. The diagonal entries were filled with minus the value of the sum of the non-diagonal ones, and each of the 10 states is set to deterministically emit one of the symbols A, B, C or D at random. To create the synthetic dataset, we generate 1000 sequences with time length 10 from the rate matrix using the Doob-Gillespie algorithm (described in Section 2.1).

Both HGEP and the EM-learned maximum likelihood outperformed the baseline. In contrast to the next two tasks, the EM approach slightly outperformed the HGEP model here. We believe this is because the synthetic data was not sufficiently rich to highlight the advantages of HGEPs.

8.2.2 RNA evolution modeling

MJPs can be used for describing the evolution of nucleotide sequences in a phylogenetic tree that shows the evolutionary relationships among different species. Each node of the tree corresponds to the nucleotide sequence of a species. While nucleotide sequences can be observed at the leaves of the tree (modern species), the information on the branch length (time since divergence from another species) and the topology of the tree cannot be observed.

We used the dataset from Cannone et al. [6] containing aligned 16S ribosomal RNA of species from the three domains of life. Aligned RNA sequences are sequences arranged based on their similarities. When two RNA sequences are aligned, the relation between two nucleotides at the same position can only be an identity, a mismatch or a gap. For instance, if we assume sequence AGGC evolves to AA-C and these sequences are aligned, then there are two similar sites (1 and 4), one mismatch at the second site and one deletion at the third site.

As a preprocessing step, a tree was constructed on a random subset of 30 species using PhyML [20], and the nucleotides at speciation events (i.e., branching events in the tree) were reconstructed using a K2P rate matrix [24] and the sum-product algorithm on trees.

We then considered the time series consisting of paths from one modern leaf to the root *for each site*. Figure 8.3(a) illustrates a sample path in the reconstructed phylogenetic tree. In this figure, node r is the root, nodes e, f, g, h are the leaves, and nodes a, b, c, d are the internal nodes reconstructed by the explained procedure. At each of these nodes there is a nucleotide sequence. However, we study each site of these sequences independently where the observed state is the nucleotide; thus, there are 4 possible observations (i.e. A, T, C, G) (see Figure 8.3(b)). For the RNA dataset, we considered 1000 sequences of nucleotides from the root to the modern leaves in the reconstructed tree. A sample sequence is denoted by a dashed rectangle in Figure 8.3(b). The number of datapoints and heldouts are indicated in Table 8.2.

For this dataset, the task is to reconstruct held-out nucleotides using only the data in the paths. Again, both HGEP and EM outperformed the baseline, and our model outperformed EM with a relative error reduction of 29%.

8.2.3 MS disease progression

The dataset we used for this experiment is obtained from a phase III clinical trial of a drug given to MS patients. The dataset tracks the progression of MS (i.e., EDSS scores) in 72 patients of the placebo arm of the trial over 3 years. This is the subsample of the original cohort of patients who enrolled early to the trial and completed 3 magnetic resonance imaging scans [39]. As in Mandel [31], the observed EDSS score of a patient at a given time is binned into three categories: 1 (EDSS ≤ 1.5); 2 (EDSS $= 2, 2.5$); and 3 for (EDSS ≥ 3). Here, we are assuming these

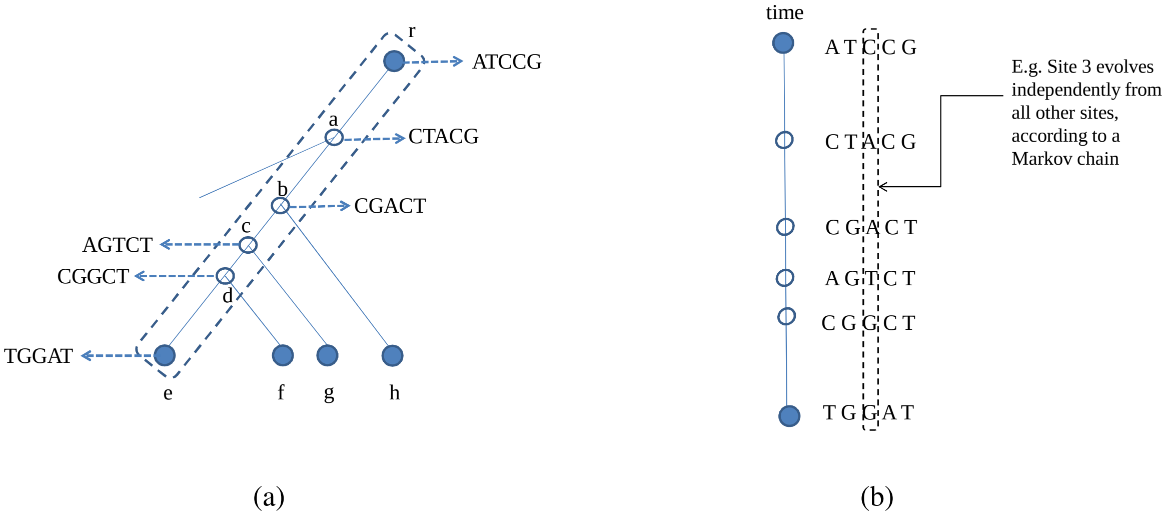


Figure 8.3: (a) A sample path in a reconstructed phylogenetic tree, (b) Evolution of the sequences for the reconstructed sample path

categorical observations can be described by an HMM with underlying, possibly, infinite hidden states that are not directly observed.

In the held-out task experiment, both HGEP and EM outperformed the baseline by a large margin. The HGEP model outperformed EM with a relative error reduction of 22%.

8.2.4 Application in estimating disease progression in MS

In this section we apply our model to the problem of comparing the disease progression in treatment and placebo arms of a clinical trial. We use the data of the placebo arm data from the previous section along with the data from the treatment arm of the same trial. Several tools are currently used in MS to describe or predict the course of the disease. Examples include expected time to a certain EDSS score [32], time to confirmed progression (i.e. reaching a certain EDSS and staying there

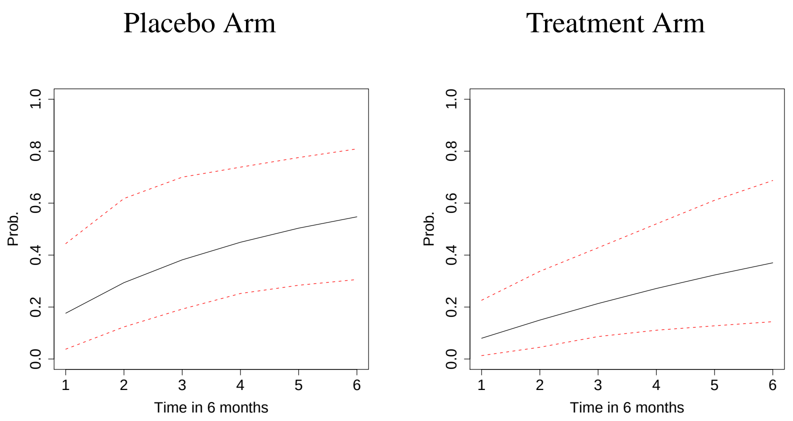
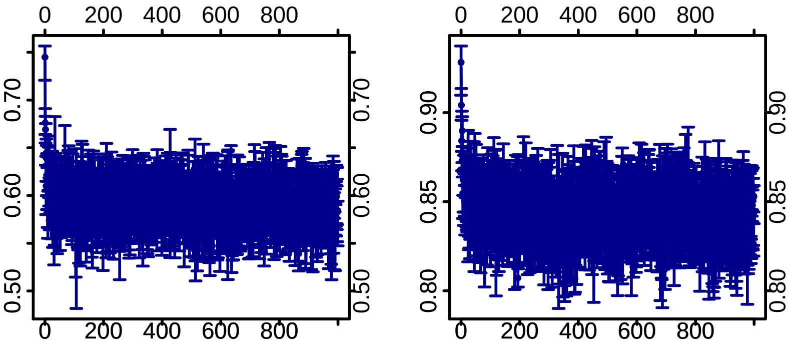


Figure 8.4: Mean predictive probability of reaching category 3 starting in category 1 (95% credible interval is provided by the dashed lines)

for 6 months or longer), and plots for the probability of confirmed progression against time [31].

We focus on the plots of probability of progression against time; these plots show the (mean predictive) probability of reaching a particular EDSS state against time, for patients who start at the same baseline EDSS state. To construct these plots, after each of the 1000 PMCMC iterations, we sample (without rejection) 100 times from the predictive distribution of the time to the particular state conditioned on the starting state. Then, using the 100 samples, for each PMCMC iteration we estimate the probability of having hit a specific EDSS state at a time point or earlier. We use 11 time points, ranging from 1 to 6 months in 15 days steps. Finally, we compute the mean of the probabilities at each of the 11 time points to summarize the results.

As mentioned observed EDSS scores are binned into three categories. Figure 8.4 shows the plot of the predicted probability of reaching category 3 starting from category 1 for the both arms of a clinical trial, along with their 95% credible intervals. The plots depict that, for instance, after 3 years the mean predicted probability of reaching category 3 for the treatment and the placebo arms are 0.37 (95% C.I.: 0.12-0.61) and 0.55 ((95% C.I.: 0.22-0.73)), correspondingly.



(a) (b)

Figure 8.5: Comparison of inference algorithms for the RNA dataset: mean acceptance rate per iteration for a proposal using 100 particles as a function of MCMC iteration using (a) predictive distribution (b) modified predictive distribution in the SMC proposal.

8.3 Comparison of Inference Algorithms

In this section to analyze the effect of using the two different proposal distributions described in Section 6.4 in the PMCMC algorithm for our experiments. The most important effect is on the acceptance rate of the samples in the PMCMC algorithm. Figure 8.5 shows that the acceptance rate in the "improved" proposal distribution is greater than that for the original proposal distribution. We have only provided the results for the RNA dataset; similar results hold for other datasets.

Moreover, for the "improved" proposal we always accept all the *particles in the SMC algorithm* so we have zero failure rate for the particles. Recall that the weight of each particle is computed based on the likelihood of the observation. For our defined likelihood the likelihood can only be 0 or 1. Without the "improved" proposal we can have many zero weighted particles in case of long sequences (i.e., high failure rate for the particles) due to the fact that for long sequences, we have higher chance of having at least one zero likelihood value.

Chapter 9

CONCLUSION

In this thesis, we introduced a model for MJPs with (potentially) infinite state space; we showed how this can be used as a prior in a Bayesian approach to the analysis of multiple sequences of data observed at discrete time points. We reviewed some basic notions and closely related models, built our model (GEP) based on the gamma process, and extended it to a hierarchical model (HGEP). We showed the GEP has some attractive properties such as conjugacy and closed form predictive distributions. Finally, we applied the model to some real world datasets; the results showed the model outperformed the maximum likelihood rate matrix estimator despite using a simple likelihood model (i.e., a Dirac delta function).

However, in our model, as in many nonparametric Bayesian models, the data likelihood (i.e., $p(\mathscr{Y})$) is nontrivial and hard to compute. As a result, performing Bayesian model comparison and computing the Bayes factor, which requires computing the data likelihood, is a challenging problem in our model. An example of this problem is the experiment of Section 8.2.4 where we want to compare the two arms of the trial; there may be some situations in which the presence of a treatment effect is less obvious from the plots and we need a statistical test for testing the treatment effect. Moreover, our inference algorithm is computationally intensive; we may improve it in terms of computational efficiency by possibly using a slice sampler.

We are currently working on some of the applications and generalizations of our model. Application of the method in models with continuous state space for the observations, proposing more efficient inference algorithms, calculating the Bayes factor for estimating the treatment effect in clinical trials, and incorporating covariates in the hierarchical GEP are issues that we are plan to address in future research. We are also interested in adding a layer of decision making to our model. This can make the model suitable for sequential decision making problems where we have partial observability of the data and the complexity of the latent structure is unknown.

Bibliography

[1] P. S. Albert. A Markov model for sequences of ordinal data from a relapsing-remitting disease. *Biometrics*, 50(1):51–60, 1994. → pages 2

[2] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. → pages 45, 48, 49, 50, 56, 57

[3] M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 14:577–584, 2002. → pages 3, 24

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. → pages 18

[5] A. Bureau, S. Shiboski, and J. P. Hughes. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22(3):441–462, 2003. → pages 1, 2

[6] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. D'Souza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Müller, et al. The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3 (1):2, 2002. → pages 76

[7] Y. Chung and D. Dunson. The local Dirichlet process. *Annals of the Institute of Statistical Mathematics*, 63(1):59–80, 2011. → pages 28

[8] J. L. Doob. Markoff chains–denumerable case. *Transactions of the American Mathematical Society*, 58(3):455–473, 1945. → pages 7

[9] D. Dufresne. G distributions and the beta-gamma algebra. *Electronic Journal of Probability*, 15(71):2163–2199, 2010. → pages 33

[10] D. Dunson. Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7(4):551–568, 2006. → pages 27

[11] P. Erdős and A. Rényi. On the evolution of random graphs. *Evolution*, 5(1): 17–61, 1960. → pages 75

[12] T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973. → pages 13, 14

[13] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning*, pages 312–319. ACM, 2008. → pages 3, 25, 26

[14] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. → pages 48

[15] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, pages 169–193. Oxford University Press, 1992. → pages 48

[16] J. Griffin. The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference*, 141(11):3648–3664, 2011. → pages 28

[17] J. Griffin and M. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006. → pages 3, 27

[18] J. Griffin and M. Steel. Stick-breaking autoregressive processes. *Journal of Econometrics*, 162(2):383–396, 2011. → pages 3, 27

[19] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001. → pages 46

[20] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52 (5):696–704, 2003. → pages 77

[21] P. Heidelberger and P. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983. → pages 48

[22] A. Hobolth and J. Jensen. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. → pages 75

[23] M. Jordan. Dirichlet processes, chinese restaurant processes and all that. In *Tutorial presentation at the NIPS Conference*. URL http://www.cs.berkeley.edu/~jordan/nips-tutorial05.ps. → pages 15

[24] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980. → pages 77

[25] J. Kingman. *Poisson Processes*. Oxford University Press, 1993. → pages 8, 9

[26] J. Kivinen, E. Sudderth, and M. Jordan. Learning multiscale representations of natural scenes using Dirichlet processes. In *IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. → pages 25

[27] D. Kroese, T. Taimre, and Z. Botev. *Handbook of Monte Carlo Methods*. Wiley, 2011. → pages 6

[28] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, 2008. → pages 45

[29] S. MacEachern. Dependent nonparametric processes. In *Section on Bayesian Statistical Science, American Statistical Association*, pages 50–55, 1999. → pages 26, 27

[30] R. MacKay. Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30(4):573–589, 2002. → pages 2

[31] M. Mandel. Estimating disease progression using panel data. *Biostatistics*, 11(2):304–316, 2010. → pages 1, 2, 77, 79

[32] M. Mandel and R. A. Betensky. Estimating time-to-event from longitudinal ordinal data using random-effects Markov models: application to multiple sclerosis progression. *Biostatistics*, 9(4):750–764, 2008. → pages 2, 78

[33] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. → pages 45

[34] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modelling*. Springer-Verlag, 1995. → pages 32

[35] L. Nieto-Barajas, P. Müller, Y. Ji, Y. Lu, and G. Mills. Time series dependent Dirichlet process. *Preprint*, 2008. → pages 28

[36] J. Pitman. *Combinatorial Stochastic Processes*. Springer-Verlag, 2006. → pages 13

[37] A. Raftery and S. Lewis. Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7(4):493–497, 1992. → pages 48

[38] A. Rodriguez, D. Dunson, and A. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008. → pages 28

[39] R. A. Rudick, E. Fisher, J. C. Lee, J. Simon, and L. Jacobs. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting multiple sclerosis. *Neurology*, 53(8):1698–1704, 1999. → pages 77

[40] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. → pages 13, 14

[41] A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, 11(2-3): 413–428, 2004. → pages 1

[42] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. → pages 17, 19, 21, 22, 23, 24

[43] A. C. Titman and L. D. Sharples. Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66(3):742–752, 2010. → pages 2

[44] W. Wei, B. Wang, and D. Towsley. Continuous-time hidden Markov models for network performance evaluation. *Performance Evaluation*, 49:129–146, 2002. → pages 1

[45] E. Xing and K. Sohn. Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis*, 2(3):501–528, 2007. → pages 25