Title
Bayesian Nonparametric Latent Feature Models

Author
Miller, Kurt Tadayuki

Publication Date
2011

Peer reviewed|Thesis/dissertation

Bayesian Nonparametric Latent Feature Models

by

Kurt Tadayuki Miller

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Thomas L. Griffiths
Professor Daniel Klein

Fall 2011

Bayesian Nonparametric Latent Feature Models

by

Kurt Tadayuki Miller

Abstract

Bayesian Nonparametric Latent Feature Models

by

Kurt Tadayuki Miller

Doctor of Philosophy in Engineering–Electrical Engineering and Computer Sciences

and the Designated Emphasis in Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

Priors for Bayesian nonparametric latent feature models were originally developed a little over five years ago, sparking interest in a new type of Bayesian nonparametric model. Since then, there have been three main areas of research for people interested in these priors: extensions/generalizations of the priors, inference algorithms, and applications. This dissertation summarizes our work advancing the state of the art in all three of these areas. In the first area, we present a non-exchangeable framework for generalizing and extending the original priors, allowing more prior knowledge to be used in nonparametric priors. Within this framework, we introduce four concrete generalizations that are applicable when we have prior knowledge about object relationships that can be captured either via a tree or chain. We discuss how to develop and derive these priors as well as how to perform posterior inference in models using them. In the area of inference algorithms, we present the first variational approximation for one class of these priors, demonstrating in what regimes they might be preferred over more traditional MCMC approaches. Finally, we present an application of basic nonparametric latent features models to link prediction as well as applications of our non-exchangeable priors to tree-structured choice models and human genomic data.

Dedicated to Melissa

Acknowledgements

My research and education would not have been possible without the help and support of advisors, colleagues, friends, and family.

I am extremely grateful for the advice and support of professors Michael Jordan and Thomas Griffiths, who have helped guide me on both academic and personal decisions and whose influence can be seen throughout this dissertation. In addition, the fantastic students and postdocs of SAIL, my great friends from my years at the Ashby house, and my teammates on my various soccer teams have all contributed to both the education as well as the fun that I have in my time at Berkeley. Part of my time at Berkeley was funded by the Lawrence Scholars Program through Lawrence Livermore National Laboratory, and to them as well as my very generous lab mentor Tina Eliassi-Rad, I owe a great deal of thanks.

In addition to the support I have had at Berkeley, I have had the privilege of collaborating with Yee Whye Teh, Finale Doshi-Velez and Jurgen Van Gael on the work presented in Sections 3.3 and 3.4. I have also benefited greatly from conversations with friends I have met at various conferences and universities.

Before Berkeley, I was first introduced to academic research while working on my Master's degree at Stanford. I worked closely with postdoc Mark Paskin, who was very instrumental in laying the groundwork for my future academic pursuits, as well as professors Sebastian Thrun and Andrew Ng. My first exposure to industrial research was courtesy of Geoffrey Barrows, who introduced me to being on the cutting edge of technology during our time together at the Naval Research Laboratory and later at Centeye.

Finally, my fiancée Melissa, parents John and Eileen, and siblings Adriane and Brendan have provided me with the love and support I have needed throughout all of this and without them, none of this would have been possible.

Contents

Chapter 1

Introduction

In many statistical problems, we observe some set of data and wish to infer various quantities related to it. This can be as simple as estimating the mean of the data or can be more complicated like estimating the entire distribution of the data. Whatever it is we wish to infer, we generally need to make some kind of assumption about the form, structure, and/or distribution of the data. Often, the more assumptions we make, the simpler it is to perform inference, but if these assumptions are false, we could be drawing incorrect inferences from the data. A common assumption is that the data comes from some simple distribution with a few unknown parameters. The simplest kind of inference then reduces to estimating the exact values of these parameters. This is often a very useful first step in understanding the data, but as our understanding of the data grows, it is desirable to reduce the number of assumptions we make and allow for richer models. We therefore often look beyond these simple parametric models to nonparametric ones. This dissertation explores how to do this in the Bayesian setting for one particular class of models.

The field of *Bayesian nonparametric statistics* seeks to combine the best of the *Bayesian* and *nonparametric* worlds. From the Bayesian world, we would like a mathematically elegant framework for updating our beliefs about unknown quantities based on any data we observe. More concretely, we would like to be able to place a prior $p(\phi)$ on unknown quantities or parameters ϕ, observe data X thought to be related to ϕ via a likelihood function $p(X|\phi)$, and then update our belief about what ϕ is via Bayes's rule $p(\phi|X) \propto p(X|\phi)p(\phi)$. This has traditionally been done in the parametric setting in which ϕ is a finite dimensional real-valued vector. From the nonparametric world, we seek to develop priors and models that allow us to draw more complex inferences as we observe more and more data. In order to do this, we cannot assume any particular fixed parametric form when modeling the data. We must have models that can grow in complexity as we observe more data.

Bayesian nonparametric methods (also commonly referred to as nonparamet-

ric Bayesian methods) combine these two paradigms by letting ϕ be an infinite-dimensional parameter. Defining the prior $p(\phi)$ on this infinite-dimensional parameter space is equivalent to defining an infinite-dimensional stochastic process. With this generality, one could develop many exotic stochastic processes as priors. However, few of them lead to reasonable predictive models for which we know how to compute the posterior distribution $p(\phi|X)$. Therefore, the trick is to develop stochastic processes over infinite-dimensional spaces in such a way that for useful likelihoods $p(X|\phi)$, we can practically compute $p(\phi|X)$.

There are many Bayesian nonparametric priors based on random infinite-dimensional objects. In the machine learning community, work has mostly focused on three of these Bayesian nonparametric priors:

1. Gaussian process
2. Dirichlet process/Chinese restaurant process and related priors
3. Beta process/Indian buffet process and related priors

The Gaussian process is a prior on the infinite-dimensional space of continuous functions and therefore is directly applicable to nonparametric regression, though it has also been successfully applied to classification as well as other domains. Of the Bayesian nonparametric priors we list above, it has the longest history with some of its main ideas going back centuries to Gauss himself with later developments in the early twentieth century. It gained in popularity in the latter half of the twentieth century in the geostatistical community under the name of Kriging (Cressie, 1993; Stein, 1999) before taking off in the machine learning community in the 1990s. For an overview of these priors, see Rasmussen and Williams (2006).

The Dirichlet process and its extensions are priors on the infinite-dimensional space of discrete distributions. They are commonly used as priors on latent class models such as those used in clustering and mixed membership models. The Dirichlet process was first introduced by Ferguson (1973), but again did not gain widespread adoption until the late 1990s/early 2000s when computational techniques and resources allowed them to be more practically applicable. While there are now various tutorials, chapters and monographs on these priors, one of the best introductions is Chapter 2 of Sudderth (2006).

The Beta process and related priors are examples of priors for Bayesian nonparametric latent feature models and are the focus of this dissertation. These are priors on the infinite-dimensional space of discrete measures (not necessarily distributions) and are commonly used as priors on binary matrices or non-negative integer valued matrices. They have the shortest history in the machine learning community, having been originally introduced by Griffiths and Ghahramani (2006) and Thibaux and

Jordan (2007), with roots in the work of Hjort (1990) and Kim (1999) in the survival analysis community. This earlier work itself was based on Lévy processese developed by Paul Lévy in the 1930s. The next chapter will provide a formal introduction to these priors and reviews the required background.

Since Bayesian nonparametric latent feature models have the shortest history, there are still many areas that need to be further developed. These areas can be broken down into three categories:

- Extensions and generalizations of the priors: we must understand the assumptions of these priors and, when they do not adequately fit our desired uses, figure out how we can extend them or generalize them to make them more broadly applicable.

- Inference algorithms: we must be able to perform posterior inference in models using these priors. As was stated earlier, it is easy to define infinite-dimensional stochastic processes, but these are not practical unless we can compute posterior distributions.

- Applications: we must understand and explore the applications for which these priors are appropriate as well as for which they are suboptimal. Without good applications, these priors will find limited interest.

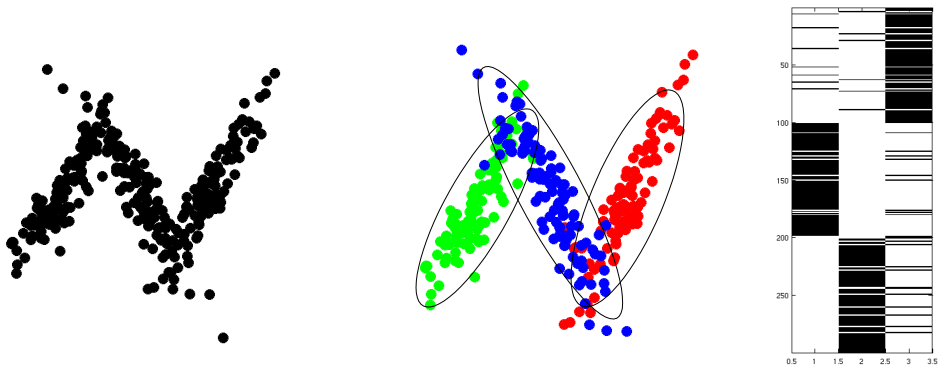
This dissertation presents our work in all three of these areas. We begin by reviewing relevant background in Chapter 2 and introducing the basic priors for Bayesian nonparametric latent feature models. Chapter 3 reviews sample-based inference algorithms and introduces our work on variational inference algorithms. Chapter 4 presents our nonexchangeable generalizations of the priors for Bayesian nonparametric latent feature models and Chapter 5 discusses inference algorithms for models using these new priors. Chapter 6 brings all of this work together by discussing our applications of these priors. We summarize our contributions in Chapter 7.

Chapter 2

Bayesian Nonparametric Latent Feature Models

In this chapter, we introduce Bayesian nonparametric latent feature models and provide the relevant background material. We start by motivating their use as well as establishing notation and background material in Section 2.1. We then review Lévy processes, one of the principal mathematical tools for developing these priors in Section 2.2. Given this background, the final two sections of this chapter review the two main classes of priors for latent feature models. In Section 2.3, we discuss priors for Bayesian nonparametric latent feature models with binary-valued latent features and in Section 2.4, we discuss priors for Bayesian nonparametric latent feature models with non-negative integer valued latent features. Knowledge of everything in this chapter is not necessary for users of these priors and models, but will be important for anyone interested in fully understanding and extending them.

In this chapter (and the rest of this dissertation), we will assume knowledge of several concepts. First, we assume the reader is familiar with probability theory. Measure theoretic probability theory at the level of Durrett (2004) or Kallenberg (1997) is sufficient, but not entirely necessary. Second, the reader should be comfortable with probabilistic graphical models and the ideas behind latent class methods such as Gaussian mixture models. To review these concepts, see Bishop (2007) or Koller and Friedman (2009). Finally, we assume the reader is familiar with the basics of Bayesian analysis as described in Gelman et al. (2003) and Markov Chain Monte Carlo as described in Robert and Casella (2004).



(a) (b) (c)

Figure 2.1: Gaussian mixture models. (a) Data generated from a Gaussian mixture model. (b) One potential set of class membership assignments and corresponding Gaussian distributions. (c) The relevant class membership matrix corresponding to (b).

2.1 Overview

Probabilistic graphical models provide a powerful formalism for working with data. Many unsupervised approaches use this framework to find latent structure in observed data that can help explain our observations.

Latent class models such as the Gaussian mixture model are a popular class of unsupervised models. We begin by giving a high level motivation for these approaches and then introduce their generalization to latent feature models.

2.1.1 Latent Class Models

In the Gaussian mixture model (GMM, also known as a Mixture of Gaussians, MoG), a special case of latent class models, we observe N data points $x_1, x_2, \ldots, x_N$ and we believe that these data points have been generated by the mixture of several different Gaussian distributions. Each data point is assumed to have been generated from a single one of these distributions. For example, if $x_i \in \mathbb{R}^2$, then our observations might look like Figure 2.1(a) where this data comes from a mixture of three Gaussians. Our goal is then to infer what the parameters are for each of the Gaussians and which data points have been generated from each Gaussian. Given the raw data in Figure 2.1(a), we might infer the latent class memberships indicated by the different colors in Figure 2.1(b) along with the corresponding Gaussians. As part of this, we wish

to infer a binary matrix Z where Z is an $N \times K$ matrix where N is the number of data points and K is the number of classes. In this matrix, there is a one (black in the figure) at $Z(i, j)$ if the i^{th} observation was generated from the j^{th} class and zero (white in the figure) otherwise. Figure 2.1(c) shows the class memberships inferred in Figure 2.1(b). For details on inference and inference algorithms, see Bishop (2007). Note that in general, we are rarely 100% sure about class memberships, so we will often infer distributions on entries in Z.

Latent class models beyond the GMM allow for more general distributions than just the normal distribution to generate each class, but all assume that there is some underlying binary matrix Z that must be inferred.

There are several issues that must be addressed with this latent class representation. First of all, we often do not know K, the number of latent classes that generate any particular data set. While there are both frequentist and Bayesian approaches to this tackling this problem, we focus on the Bayesian approaches. Within the Bayesian approaches, the Dirichlet process, one of the three Bayesian nonparametric priors we mentioned in Chapter 1, has become a very popular solution to this problem over the past ten years. Due to the success of this approach, much of the later developments in priors for Bayesian nonparametric latent feature models that we will soon describe can be related to developments of the Dirichlet Process. While knowledge of these developments is extremely useful and highly recommended for the understanding of Bayesian nonparametric latent feature models and priors, we will not review this prior work since it is not required background. Unfortunately, there are no concise reviews of the relevant work, but the interested reader can begin with recent review articles such as Teh and Jordan (2010) and Teh (2010), or the longer book by Hjort et al. (2010).

The second issue is that while latent class models are excellent models across a wide variety of data, they are not always the best choice of models. For example, when modeling various kinds of human data, if we used a latent class model, then it would be equivalent to saying that there are certain classes of people and that each person only belongs to a single group. When taken to its extreme and each person belongs to his/her own group, this would allow us to model people we have seen very well. However, this would not allow us to generalize whatever we learn to apply to people we have not seen yet. Therefore, we would like classes to correspond to multiple people so that our results can generalize. In order to explain people well though, the classes we would need to infer would be very specific and we would need a large amount of data to learn those classes well. In addition, if we learned the characteristics of each class independently, we would fail to capture the fact that different classes share different characteristics. We would ideally like to learn a more compact representation that captures these overlapping characteristics. This is precisely the point of latent feature models.

2.1.2 Latent Feature Models

Latent feature models address the last issue brought up in the previous section. That is, latent feature models allow us to learn a compact representation that can simultaneously explain our observations as well as any unobserved data. Just like latent class models, they are not applicable to every kind of data, but there are many data sets that are well modeled with latent features.

Latent feature models generalize the form of the latent matrix Z that we wished to infer in the previous section. In latent class models, Z is a binary matrix with each row corresponding to each data point and each column corresponding to a class. There can be only one non-zero entry in each row, but each column ideally has multiple non-zero entries. In latent feature models, each row still corresponds to a single data point, but now the columns correspond to different features and each data point may be possess different amounts of each of these features. In general, these can be real-valued features with many non-zero entries in every row. However, in order to have a practical model, each row of Z can only have a finite number of non-zero entries and it is hard to directly work with a real-valued process that has this kind of sparsity, so it is hard to have a real-valued nonparametric prior. In this dissertation, we therefore restrict our attention to priors for binary and non-negative integer valued features since this is a reasonble place to start and we will show how to attain the desired sparsity of Z. These kinds of priors can then be combined with real valued processes to generate nonparmetric real-valued priors. Therefore, we will work with binary or non-negative integer valued matrices in which every row is now allowed to have multiple non-zero entries. It's as simple as that! The rest of this dissertation flushes out this simple idea.

There are two main kinds of priors for Bayesian nonparametric latent feature models we will discuss. In the first type, Z is still a binary matrix as described above, so data points either have or do not have the feature. In the second type, Z is a non-negative integer valued matrix in which entries are the number of times each data point has that feature. We will discuss these two types in more detail in Sections 2.3 and 2.4.

What is the interpretation of the columns of Z in these latent feature models? Going back to the human data data example, in the binary-valued latent feature models, the columns might correspond to features that humans either do or do not possess that we wish to infer. For example, if we had no prior information about people, the unobserved binary features that we wish to infer might be "UC Berkeley student," "soccer player," and "lives in California." Each of these may have some effect on our observed data and humans may have any number of these features. As an example of a non-negative integer valued features, there might be a feature such as "number of cars owned." These are most often used as counts of various attributes.

Note that any latent class model can be represented by a latent feature model in which each row is restricted to have only a single non-zero entry, so latent class models can be seen as a special instance of latent feature models. In addition, the exact opposite is also true. Anything represented by a latent feature model can also be represented by a latent class model. Sticking to binary features, it is clear that with the three binary features listed previously, we could easily construct $2^3 =$ eight classes and use a latent class model to have an equally expressive model. Thus, we can see that latent class models, by using exponentially many more classes, can explain the same thing as much more compact latent feature models, so latent feature models can be seen as a special case of latent class models. However, as the number of features K grows, we would need 2^K classes to explain the same thing as a latent feature model with K features. Furthermore, we might think all "soccer players" should share some attribute which is easy to do if each feature has an associated parameter, but harder to do when using 2^K classes.

Given that latent class models and latent feature models are just as expressive as each other and that latent feature models furthermore have much more compact representations and allow for easier parameter sharing, it might seem that we should always use latent feature models. However, as is often the case, learning these richer, more compact and expressive models is more computationally intensive, so if data is well explained by latent classes, it is more efficient to use latent class models.

As a final note, there are several naming variations used for "latent feature models." Some people also refer to the models developed in this section as "multiple class models," "binary factor models," "factorial models," "distributed representations," and several other variations along these lines. In addition, people will also call other, related kinds of models "latent feature models." For this entire dissertation, when we discuss "latent feature models," we are referring to the models motivated in this section.

2.1.3 Notation

Now that we have motivated latent feature models, we must explain how they work and we must again tackle the problem that K, the number of features, is not known to us in advance. For now, assume that knowledge of K is not an issue. As mentioned in Chapter 1, one of the big advantages of Bayesian nonparametric approaches is that by using rich priors in our models, inference of K is part and parcel of the posterior inference process. This idea will be further developed in later sections.

We assume there are N entities that we will observe data about. Our variables will be the following:

- Observations:

  - X: Our observations associated with the N entities. Often, $X = (x_1, \ldots, x_N)$, that is, X is composed of separate observations x_i, one for each entity i. We will see examples in Sections 6.1 and 6.2 in which this is not true, but for now, assume $X = (x_1, \ldots, x_N)$.

- Unknowns:

  - Z: The $N \times K$ latent feature matrix as described earlier.
    - Let z_i correspond to the i^{th} row of Z and therefore the feature vector corresponding to the i^{th} entity.
    - We will occasionally need to refer to a particular column, and will use the index k to signify this, so z_k will correspond to the k^{th} column. Besides the index, context will help determine if we are referring to a row or column.
    - Let z_{ik} be the (i, k) entry of the matrix.
    - During inference, we will need to refer to the matrix except particular entries. Let Z_{-ik} be all of Z except the (i, k) entry.
    - If we are just referring to the k^{th} column z_k, let $z_{(-i)k}$ be all of z_k except z_{ik}.
  - θ: Additional parameters describing how Z influences X. This is going to be problem specific depending on what our observations are as well as what the likelihood model is.

We will work within the Bayesian framework, so we must place a prior on our unknowns Z and θ. We assume an independent prior $p(Z, \theta) = p(Z)p(\theta)$, so we must therefore define:

- $p(X|Z, \theta)$: The likelihood model. This will always be application specific and must be tailored to X.

- $p(\theta)$: The prior on θ. Again this will be application specific.

- $p(Z)$: The prior on Z. This can be developed independent of the application and is the focus of much of this dissertation, but as we discuss starting in Chapter 4, it is often helpful if this is adapted to be more suitable for particular assumptions about our observations.

Then, given our observations X, our goal will be to infer Z and θ via Bayes's rule:

$$p(Z, \theta|X) \quad \propto \quad p(X|Z, \theta)p(\theta)p(Z).$$

Sometimes we are interested in the posterior distribution of Z and θ itself and sometimes we wish to use it to do prediction on any parts of X we might not have observed base on the parts we have observed. We do this by integrating over the posterior distribution

$$p(X_{\text{unobserved}}|X_{\text{observed}}) = \int p(X_{\text{unobserved}}|Z,\theta)dP(Z,\theta|X_{\text{observed}}).$$

Since $p(X|Z,\theta)$ and $p(\theta)$ are application specific, our main focus will be on $p(Z)$ until we get to Chapter 6, the applications chapter.

2.1.4 Exchangeability and De Finetti's Theorem

Before we dive into the math needed to define the two Bayesian nonparametric latent feature priors, we first discuss *exchangeability*, a property which will be used in the basic prior for Z. Recall that z_i refers to the i^{th} row of Z.

Definition 2.1.1. A finite set of observations $z_1, \ldots, z_N$ is *exchangeable* if every permutation of $z_1, \ldots, z_N$ has the same joint distribution as every other permutation. An infinite collection is called exchangeable if every finite subcollection is exchangeable. (Schervish, 1995)

This just means that for every set of entities, the order we see them should not affect our prior belief about the kinds of Z. This is often a valid assumption and is widely used in many probabilistic models. All of the initial development of Bayesian nonparametric latent feature models has assumed infinite exchangeability and this has a very important consequence, De Finetti's theorem.

Theorem 2.1.2 (De Finetti's theorem)**.** *$\{z_i\}_{i=1}^{\infty}$ is infinitely exchangeable if and only if there is a random probability measure P such that for any n,*

$$p(z_1, \ldots, z_n) = \int \left[\prod_{i=1}^{n} p(z_i|\theta)\right] dP(\theta).$$

(Schervish, 1995)

Therefore, if we know that Z is exchangeable, which is our assumption here, then there is a distribution P known as the De Finetti mixing distribution such that, conditional on θ drawn from P, the z_i are i.i.d. For the priors we consider here, the mixing distribution will be a Lévy process, which is the subject of the next section.

In Chapter 4, we will discuss models in which the exchangeability assumption is violated and how to adapt our priors to handle this fact. Until then, we will assume all models are exchangeable.

2.2 Lévy Processes

As mentioned in Chapter 1, priors for Bayesian nonparametric models work by defining an infinite-dimensional stochastic process that serves as a prior on our unknown parameter ϕ. In the case of latent feature priors, this unknown parameter ϕ is the Z described in previous sections. It turns out that the infinite-dimensional stochastic processes that we will define as priors on Z are special cases of a much older family of stochastic processes known as *Lévy processes*. There are many books and papers written about Lévy processes, but for a background sufficient for the topics discussed here, we have consulted the combination of books by Durrett (2004), Fristedt and Gray (1996), Sato (1999), and Kingman (1993). This section reviews the relevant background so that in Sections 2.3 and 2.4 we can discuss all relevant aspects of latent feature priors without needing to review these concepts.

2.2.1 Definitions and Theorems

Definition 2.2.1. A *Lévy process* in $\mathbb{R}$ or $\mathbb{R}^+$, respectively, is a right-continuous function Y from $[0, \infty)$ to $\mathbb{R}$ or $\mathbb{R}^+$ for which $Y_0 = 0$ a.s. and Y has stationary, independent increments. Let Y_t be the value of Y at t. (Fristedt and Gray, 1996)

Note that since Lévy processes have stationary, independent increments, they are infinitely divisible. For priors for Bayesian nonparametric latent feature models, we are only interested in the special case of Lévy processes in $\mathbb{R}^+$, which are non-decreasing functions also known as *subordinators*.

Lévy processes have a special representation and decomposition that are very important for understanding and simulating them. First we need a definition.

Definition 2.2.2. A measure ν on $\mathbb{R} \setminus \{0\}$ is called a *Lévy measure* if

$$\int_{\mathbb{R}\setminus\{0\}} (y^2 \wedge 1)\nu(dy) \quad < \quad \infty. \tag{2.1}$$

A measure ν on $\mathbb{R}^+$ is called a *Lévy measure* if

$$\int_{(0,\infty)} (y \wedge 1)\nu(dy) \quad < \quad \infty. \tag{2.2}$$

(Fristedt and Gray, 1996)

This means that ν is a Lévy measure if for all ϵ, there is finite mass more than ϵ away from zero. ν is allowed to have infinite mass near the origin, but Equation (2.1) defines how fast ν is allowed to grow near the origin. Given this definition, we can now introduce the Lévy-Khinchine representation of Lévy processes.

Theorem 2.2.3 (Lévy-Khinchine Representation Theorem)**.** *There is a one-to-one correspondence between all infinitely divisible distributions (and therefore Lévy processes) Y and the set of triples (η, σ, ν) where $\eta \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, and ν is a Lévy measure such that for all t, the characteristic function of Y_t is*

$$E[e^{iuY_t}] \quad = \quad \exp(-\psi(u,t))$$

where

$$\psi(u,t) \quad = \quad -i\eta ut + t\frac{\sigma^2 u^2}{2} + t\int_{\mathbb{R}\setminus\{0\}} (1 - e^{iuy} + it\chi(y))\nu(dy)$$

where $\chi(\cdot)$ can take one of several forms, one of which is

$$\chi(x) \quad = \quad (x \wedge 1) \vee (-1). \tag{2.3}$$

A special case of this occurs when Y is a subordinator. In this case, there is a one-to-one correspondence between all infinitely divisible functions on $\mathbb{R}^+$ and pairs (ξ, ν) where $\xi \in \mathbb{R}^+$ and ν is a Lévy measure for $\mathbb{R}^+$ such that for all t, the moment generating function of Y_t is

$$E[e^{-uY_t}] \quad = \quad \exp(-\theta(u,t)) \tag{2.4}$$

where

$$\theta(u,t) \quad = \quad \xi ut + t\int_{(0,\infty)} (1 - e^{-uy})\nu(dy). \tag{2.5}$$

(Fristedt and Gray, 1996)

There are also multivariate extensions of this representation that can be found in Sato (1999).

By examining the components of ψ and θ and matching them to characteristic and moment generating functions of known distributions, one can guess at least part of the next result, the Lévy-Itō decomposition. The first part of ψ corresponds to the characteristic function of a constant function, the second part is from a Gaussian distribution, and the last part is related to a compound Poisson process. Intuitively, when specializing to the case of the subordinator, the Gaussian part disappears and the last term reduces exactly to a compound Poisson process.

Theorem 2.2.4 (Lévy-Itō Decomposition Theorem)**.** *Let Y_t be a Lévy process on $\mathbb{R}$ with triple (η, σ, ν) as defined in Theorem 2.2.3. Let (X, W) be an independent pair where W is standard Brownian motion and X is a Poisson point process in*

$(0,\infty)\times(\mathbb{R}\setminus\{0\})$ *whose intensity measure is* $\lambda\times\nu$ *where* λ *is the Lebesgue measure. Then there exists a sequence of* $\epsilon_k\downarrow 0$ *such that*

$$Y_t \stackrel{d}{=} \eta t+\sigma W_t+ \lim_{k\to\infty}\left[\int_{(-\infty,-\epsilon_k]\cup[\epsilon_k,\infty)} xX((0,t]\times dx)-t\int_{(-\infty,-\epsilon_k]\cup[\epsilon_k,\infty)} x\nu(dx)\right] \tag{2.6}$$

where $\chi(\cdot)$ *is as in Equation* (2.3).

For the case when Y_t *is a subordinator, then* ν *satisfies Equation* (2.2)*, and the Lévy-Itō decomposition simplifies. Now let* Y_t *be a subordinator and the pair* (ξ,ν) *be as defined in Theorem 2.2.3. Let* X *be a Poisson process in* $(0,\infty)\times(0,\infty]$ *whose intensity measure is* $\lambda\times\nu$ *where* λ *is again the Lebesgue measure. Then*

$$Y_t \stackrel{d}{=} \xi t+\int xX((0,t]\times dx). \tag{2.7}$$

(Fristedt and Gray, 1996)

We can see that the first term in Equation (2.6) corresponds to a drift term and the second term corresponds to standard Brownian motion. The final term is a compound Poisson process centered about zero by subtracting off the mean. This limit is called a compensated sum of jumps and for this reason, the Lévy measure ν is also referred to as a compensator. The reason for the compensation is that the limit does not necessarily exist unless we compute the indicated difference. This last term is often broken up into two parts, a compound Poisson process defined on $(-\infty,-\epsilon]\cup[\epsilon,\infty)$ for some $\epsilon>0$ and a compensated Poisson process defined on $(-\epsilon,\epsilon)$.

In the case of subordinators, which is the case we are interested in, things become even simpler. By the Lévy-Itō decomposition, a subordinator is the combination of a constant drift term and a compound Poisson process. The priors we construct only rely on pure-jump processes, so we further simplify to only looking at the compound Poisson process part. Specializing all the earlier results to this case, this means that for these pure-jump subordinator, Y_t, the only unknown is ν, a Lévy measure satisfying Equation (2.2). We also reinterpret Y_t to be $Y(0,t]$, the measure assigned to the interval $(0,t]$. Since Y can be constructed from a non-negative Poisson process, this defines a random measure and we can more generally talk about $Y(A)$ for Borel sets A. Since Y has stationary independent increments, $Y(A)$ and $Y(B)$ are independent for disjoint sets A and B and identically distributed if $\lambda(A)=\lambda(B)$.

Using the convention that the Bayesian nonparametric latent feature community has adopted, we redefine ν to be what was previously the product measure $\lambda\times\nu$. By doing so, we can allow λ to be different than the Lebesgue measures on some space Ω, but if it is not a multiple of the Lebesgue measure, the resulting process will not

be a Lévy process since it will not have stationary increments. Instead, it will be a *completely random measure* as defined by Kingman (1967).

Definition 2.2.5. A random measure Φ is a *completely random measure* if, for any finite collection $A_1, \ldots, A_n$ of disjoint sets, the random variables $\Phi(A_1), \Phi(A_2), \ldots, \Phi(A_n)$ are independent. (Kingman, 1967)

All of the following results hold for both pure-jump completely random measures and pure-jump subordinators and for this reason, some people prefer to discuss completely random measures since they allow for more flexible, non-stationary representations.

For Y defined as above, the joint measure $\nu(d\omega, dy)$ is now defined on $\Omega \times [0, \infty)$ and the Lévy-Khinchine representation theorem now says that

$$E[e^{-uY(A)}] \;=\; \exp\left(-\int_{(0,\infty)}\int_{A}(1-e^{-uy})\nu(d\omega, dy)\right). \tag{2.8}$$

It is easy to see that for the special case of the Lebesgue measure for what used to be λ, this reduces to Equations (2.4) and (2.5).

The corresponding Lévy-Itō decomposition says that there is a Poisson process X on $\Omega \times [0, \infty)$ with intensity measure ν such that

$$Y(A) \;\stackrel{d}{=}\; \int xX(A \times dx).$$

Therefore, to define our priors, we must identify ν and then know how to simulate from these Lévy processes and perform posterior updates on them.

In order to perform posterior updates, we must address one more issue and that is the issue of having atoms in ν. Atoms are not allowed in the base measure of Lévy processes since we must have stationary increments, but they are permitted in completely random measures. Furthermore, in completely random measures, since what happens in any location or subset of locations is independent of what happens elsewhere, we can always separately reason about the continuous part of ν and any atoms. We will let ν' be the part of the Lévy measure with atoms. We therefore assume that when discussing ν for the rest of this section, it has no atoms and that we are independently reasoning about the atoms produced by ν'.

How do we deal with atoms in ν'? Let Y' be a completely random measure with a Lévy measure ν' with discrete support. The atoms in ν', the points at which $\nu'(\{\omega\}, [0, \infty)) > 0$, are known as *fixed points of discontinuity* because they are the locations at which $p(Y'(\{\omega\}) > 0) > 0$.

For fixed points of discontinuity in the case of the beta and to-be-introduced Bernoulli processes, we must have that $\nu'(\{\omega\}, (0, \infty)) \leq 1$. Call the set of all these

points $\mathcal{D}$. We can then construct Y' as $Y' = \sum_{\omega \in \mathcal{D}} p_\omega \delta_\omega$. ν' determines the values of p_ω as

$$\begin{aligned} p(p_\omega \in A) &= \nu'(\{\omega\} \times A) \qquad \forall A \subset (0, \infty) \\ p(p_\omega = 0) &= 1 - \nu'(\{\omega\} \times (0, \infty)). \end{aligned}$$

For the gamma and Poisson processes defined in Section 2.4.1, we will have a different, be equally simple relationship. Therefore, reasoning about the discontinuities in ν' is straightforward. For the rest of this section, we focus on the continuous part ν.

2.2.2 Lévy Process Take-Away Message

Section 2.2.1 had some rather technical definitions and theorems that are good background knowledge when working with priors Bayesian nonparametric latent feature models. If we are not concerned with all the mathematical detail, what is the take-away message?

The take-away message is that when working with random measures defined by pure-jump non-negative Lévy processes or a completely random process, we must only identify ν. Then we can equivalently work with a Poisson process with intensity measure ν.

To better understand all these definitions, we now visualize pure-jump non-negative Lévy processes and how these relate to Poisson processes with base measure ν. For concreteness, we define ν to be the Lévy measure for the beta process defined on a space $\Omega \times [0, 1]$, which we will fully introduce in Section 2.3.1,

$$\nu(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp B_0(d\omega). \tag{2.9}$$

Here B_0 is known as the base measure (instead of using the Lebesgue measure) and c is known as the concentration parameter. In general, c can be a function of ω, but this is not commonly used in latent feature priors. Note that by using a more generic B_0, the domain Ω can also be more general than $[0, \infty)$ and in the cases when B_0 is not a constant multiple of the Lebesgue measure, the name "completely random measure" is more appropriate than Lévy process. Also, the term "subordinator" is used only when the domain is $[0, \infty)$.

Let X be a Poisson process with base measure ν on the space $\Omega \times [0, 1]$ and for concreteness, let B_0 be the uniform distribution on $[0, 1]$, thus restricting X to be a Poisson process on $[0, 1] \times [0, 1]$. The measure ν can be seen in Figure 2.2(a). Note that since ν has the term $cp^{-1}(1-p)^{c-1}$, it is an infinite, improper beta measure in p and therefore has infinite mass. The result is that the Poisson process X drawn with intensity measure ν seen in Figure 2.2(b) has a countably infinite number of points.

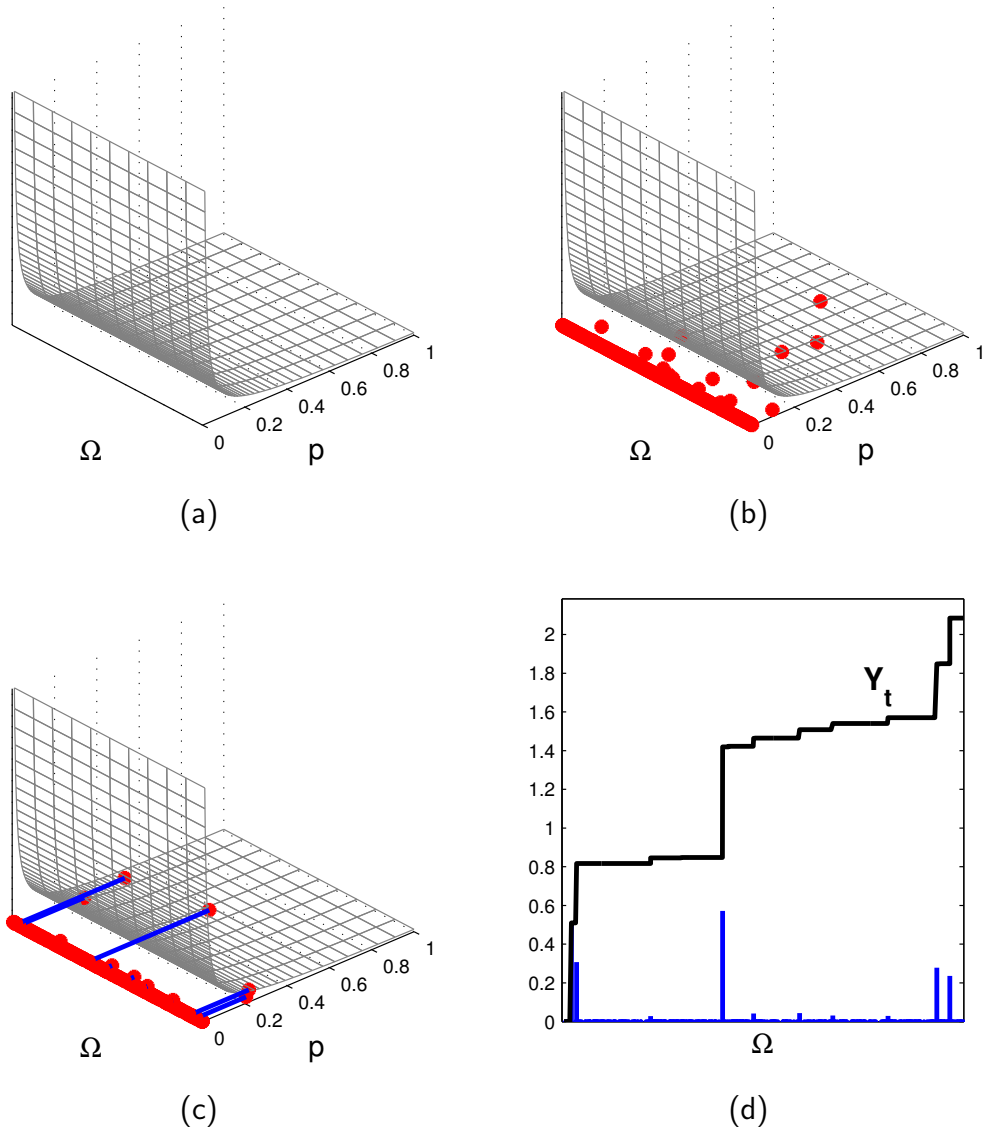


Figure 2.2: Visualization of a Lévy process. (a) The Lévy measure ν from Equation (2.9) with B_0 the uniform distribution on $[0, 1]$. (b) A random Poisson process X drawn with intensity measure ν. Note that since ν is improper, it has infinite mass, so X has a countably infinite there are a countably infinite number of points. By Campbell's theorem (Section 2.2.3), though, for any $\epsilon > 0$, there are only a finite number of points with p greater than ϵ. (c) The Poisson process X with the heights of p made explicit since those are what we sum over in Equation (2.10) to get Y. (d) The corresponding Lévy process where Y is the results of summing over the increments corresponding to the heights of X, shown in blue.

By Campbell's theorem (Section 2.2.3), though, for any $\epsilon > 0$, there are only a finite number of points with p greater than ϵ. By the Lévy-Itō decomposition, Y is the result of integrating the heights (p) of all the points in X. We make the heights of X explicit in Figure 2.2(c) and show the resulting Y in Figure 2.2(d).

Since Y is equivalent to the integral of a discrete Poisson process which has a countably infinite number of points, we can represent Y as a discrete measure:

$$Y = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}, \tag{2.10}$$

where $\{(\omega_k, p_k)\}_{k=1}^{\infty}$ are the random points of the Poisson process X. Due to the definition of ν, all p_k must be in the range of $[0, 1]$.

Thus, we have gone from mathematically elegant, but potentially complex Lévy processes to the special case of pure-jump non-negative Lévy processes in which we only need to define the Lévy measure ν which in turn gives us the discrete representation as seen in Figure 2.2 and made concrete by Equation (2.10).

2.2.3 Campbell's Theorem

As we have mentioned, X will have an infinite number of points when ν is an improper Lévy measure. How can we be sure Y is a finite measure? This is addressed by Campbell's Theorem.

Theorem 2.2.6 (Campbell's Theorem)**.** *Let X be a Poisson process on $\Omega \times (0, \infty]$ with mean measure ν. Then the sum*

$$\Sigma = \sum_{(\omega,p)\in X} p$$

is absolutely convergent if and only if

$$\int_{\Omega} \int_{0}^{\infty} p \cdot \nu(d\omega, dp) < \infty.$$

(Kingman, 1993)

In the beta process as well as the gamma process introduced in Section 2.4.1, we start with an improper beta and gamma distribution in the Lévy measure, but by multiplying by p in Campbell's theorem, they become integrable, so Y has finite measure.

2.2.4 Inverse Lévy Measure

The final issue we wish to address before moving on to priors for the latent feature models is how to get we samples from X given ν? The *inverse Lévy measure* algorithm by Wolpert and Ickstadt (1998b) is a generic inference technique for pure-jump non-negative Lévy processed that allows us to sample X by generating the points in X in decreasing order of p. For particular instances of ν, as we will discuss shortly, the inverse Lévy measure has a closed form solution, but it often does not.

We start by assuming ν decomposes into a product measure on $\Omega \times (0, \infty]$

$$\nu(d\omega, dp) \quad = \quad A(dp)B(d\omega).$$

This is a valid assumption for all latent feature priors discussed in this dissertation. Let $\alpha = B(\Omega)$. We will generate $X = \{(\omega_i, p_i)\}_{i=1}^{\infty}$ in decreasing order of p_i. Since this is a product measure, we sample the p_i independently from a Poisson process with intensity αA and then sample each ω_i independently from B/α.

We are therefore just left with sampling a Poisson process with intensity αA in decreasing order. We can do this by first generating points τ_i from a standard unit-rate Poisson process on $[0, \infty)$. We then use the mapping theorem for Poisson processes (Kingman, 1993) to transform these points into p_i. This is as simple as setting

$$p_i \quad = \quad \inf\{u \geq 0 : \alpha A([u, \infty)) \leq \tau_i\}. \tag{2.11}$$

For general ν, this will not have a closed form solution, but it is a single dimensional estimation problem that can be solved numerically.

To make this more concrete, we will describe the inverse Lévy measure for the beta process ν in Equation (2.9) when $c = 1$, a special case for which an elegant closed form solution is known:

$$\nu(d\omega, dp) \quad = \quad p^{-1}dpB_0(d\omega). \tag{2.12}$$

Let $\alpha = B_0(\Omega)$. Then given the points $\{\tau_i\}_{i=1}^{\infty}$ from a unit-rate Poisson process on $[0, \infty)$, we convert these to points from a Poisson process with rate ν as shown in Figure 2.3 and discussed below.

We start by calculating the distribution of p_1. ν is continuous and defined on $[0, 1]$, so plugging in ν into Equation (2.11) gives us p_1 satisfies

$$\begin{aligned} \tau_1 &= \alpha \int_{p_1}^{1} p^{-1}dp \\ &= -\alpha \log(p_1). \end{aligned}$$

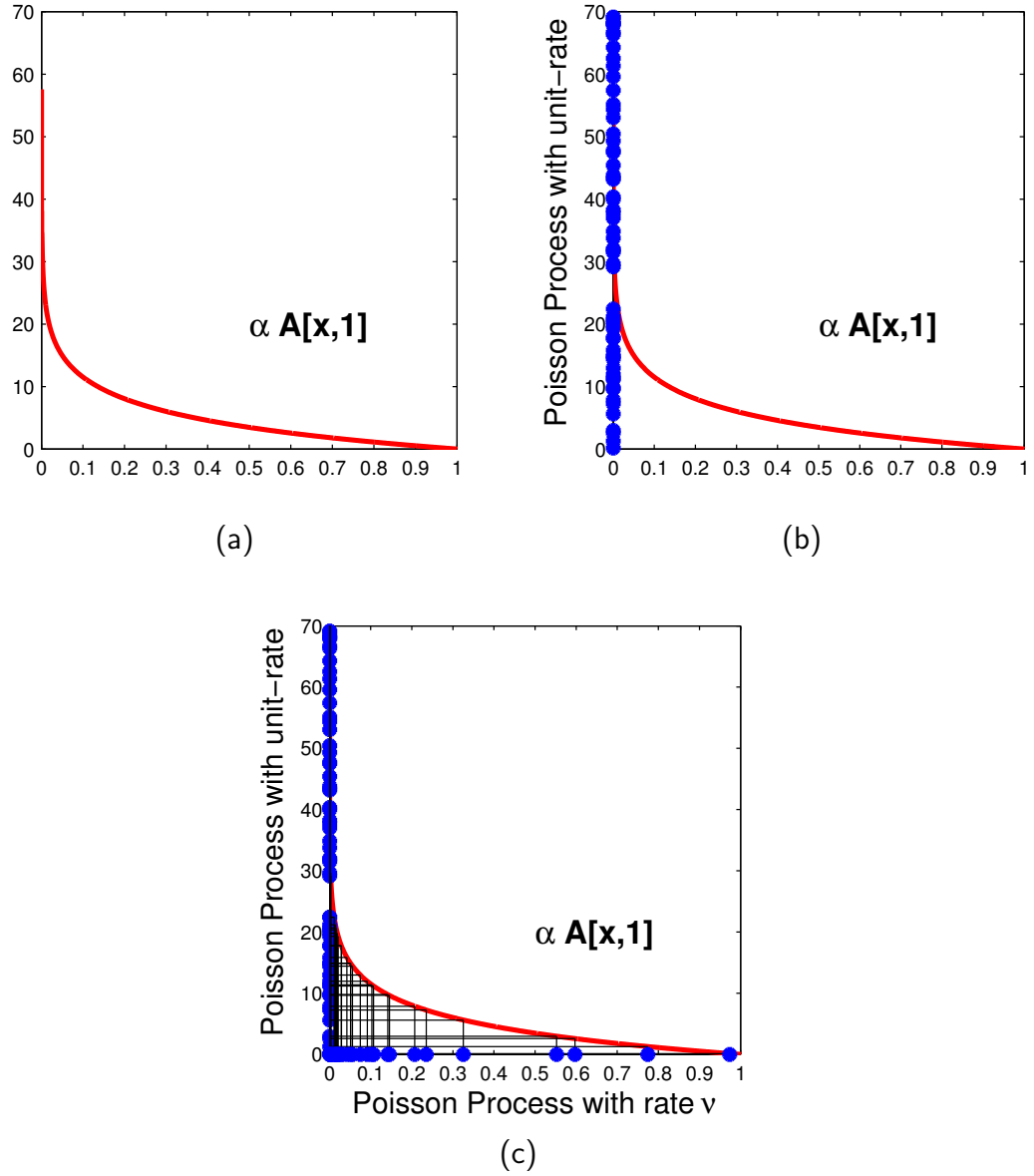


Figure 2.3: A demonstration of the inverse Lévy measure algorithm with ν as in Equation (2.12). (a) $\alpha A[x, 1]$. Note that $\alpha A[1, 1] = 0$ and for $\epsilon > 0$, $\lim_{\epsilon \downarrow 0} \alpha A[\epsilon, 1] = \infty$. (b) A unit-rate Poisson process τ is shown along the y-axis extending to ∞. (c) Finally, to get our desired Poisson process, p with intensity measure ν, we use the continuous mapping theorem to map $\{\tau_i\}_{i=1}^{\infty}$ to $\{p_i\}_{i=1}^{\infty}$ though Equation (2.11). This corresponds to finding p_i such that $\alpha A[p_i, 1] = \tau_i$ for each i. Since there are an infinite number of points in $\{\tau_i\}_{i=1}^{\infty}$, there are an infinite number of points in $\{p_i\}_{i=1}^{\infty}$, almost all of which are arbitrarily close to the origin.

Since τ_1, the first point in a unit-rate Poisson process is an Exponential(1) random variable, using the change of variable formula, the distribution of p_1 is then calculated to be

$$\begin{aligned} p(p_1) &= p(\tau_1(p_1)) \left| \frac{d\tau_1(p_1)}{dp_1} \right| \\ &= \exp\left(\alpha \log(p_1)\right) \left| \frac{d}{dp_1} \left(-\alpha \log(p_1)\right) \right| \\ &= \alpha p_1^{\alpha-1}. \end{aligned}$$

In other words, $p_1 \sim \text{Beta}(\alpha, 1)$.

We now compute the distribution $p(p_i|p_{i-1})$. Given, τ_{i-1}, in the unit-rate Poisson process, we know that the distribution of $\tau_i - \tau_{i-1}$ is again an Exponential(1) random dom variable. As before, we can calculate $\tau_i = -\alpha \log(p_i)$ and $\tau_{i-1} = -\alpha \log(p_{i-1})$. Therefore,

$$\begin{aligned} p(p_i|p_{i-1}) &= p(\tau_i(p_i)|\tau_{i-1}(p_{i-1})) \left| \frac{d\tau_i(p_i)}{dp_i} \right| \\ &= \exp\left(\alpha \log(p_i) - \alpha \log(p_{i-1}\right) \left| \frac{d}{dp_i} \left(-\alpha \log(p_i)\right) \right| \\ &= \alpha \frac{p_i^{\alpha-1}}{p_{i-1}^{\alpha}}. \end{aligned}$$

Note that since $p_i < p_{i-1}$, then we can also look at the ratio $v_i = p_i/p_{i-1}$. Then

$$\begin{aligned} p(v_i|p_{i-1}) &= p(p_i(v_i)|p_{i-1}) \left| \frac{dp_i(v_i)}{dv_i} \right| \\ &= \alpha \frac{(p_{i-1}v_i)^{\alpha-1}}{p_{i-1}^{\alpha}} \left| \frac{d}{dv_i} p_{i-1}v_i \right| \\ &= \alpha v_i^{\alpha-1}. \end{aligned}$$

So the ratio $v_i = p_i/p_{i-1}$ is independent of the value of p_{i-1} and is $\text{Beta}(\alpha, 1)$ distributed. In other words, in the special case of ν in Equation (2.12), we can sample p_i in decreasing order by sampling

$$\begin{aligned} v_i &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1) \\ p_i &= \prod_{k=1}^{i} v_k. \end{aligned}$$

This gives us the stick breaking representation for the beta process that we discuss in Section 2.3.2, which was originally derived in an alternate manner.

2.3 Priors for Binary Latent Feature Models

We are now finally ready to introduce our first prior for Bayesian nonparametric latent feature models. This section reviews priors for binary latent features. Section 2.4 reviews priors for non-negative integer valued latent features.

In Section 2.1, we motivated the development of a prior $p(Z)$ on $N \times K$ binary matrices. However, since Z represents unobserved features that we wish to infer, we rarely know K, so in practice, we would like a prior that places positive mass on all possible $N \times K$ binary matrices, whatever the value of K is. We do this by developing a prior on these latent feature matrices such that K is allowed to be infinite, but such that all mass is placed on matrices having a finite number of non-zero entries. If we do this, then we will have a prior for a Bayesian nonparametric binary latent feature model.

In the next three subsections, we develop three equivalent ways to define an exchangeable stochastic process that serves as a nonparametric prior on Z. In Section 2.3.1, we introduce how the beta process (BP) can be used to create a prior on Z. We then show how we can use the inverse Lévy measure and the corresponding stick breaking process to more practically do this in Section 2.3.2. Understanding how the beta process is used to generate priors on Z, we then show how we can marginalize out the beta process to directly get a prior on Z known as the Indian buffet process (IBP) in Section 2.3.3. We then review at a very high level several extensions to the basic priors that have resulted from these developments in Section 2.3.4.

The order we present the ideas here is helpful for understanding, but different than the historical order they were developed. The BP was originally developed by Hjort (1990) and later clarified by Kim (1999) in the field of Bayesian survival analysis. The IBP was independently developed by Griffiths and Ghahramani (2006) motivated by a prior for Bayesian nonparametric latent class models, the so-called Chinese restaurant process that we discuss in Section 2.4.3. It was not until later that anyone realized the IBP and BP were related. This connection was made clear by Thibaux and Jordan (2007) as we discuss in Section 2.3.3. The stick breaking process turns out to be identical to the inverse Lévy measure of the beta process, but was developed independently based on the IBP by Teh et al. (2007) in parallel with Thibaux and Jordan (2007) discovering the relationship between the BP and the IBP.

2.3.1 The Beta Process

We introduced the beta process (BP) in Section 2.2.2. The BP is the Lévy process with the Lévy measure in Equation (2.9)

$$\nu(d\omega, dp) \quad = \quad cp^{-1}(1-p)^{c-1}dpB_0(d\omega)$$

where B_0 is our base measure and c is a concentration parameter. In addition, we showed in Figure 2.2 what this process looks like when B_0 is the Uniform$[0, 1]$ measure. This process was originally developed by Hjort (1990) using a limiting argument for beta distributions. It later was clarified by Kim (1999). The developments in this section follow the work by Thibaux and Jordan (2007) who originally showed how the BP could be used as a nonparametric latent feature prior and whose work more closely followed Kim (1999) than Hjort (1990). This work was later mirrored by Paisley and Carin (2009), whose work was primarily influenced by the BP representation by Hjort (1990).

We rename the random measure Y generated by the BP to be B as is done by Thibaux and Jordan (2007). Using Equation (2.9), we can see that

$$B \quad = \quad \sum_{k=1}^{\infty} p_k \delta_{\omega_k},$$

where $\{\omega_k, p_k\}_{k=1}^{\infty}$ are the countably infinite points generated by a Poisson process with intensity ν. Each of the atoms in B then has location ω_k and weight p_k where $p_k \in [0, 1]$. For a set S, this gives us the measure $B(S) = \sum_{k:\omega_k \in S} p_k$. We write $B \sim \text{BP}(c, B_0)$ to denote that B is drawn from a beta process. We formalize this in the next definition.

Definition 2.3.1. Let B_0 be a measure on Ω and $c > 0$ be a concentration parameter. The *beta process* B, written $B \sim \text{BP}(c, B_0)$, is the completely random measure with Lévy measure

$$\nu(d\omega, dp) \quad = \quad cp^{-1}(1-p)^{c-1}dpB_0(d\omega).$$

We then define the Bernoulli process to be:

Definition 2.3.2. Let H be a measure on Ω. The *Bernoulli process* z with *hazard rate* H, written $z \sim \text{BeP}(H)$, is the completely random measure with Lévy measure

$$\nu(d\omega, dp) \quad = \quad \delta_1(dp)H(d\omega)$$

where δ_1 is the delta function at one.

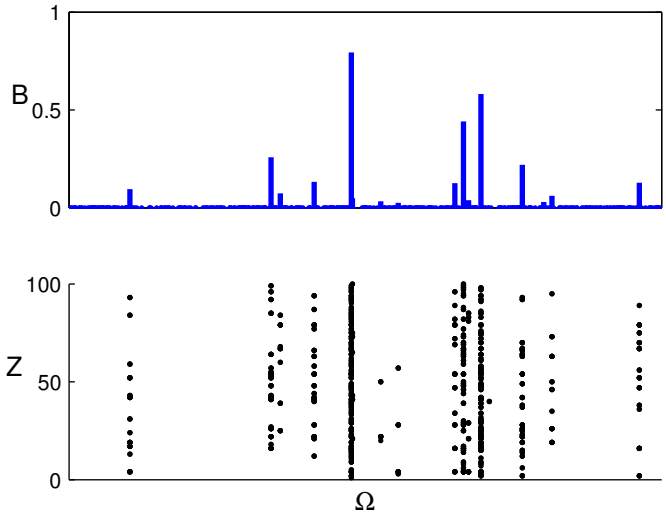


Figure 2.4: We visualize $B \sim \mathrm{BP}(c, B_0)$ in the top figure and directly below it draw $z_i \sim \mathrm{BeP}(B)$ for $i = 1, \ldots, 100$. A black dot in the i^{th} row indicates that z_i has the corresponding feature. As we can see, ω_i with a large p_i naturally are present in more z_i.

By having the delta function at one, all points in z must have value one. All points that do not appear in z have value zero. If H is a continuous measure, this means the locations of the atoms of z are a Poisson process with intensity H. This takes the form $z = \sum_{k=1}^{N} \delta_{\omega_k}$ where $N \sim \mathrm{Poisson}(H(\Omega))$ and each of the ω_k are sampled i.i.d. from the distribution $H/H(\Omega)$. If H is discrete and takes the form $H = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$, then we define z to be $z = \sum_{k=1}^{\infty} b_k \delta_{\omega_k}$ where $b_k \sim \mathrm{Bernoulli}(p_k)$. This is straightforward from our discussion of fixed points of discontinuity in Section 2.2.1.

Putting all of this together, our full stochastic process for generating the binary latent feature matrix Z is

$$\begin{aligned} B &\sim \mathrm{BP}(c, B_0) \\ z_i | B &\sim \mathrm{BeP}(B) \qquad i = 1, \ldots, N. \end{aligned}$$

We visualize this in Figure 2.4.

Since z_i takes on values zero and one, we can store it in an infinitely long binary row vector where each column corresponds to ω_k. In addition, since B has finite mass by Campbell's theorem, z_i itself will have a finite number of non-zero entries, so we can store all rows of $Z = \{z_1, \ldots, z_N\}$ in a finite amount of space by only storing non-zero features.

Now we can see that *any* binary matrix Z with N rows and a finite number of non-zero columns will have positive probability under this prior, meaning we have defined a valid nonparametric prior. This is our first step towards having a useful nonparametric prior. What do we do with all the $\{\omega_i\}$? Often, each feature will have an associated parameter which we can let be ω_i, so B_0 is the prior on these feature values. Or, if we let B_0 be uniform on $[0, 1]$, we can use ω_i as random number generator for computing values associated with each feature. For simplicity, when discussing the Indian buffet process in Section 2.3.3, we leave out mentions of ω_i, but we can always add them back in. In Chapter 3, we discuss how we can use this prior to perform posterior inference, allowing this to be used in practical models, and in Chapter 6, we complete the story by demonstrating applications of these priors.

Before continuing on to other representation of this prior on Z, we address posterior updates to B based on Z as this will be useful in both Section 2.3.3 as well as when we discuss inference algorithms in Chapter 3.

If we observe $z_1, \ldots, z_N$, what is our posterior belief about B? Theorem 3.3 of Kim (1999) answered this question. In the notation of Thibaux and Jordan (2007), the posterior takes the form:

$$\begin{aligned} B|z_1, \ldots, z_N &\sim \mathrm{BP}(c+N, B_N) \qquad (2.13) \\ \text{where } B_N &= \frac{c}{c+N}B_0 + \frac{1}{c+N}\sum_{i=1}^{N} z_i = \frac{c}{c+N}B_0 + \sum_j \frac{m_{N,j}}{c+N}\delta_{\omega_j} \end{aligned}$$

where ω_j are the set of all atoms present in $z_1, \ldots, z_N$ and $m_{N,j} = \sum_{i=1}^{N} z_i(\{\omega_j\})$ is the number of times ω_j occurs in all the z_i. In other words, in the posterior the weight of the continuous measure B_0 is reduced and we add in a discrete measure located at the atoms of the z_i weighted by how many times they occur.

An equivalent, but slightly clearer way of thinking about this that more closely parallels the presentation of Kim (1999) is that we start off with a beta process B with Lévy measure

$$\nu(d\omega, dp) = cp^{-1}(1-p)^{c-1}dpB_0(d\omega).$$

We then draw z_i from the resulting discrete measure. For each atom in B, we draw ω_j with probability p_j. If we have drawn a point ω_j a total of $m_{N,j} > 0$ times after N draws, then the the posterior distribution of p_j at ω_j is the Beta$(m_{N,j}, N - m_{N,j} + c)$ distribution

$$\frac{\Gamma(N+c)}{\Gamma(m_{N,j})\Gamma(N-m_{N,j}+c)}p^{m_{N,j}-1}(1-p)^{N-m_{N,j}+c-1},$$

just as if we had started off with a Beta$(0, c)$ prior and then observed N observations $z_1, \ldots, z_N$ from a Bernoulli distribution. Since we have observed at least one z_i at location ω_j, we know that B must have had an atom at ω_j, resulting in this discrete posterior.

For the continuous part of the posterior Lévy measure, the probability of not seeing any z_i at some ω_j is $(1 - p_j)^N$, so the continuous part of the Lévy measure is

$$cp^{-1}(1-p)^{c+N-1}dpB_0(d\omega).$$

This is similar to starting with a Beta$(0, c)$ prior and then not observing anything after N attempts. The resulting posterior is still improper, thus resulting in a posterior beta process. Combining these two parts gives us the posterior Lévy measure.

2.3.2 The Stick Breaking Process

The beta-Bernoulli process prior on Z presented in the previous section tells us the distribution of B and Z, but it does not tell us how to actually generate samples form this prior. In order to use it in models, we need to know how to get samples of both B and Z. As discussed in Section 2.2.4, the inverse Lévy measure algorithm of Wolpert and Ickstadt (1998b) is precisely such a way to draw the atoms of $B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$ in strictly decreasing order of p_k. To do this, we let $\alpha = B_0(\omega_k)$ and assume $c = 1$, which is the special case corresponding to the Indian buffet process introduced in the next section[1]. Then we draw

$$\begin{aligned} v_k &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1) \\ p_k|\{v_i\}_{i=1}^{\infty} &= \prod_{i=1}^{k} v_i \\ \omega_k &\overset{\text{i.i.d.}}{\sim} B_0/\alpha. \end{aligned}$$

This is referred to as the stick breaking process because we start with a stick of length one. Then we break off a Beta$(\alpha, 1)$ piece of this stick. This is p_1. Then we break off a Beta$(\alpha, 1)$ fraction of p_1 and call this p_2, and continue this process where each p_k is a Beta$(\alpha, 1)$ fraction of p_{k-1}. This construction is due to Teh et al. (2007) who derived it based on a limiting argument for the IBP, although the derivation in Section 2.2.4 is more direct. An alternate stick breaking construction based on a different limiting argument has also been derived by Paisley et al. (2010). This construction has recently been rederived and extended by Broderick et al. (2011) by directly examining the underlying random measure rather than with a limiting

[1]For general c, there is no closed form stick breaking process.

argument.

We can continue the stick breaking process arbitrarily long, although for practical reasons, we are only able to store a finite number of atoms of B. Given these atoms, it is very easy to sample the Bernoulli process z_i via a Bernoulli sampler. This gives us a practical way to get samples of B and Z.

2.3.3 The Indian Buffet Process

We have seen how to generate $Z = (z_1, \ldots, z_N)$ via

$$p(z_1, \ldots, z_N) = \int \left[\prod_{i=1}^{N} p(z_i|B) \right] dP(B),$$

where B is drawn from the BP and z_i are drawn conditionally independently via the BeP. Due to these conditionally independent draws, the rows $(z_1, \ldots, z_N)$ are exchangeable as is seen by checking De Finetti's theorem in Section 2.1.4.

The Indian Buffet Process (IBP) is a way to draw Z directly without first needing to sample B developed by Griffiths and Ghahramani (2006). It was derived without any knowledge of the BP and it came as a revelation when the BP was shown to be the De Finetti mixing distribution for the IBP by Thibaux and Jordan (2007).

We first introduce the IBP itself, then we discuss how to derive in two different ways, the first using the beta-Bernoulli process, the second involving the original derivation without knowledge of the BP.

The Indian Buffet Process The IBP sets $c = 1$ in the BP and has a single parameter α which is equivalent to $B_0(\Omega)$ in the BP and is a culinary metaphor[2] describing how to generate the non-zero columns of Z. In this metaphor, each row of Z corresponds to a customer at an Indian buffet and each column corresponds to one of infinitely many dishes. z_{ik}, the entry at (i, k), will be one if the i^{th} customer tastes the k^{th} dish and zero otherwise. We fill in the matrix as shown in Figure 2.5 and as described without the culinary metaphor below.

To generate matrices Z from the IBP(α), start with an all-zero matrix and perform the following:

- In the first row, mark the first Poisson(α) columns as one. Leave the rest all zero.

[2]The Bayesian nonparametric community has been keen on finding culinary metaphors for new processes since the Chinese Restaurant Process was developed as a culinary metaphor for sampling from the Dirichlet Process.

- Now assuming we have filled in the first $i-1$ rows, fill in the i^{th} row as follows:
  - Look at each non-zero column. If there are m_k non-zero entries in the k^{th} column, set the (i,k) entry to one with probability m_k/i.
  - Now add an additional Poisson(α/i) ones after that last non-zero column.

See Figure 2.5 to visualize this.

We initially claimed that the IBP is an exchangeable prior, so that permuting the rows of Z does not change the probability of seeing Z, but it is clear given the description of the IBP and the figures in Figure 2.5 that as described, it is not exchangeable. This is remedied by remembering that the order of the columns themselves do not have any real meaning – they are just placeholders for some feature. In the culinary metaphor, the dishes themselves might have been arranged arbitrarily. So instead of placing a prior on Z itself, we are placing a prior on equivalence classes of Z, which we denote by $[Z]$. $[Z]$ is the class of all matrices that map to the same *left-ordered form*. The left-ordered form of Z is the matrix that results from permuting the columns so that they are sorted by their binary values. It is clear that multiple matrices have the same left-ordered form and that for all of these, permuting the columns does not affect how many features overlap between different people. When we place the IBP prior on these equivalence classes, then the prior is exchangeable. For details, see Griffiths and Ghahramani (2006).

If we wish for a parameter ω_k to be associated with each column as in the beta process, then after sampling Z, we can sample ω_k independently from $B_0/B_0(\alpha)$.

Derivation 1 We now derive how to get the IBP from the beta-Bernoulli process construction of $p(Z)$. To correspond exactly to the IBP, we set $c=1$, but leave c as a variable in the below derivation.

For the first person to enter the the restaurant corresponds to z_1. So we wish to sample z_1 from the distribution

$$p(z_1) \;=\; \int p(z_1|B)dP(B),$$

where $B \sim \text{BP}(c, B_0)$. Thibaux and Jordan (2007) showed that this marginal distribution of z_1 is $\text{BeP}(B_0)$. The reasoning is that by construction, z_1 takes values in $\{0,1\}$ and is a completely random measure, so it is a Bernoulli process. These are characterized entirely by their hazard rate, which is their expectation. In this case, the hazard function is therefore $\mathbb{E}[z_1] = \mathbb{E}[\mathbb{E}[z_1|B]] = \mathbb{E}[B] = B_0$.

As mentioned right after Definition 2.3.2, a Bernoulli process with a continuous hazard function is a Poisson process with $N \sim \text{Poisson}(B_0(\Omega))$ points where the

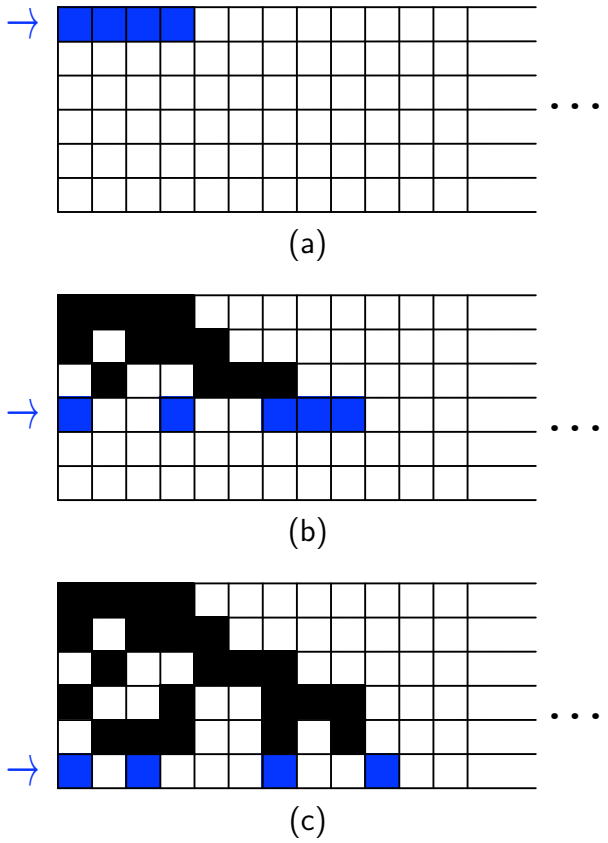


Figure 2.5: A demonstration of the Indian Buffet Process. We start with an empty (all-zero) matrix Z where each row corresponds to a person entering the buffet and each column corresponds to a dish. We fill in Z row by row. (a) The first customer to enter the restaurant tries the first $\text{Poisson}(\alpha)$ dishes, which is recorded by changing the corresponding entries of Z to one. (b) and (c) For the i^{th} customer, there are two steps. The first step is trying at all previously sampled dishes with probability proportional to the number of people who have previous tried it (details in the main text). The next step is to try a $\text{Poisson}(\alpha/i)$ number of new dishes.

locations of those points are drawn independently from $B_0/B_0(\Omega)$. Since we have defined $\alpha = B_0(\Omega)$, this means z_1 is a draw of Poisson(α) points as in the IBP.

For the i^{th} person to enter the restaurant, we wish to sample z_i from the posterior distribution

$$p(z_i|z_1,\ldots,z_{i-1}) \;=\; \int p(z_i|B)dP(B|z_1,\ldots,z_{i-1}).$$

Recalling Section 2.3.1, then by Equation (2.13), the posterior distribution of B given $z_1,\ldots,z_{i-1}$ is:

$$\begin{aligned} B|z_1,\ldots,z_{i-1} &\sim \text{BP}(c+i-1, B_{i-1}) \\ \text{where } B_{i-1} &= \frac{c}{c+i-1}B_0 + \frac{1}{c+i-1}\sum_{j=1}^{i-1} z_j = \frac{c}{c+i-1}B_0 + \sum_j \frac{m_{i-1,j}}{c+i-1}\delta_{\omega_j} \end{aligned}$$

By the same reasoning as for z_1, we can therefore show that $z_i \sim \text{BeP}(B_{i-1})$. Since B_{i-1} is a mixed continuous and discrete measure, we deal with z_i for the continuous and discrete parts separately.

For the discrete part, following the text after Definition 2.3.2, sampling from this is equivalent to sampling each ω_j independently with probability $\frac{m_{i-1,j}}{c+i-1}$. Here $m_{i-1,j}$ is the number of times feature j was present in $z_1,\ldots,z_N$, (i.e. m_k, the number of times the dish was previously sampled in the IBP description above) and $c = 1$, which means we sample each of the previously sampled dished with probability m_k/i.

For the continuous part $\frac{c}{c+i-1}B_0$ with $c = 1$, then similar to z_1, this corresponds to sampling a Poisson $\left(\frac{c}{c+i-1}B_0(\Omega)\right)$ = Poisson (α/i) number of new dishes.

Therefore the steps for sampling each z_i as derived by marginalizing out B are equivalent to those of the IBP and it becomes clear how the mix of a continuous and discrete posterior distribution of B directly affect the two aspects of sampling z_i in the IBP. Also, in this formulation, it is clear that the order of the columns is arbitrary, so this defines a prior on equivalence classes of matrices.

Derivation 2 The alternate derivation of the IBP is the original derivation in Griffiths and Ghahramani (2006) and proceeds by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. This turns out to be equivalent to the previous derivation.

We first introduce a finite beta-Bernoulli prior for generating an $N \times K$ matrix Z in which both N and K are finite. To generate Z from this prior, we sample

$$\begin{aligned} \pi_k &\sim \text{Beta}(\alpha/K, 1) \qquad k \in \{1,\ldots,K\} \\ z_{ik} &\sim \text{Bernoulli}(\pi_k) \qquad i \in \{1,\ldots,N\}, k \in \{1,\ldots,K\} \end{aligned}$$

where α is a parameter. As $K \to \infty$, the π here will correspond to the p in the BP, but we keep the notation as π to be consistent with Griffiths and Ghahramani (2006). Conditioned on π_k all entries of the k^{th} column are independent Bernoulli samples.

For finite K, we get that

$$\begin{aligned} p(Z|\pi) &= \prod_{k=1}^{K}\prod_{i=1}^{N} p(z_{ik}|\pi_k) \\ &= \prod_{k=1}^{K} \pi_k^{m_k}(1-\pi_k)^{N-m_k} \end{aligned}$$

where $m_k = \sum_{i=1}^{N} z_{ik}$ is the number of objects with feature k.

We note that the prior on π_k is conjugate to this likelihood, so we can integrate out each π_k to get

$$p(Z) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma\left(m_k + \frac{\alpha}{K}\right)\Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}.$$

As we take $K \to \infty$, this will define a prior on Z such that each row only has a finite number of non-zero entries. However, by construction, all of these matrices will be extremely sparse and as the number of columns goes to infinity, the probability of any particular configuration will go to zero. To fix this problem, Griffiths and Ghahramani (2006) defined the notion of equivalence classes of matrices, the *left-ordered form* we mentioned earlier. Since multiple matrices map to the same equivalence class, we will see that even though the probability of any particular matrix goes to zero, the probability of its equivalence class does not. The number of matrices in the equivalence class of Z can be shown to be

$$\begin{pmatrix} K \\ K_0 \ \cdots \ K_{2^{N-1}} \end{pmatrix} = \frac{K!}{\prod_{h=0}^{2^{N-1}} K_h!}$$

where K_h is the number of columns having the binary representation of the number h in Z. It is easy to see that each of these matrices is going to have the same likelihood under the prior, so if we define the class $[Z]$ to be all matrices having the same

left-ordered form as Z, then

$$\begin{aligned} p([Z]) &= \sum_{Z \in [Z]} p(Z) \\ &= \frac{K!}{\prod_{h=0}^{2^{N-1}} K_h!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma\left(m_k + \frac{\alpha}{K}\right)\Gamma\left(N - m_k + 1\right)}{\Gamma\left(N + 1 + \frac{\alpha}{K}\right)}. \end{aligned}$$

Taking the limit as $K \to \infty$, we get

$$p([Z]) = \frac{\alpha^{K^+}}{\prod_{h=1}^{2^{N-1}} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K^+} \frac{(N - m_k)!(m_k - 1)!}{N!}$$

where K^+ is the number of non-zero columns and $H_N = \sum_{j=1}^{N} \frac{1}{j}$ is the N^{th} harmonic number. This is exactly the prior distribution of the equivalence class $[Z]$ under the Indian Buffet Process. For details on taking the limit, see Griffiths and Ghahramani (2006).

We should also note that the distribution of the $\{\pi_k\}_{k=1}^{K}$ converges to the distribution of $\{p_k\}_{k=1}^{\infty}$ of B drawn from the beta process as $K \to \infty$. Therefore, to get a finite approximation to B, we can now either get draws from a truncated stick breaking process (truncated meaning we just stop after some K and assume all smaller p_k are zero) or we can draw the set $\{\pi_k\}_{k=1}^{K}$ independently from Beta$(\alpha/K, 1)$.

Derivation summary We have described two ways to derive the IBP, one using the Lévy process framework and the other using a limiting argument on a parametric class of priors. Both of these produce the same prior and demonstrate contrasting ways to generate them. Various extensions to these priors proceed by extending one or both of these derivations, so they are both important to understand.

2.3.4 Extensions

Since the original development, there have been several variations and extensions to the BP/IBP. We do not go into depth about any of these, merely highlighting some of the recent work done in this field.

One way to extend the BP/IBP is to develop a two-parameter version. In the original prior, there is only one parameter, α. Both Ghahramani et al. (2006) and Thibaux and Jordan (2007) looked at their formulations (the IBP and BP, respectively) and identified distinct ways to introduce another parameter beyond α in such a way that the original prior was a special case. This allowed for more flexibility in the sharing of features or concentrations of features.

A development after identifying the relationship between the IBP and the BP was to develop the hierarchical beta process (HBP) (Thibaux and Jordan, 2007). This work mirrors the work done on the hierarchical Dirichlet process (HDP) by Teh et al. (2005). In the HBP, we wish to draw latent features for two (or more) non-overlapping groups of objects. One way to attempt this would be to use a common base measure B_0 and draw $B^{(1)} \sim \text{BP}(c, B_0)$ for the first group and $B^{(2)} \sim \text{BP}(c, B_0)$ for the second group and use these to generate the respective features for each group. However, since B_0 is continuous, $B^{(1)}$ and $B^{(2)}$ would not share any atoms and hence would not have any features in common. In order for $B^{(1)}$ and $B^{(2)}$ to share atoms, B_0 must be discrete, but in order to have a nonparametric prior allowing for an unbounded number of features, B_0 must have countably infinite support. The solution is to add an extra layer in the hierarchy and have B_0 itself be a random draw from a beta process. This allows for there to be an infinite number of unknown features that are shared across groups.

The HBP is an example of a prior that induces dependencies across different instantiations of beta processes. The dependent IBP of Williamson et al. (2010a) seeks to also induce dependencies across latent features induced by known covariates related to our observations.

Other priors have addressed known limitations or assumptions of the BP/IBP. One such limitation is that the weights $\{p_k\}_{k=1}^{\infty}$ drawn from the beta process in order of decreasing size shrink exponentially quickly in expectation. For some applications, this is not a realistic assumption and we might hope for a slower decrease to capture phenomena like power laws. To address this issue, Teh and Görür (2009) created the IBP with power-law behavior. They demonstrate how to do this by modifying the Lévy measure and derive the corresponding marginalized representation. This work was partially inspired by properties of the Pitman-Yor process, an extension of the Dirichlet Process (Pitman and Yor, 1997). An alternate power-law approach based on a stick-breaking construction was recently introduced by Broderick et al. (2011).

Another extension addresses the lack of correlations between the features themselves since in the BP, the ω_k are sampled independently. Doshi-Velez and Ghahramani (2009a) developed a prior that introduced correlations in the features.

Austerweil and Griffiths (2010) develop the transformed IBP, which allows ω_k to be slightly transformed across different instantiations. This work was inspired by the transformed Dirichlet process by Sudderth et al. (2005).

A large focus of later chapters in this dissertation is on developing non-exchangeable priors, that is, priors on Z such that the rows are not exchangeable. An early example of one of these priors is the Markov IBP by Van Gael et al. (2009), which allows the rows of Z to be related in a Markov chain. In Chapter 4, we introduce our own Markov non-exchangeable variation of the IBP that improves upon the prior by Van Gael et al. (2009) along with several other non-exchangeable variations.

2.4 Priors for Integer Valued Latent Feature Models

The second class of priors for nonparametric latent feature models we wish to discuss are for non-negative integer valued latent feature models. This class of priors is even more recent than the binary latent feature priors, with the first prior specifically for non-negative latent features developed by Titsias (2008). However, as with most good ideas, it built upon closely related work. Wolpert and Ickstadt (1998a) derived many of the necessary results for using these priors and these results themselves use the gamma process, which is a Lévy process older than the beta process that was even used in original definition of the Dirichlet process (Ferguson, 1973).

Much of the description here mirrors the work done for binary latent feature priors. Over the course of this section, we will introduce multiple ways to generate a prior $p(Z)$ on non-negative integer valued matrices with an unbounded number of non-zero columns. Just as for the BP/IBP, we will first introduce a completely random measure and discuss how to use it to generate non-negative integer valued Z in Section 2.4.1. We then discuss stick breaking in Section 2.4.2 before describing how to integrate out the completely random measure in Section 2.4.3. Section 2.4.4 finishes with some extensions that can be done with these priors.

2.4.1 The Gamma Process

Definition 2.4.1. The *gamma process* (GP) is a completely random measure with Lévy measure

$$\nu(d\omega, dp) \quad = \quad c\frac{e^{-cp}}{p}dpB_0(d\omega).$$

We write $B \sim \mathrm{GP}(c, B_0)$ to denote that B is drawn from a GP with base measure B_0 and concentration parameter c.

Just like the BP had an improper beta Lévy measure which lead B to have a countably infinite number of points, the GP has an improper gamma measure, which means that B here will also have a countably infinite number of points, all but a finite number of which will have mass arbitrarily close to zero. We know that B will be a finite measure again by Campbell's theorem from Section 2.2.3. Whereas the weights p in the BP were in $[0, 1]$, the weights p in the GP are in $[0, \infty)$. Unlike for the beta process, the measure B has a simple characterization due to the aggregation property of the gamma distribution

$$B(S) \quad \sim \quad \mathrm{Gamma}(cB_0(S), c) \qquad \forall S \subset \Omega. \tag{2.14}$$

For a proof of this result, see Section 3 of Dufresne et al. (1991).

Just as the beta distribution is the conjugate prior for the binary-valued Bernoulli distribution, the gamma distribution is the conjugate prior for the non-negative integer valued-Poisson distribution. Hence, following that analogy, to get our nonparametric prior on Z, we will now sample $z_i \sim \mathrm{PP}(B)$ where PP stands for Poisson process. z_i is now a vector with the counts of the number of times each atom in B is present in the i^{th} entity. Thus, our full stochastic process for generating the latent feature matrix Z is

$$\begin{aligned} B &\sim \mathrm{GP}(c, B_0) \\ z_i|B &\sim \mathrm{PP}(B) \qquad i = 1, \ldots, N. \end{aligned}$$

As we did for the BP, if we only observe $z_1, \ldots, z_N$, we wish to update our posterior belief of B. In the notation of Thibaux (2008), Wolpert and Ickstadt (1998a) showed that this is just

$$\begin{aligned} B|z_1, \ldots, z_N &\sim \mathrm{GP}(c + N, B_N) \\ \text{where } B_N &= \frac{c}{c+N} B_0 + \frac{1}{c+N} \sum_{j=1}^{N} z_j. \end{aligned} \tag{2.15}$$

In other words, just like for the BP, the posterior the weight of the continuous measure B_0 is reduced and we add in a discrete measure located at the atoms of the z_i weighted by how many times they occur.

An equivalent, but slightly clearer way of thinking about this that makes explicit what is going on is that we start off with a gamma process B with Lévy measure

$$\nu(d\omega, dp) = cp^{-1}e^{-cp}dpB_0(d\omega).$$

We then draw z_i from the resulting discrete measure. For each atom in B, we draw ω_j a Poisson(p_j) number of times. If we have drawn a point ω_j a total of $m_{N,j} > 0$ times after N draws, then the the posterior distribution of p_j at ω_j is the proper Gamma($m_{N,j}, N + c$) distribution

$$\frac{(c+N)^{m_{N,j}}}{\Gamma(m_{N,j})} p_j^{m_{N,j}-1} e^{-(c+N)p_j},$$

just as if we had started off with a Gamma$(0, c)$ prior and then observed N observations $z_1, \ldots, z_N$ from a Poisson distribution. Since we have observed at least one z_i at location ω_j, we know that B must have had an atom at ω_j, resulting in the discrete posterior.

For the continuous part of the posterior Lévy measure, the probability of not

seeing any z_i at ω_j is e^{-Np_j}, so the continuous part of the Lévy measure is

$$cp^{-1}e^{-(c+N)p}dpB_0(d\omega).$$

This is similar to starting with a Gamma$(0, c)$ prior and then not observing anything after N attempts. The resulting posterior is still improper, thus resulting in a posterior gamma process. Combining these two parts gives us the posterior Lévy measure.

2.4.2 The Stick Breaking Process

Again, following the development of the BP, we wish to sample $B \sim \text{GP}(c, B_0)$ using a stick breaking construction. However, most likely due to the fact that the use of the GP is relatively young in the nonparametric latent feature community, no stick breaking construction with closed form has been presented for this prior. We can always use the inverse Lévy measure algorithm to sample B, but we might hope for a closed form solution for special cases, just like for the BP.

While it is unlikely that a stick breaking construction for the GP that generates sticks in decreasing size exists, we can leverage the stick breaking construction of the Dirichlet Process (DP) developed by Sethuraman (1994) to get a size-biased sample of the sticks of the GP. In such a sample, the sticks will be drawn in decreasing order of expectation (though the sticks themselves are most likely not of strictly decreasing size).

Let $\alpha = B_0(\Omega)$. In the stick breaking construction for the DP, we sample

$$\begin{aligned} v_k &\sim \text{Beta}(1, \alpha) \\ p_k &\sim v_k \prod_{i=1}^{k-1}(1 - v_k). \end{aligned}$$

It is well known that the DP is a normalized gamma process, i.e., if $B \sim \text{GP}(1, B_0)$, then $B/B(\Omega)$ is a DP. This fact was originally mentioned by Ferguson (1973) and has been used successfully in papers extending the DP such as those by Rao and Teh (2009) and Lin et al. (2010).

Since the DP is a normalized GP, we expect it to have unit mass, which it does since $\sum_{k=1}^{\infty} p_k = 1$. This means that given the sticks from a DP, all we need to do is scale all this sticks up to be a draw from a GP. Using Equation 2.14, we know that $B(\Omega) \sim \text{Gamma}(\alpha, 1)$ and that by properties of the gamma distribution, $B(\Omega)$ is independent of the normalized measure $B/B(\Omega)$. Therefore, to get a stick breaking construction for GP, we use the stick breaking construction for the DP and scale the sticks up by a random draw from Gamma$(\alpha, 1)$.

2.4.3 The Infinite Gamma Poisson Feature Model

Similar to the BP, we have seen how to generate $Z = (z_1, \ldots, z_N)$ via

$$p(z_1, \ldots, z_N) \;=\; \int \left[\prod_{i=1}^{N} p(z_i|B) \right] dP(B),$$

where B is drawn from the GP and z_i are drawn conditionally independently via a PP. Therefore, the rows $(z_1, \ldots, z_N)$ are also exchangeable as is seen by checking De Finetti's theorem in Section 2.1.4.

The Infinite Gamma Poisson Feature Model (IGPFM) is a way to directly sample Z without first needing to sample B, just like the IBP is for the BP, and was developed by Titsias (2008). It was again derived without knowledge of the GP, but the GP can be shown to be the De Finetti mixing distribution for the IGPFM (Thibaux, 2008).

We first introduce an earlier culinary metaphor for sampling partitions called the Chinese restaurant process that will be used by the IGPFM. Next we introduce the IGPFM itself, then show how to derive it in two different ways, the first from the GP and the second as Titsias (2008) did without knowledge of the GP.

The Chinese Restaurant Process In the Bayesian nonparametric world, there are several examples of processes and their De Finetti mixing distribution. We have already discussed the IBP and the BP and this section discusses the IGPFM and the GP. The first widely used pair were the Chinese restaurant process (CRP) and the Dirichlet process (DP). As we have already mentioned, the DP is a normalized GP, so before diving into the marginal representation of the IGPFM, it is useful to review the CRP.

The CRP is a method of sampling clusterings of points when the underlying clustering is derived from the DP. The CRP's name goes back to the 1980s (Aldous, 1983) and is related to the earlier work by Blackwell and MacQueen (1973) as well as the Ewens sampling formula (Ewens, 1972). The CRP is a culinary metaphor for how m people sit at tables in a Chinese restaurant. All people who sit at the same table are assigned to the same partition and in this way, the distribution over seating configurations determines a distribution over partitions. There is a single parameter α for this process.

The m people are assigned to tables in an incremental fashion as follows:

- The first customer starts a new table.
- The i^{th} customer sits at each of the occupied tables with probability proportional to the number of people sitting there and sits at a new, unoccupied table with probability proportional to α. To be more precise, if the k^{th} occupied table has

$m_{i-1,k}$ of the people $1, \ldots, i-1$ sitting at it, then the i^{th} person sits at it with probability $\frac{m_{i-1,k}}{i-1+\alpha}$ and sits at a new table with probability $\frac{\alpha}{i-1+\alpha}$.

If we only care about the sizes of the partitions resulting from this process, and we let c_j be the number of partitions with j customers, then the probability assigned by the CRP to a partition with $(c_1, c_2, \ldots, c_m)$ (since necessarily $c_i = 0$ for all $i > m$) is

$$p(c_1, \ldots, c_m) = \frac{m!}{\prod_{j=1}^{m} c_j! j^{c_j}} \alpha^K \frac{\Gamma(\alpha)}{\Gamma(\alpha + m)}$$

where $K = \sum_{j=1}^{m} c_j$, the number of occupied partitions. This process will be used by the IGPFM.

The Infinite Gamma Poisson Feature Model The Infinite Gamma Poisson Feature Model (IGPFM) samples Z when $c = 1$ and $\alpha = B_0(\Omega)$. Unlike the IBP, there is no culinary metaphor, just a generative process. To generate a matrix Z from the IGPFM(α), start with an all-zero matrix and perform the following:

- In the first row, we first decide the total count of features we will add and then we decide how to split this count up into individual features. We do this by sampling g_1, a Negative Binomial NB($\alpha, 1/2$) number of features and then partition g_1 according to the CRP(α). The partitions become the new features and the counts in the partitions are the counts entered in the matrix.

  For example, we might first draw $g_1 = 7$ from the NB($\alpha, 1/2$), which means that the first row will have a total of seven (not necessarily unique) features. We then run the CRP to see how this count will be partitioned into unique features. The CRP might split these seven features up across three unique features having counts $(4, 1, 2)$, so the first row would be $(4, 1, 2, 0, 0 \ldots)$.

- Now assuming we have filled in the first $i-1$ rows, we fill in the i^{th} row as follows:

  - First look at all features that are present in $z_1, \ldots, z_{i-1}$. Let $m_{i-1,k}$ be the number of times the k^{th} feature has occurred in $z_1, \ldots, z_{i-1}$, that is, $m_{i-1,k} = \sum_{j=1}^{i-1} z_{jk}$. Draw $z_{ik} \sim \text{NB}\left(m_{i-1,k}, \frac{i}{i+1}\right)$. Thus, the features that are more prevalent in $1, \ldots, i-1$ have a higher chance of being selected than those features which are not.
  - Now select the total count of new features g_i that will be unique to the i^{th} row from NB$\left(\alpha, \frac{i}{i+1}\right)$, and distribute this into unique features according to the CRP(α).

As in the IBP, this process in itself is not exchangeable, but by equivalence classes based on left-ordered forms for non-negative integer valued matrices, we get the same kind of exchangeability result.

Derivation 1 The first derivation of the IGPFM is due to Thibaux (2008), showing that it results from marginalizing out the gamma-Poisson process construction of $p(Z)$. We let $c = 1$ and $\alpha = B_0(\Omega)$.

For the first row of Z, z_1, we sample z_1 from

$$p(z_1) \sim \int p(z_1|B)dP(B)$$

where $p(z_1|B)$ is a Poisson process and the prior on B is $\mathrm{GP}(c, B_0)$.

By Equation (2.14), the measure of $B(\Omega) \sim \mathrm{Gamma}(B_0(\Omega), 1)$. Therefore, recalling that if $x|\lambda \sim \mathrm{Poisson}(\lambda)$ and $\lambda \sim \mathrm{Gamma}(a, b)$, then by marginalizing out λ, we get $x \sim \mathrm{NB}\left(a, \frac{b}{1+b}\right)$, then we can see that the total count of points in z_1 is $\mathrm{NB}(\alpha, 1/2)$. Let this be g_1 as in the IGPFM.

Now we need to know how to allocate the g_1 points in z_1. Given g_1 and B for any B, we note that z_1 is a set of g_1 independent draws from $B/B(\Omega)$, which is itself a DP. Marginalizing out B therefore gives us that z_1 is a draw of g_1 points from the $\mathrm{CRP}(\alpha)$.

For the i^{th} row of Z, z_i, we sample z_i from the conditional distribution

$$p(z_i|z_1, \ldots, z_{i-1}) \sim \int p(z_i|B)dP(B|z_1, \ldots, z_{i-1})$$

where $p(z_i|B)$ is a Poisson process and the posterior of B given $z_1, \ldots, z_{i-1}$ is as given by Equation (2.15):

$$\begin{aligned} B|z_1, \ldots, z_{i-1} &\sim \mathrm{GP}(c + i - 1, B_{i-1}) \\ \text{where } B_{i-1} &= \frac{c}{c+i-1}B_0 + \frac{1}{c+i-1}\sum_{j=1}^{i-1} z_j. \end{aligned}$$

B_{i-1} is a mixed continuous and discrete measure, so since B and z_i are completely random measures, we treat the continuous and discrete parts of B_{i-1} separately.

For the discrete part, we treat each atom in $\sum_{j=1}^{i-1} z_j$ independently (which is valid since this is a completely random measure). Using Equation (2.14) on each atom, we see that for the k^{th} atom ω_k, if it has appeared $m_{i-1,k}$ times in $z_1, \ldots, z_{i-1}$, then $B(\{\omega_k\})|z_1, \ldots, z_{i-1} \sim \mathrm{Gamma}(m_{i-1,k}, c + i - 1)$. z_{ik} is drawn from a Poisson distribution conditioned upon $B(\{\omega_k\})$, which means that when we marginalize out

$B(\{\omega_k\})$, we get $z_{ik}|z_1,\ldots,z_{i-1} \sim \text{NB}\left(m_{i-1,k}, \frac{i}{i+1}\right)$.

For the continuous part, the updated c is $c+i-1$ and we have shrunk the base measure B_0 by $\frac{c}{c+i-1}$. Equation (2.14) therefore tells us that on the continuous part of the posterior, $B(\Omega)|z_1,\ldots,z_{i-1} \sim \text{Gamma}(cB_0(\Omega), i+c-1)$, which means that for $c=1$ and $\alpha = B_0(\Omega)$, the total count of new features in z_i is $\text{NB}\left(\alpha, \frac{i}{i+1}\right)$. Call this count g_i. By the same argument as for z_1, these counts are partitioned according to the $\text{CRP}(\alpha)$.

Therefore the steps for sampling each z_i as derived by marginalizing out B are equivalent to those of the IGPFM and it becomes clear how the mix of a continuous and discrete posterior distribution of B directly affect the two aspects of sampling z_i in the IGPFM.

Derivation 2 The second derivation is the original derivation by Titsias (2008) and proceeds by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. This turns out to be equivalent to the previous derivation.

We first introduce a finite gamma-Poisson prior for generating an $N \times K$ matrix Z in which both N and K are finite. To generate Z from this prior, we sample

$$\begin{aligned} \lambda_k &\sim \text{Gamma}\left(\frac{\alpha}{K}, 1\right) \qquad k \in \{1,\ldots,K\} \\ z_{ik} &\sim \text{Poisson}(\lambda_k) \qquad i \in \{1,\ldots,N\}, k \in \{1,\ldots,K\} \end{aligned}$$

where α is a parameter. In the infinite limit, λ become equivalent to the p of the GP. Conditioned on λ_k, all entries of the k^{th} column are independent Poisson samples.

This gives us that

$$\begin{aligned} p(Z|\lambda) &= \prod_{n=1}^{N}\prod_{k=1}^{K} \frac{\lambda_k^{z_{nk}} \exp(-\lambda_k)}{z_{nk}!} \\ &= \prod_{k=1}^{K} \frac{\lambda_k^{m_k} \exp(-N\lambda_k)}{\prod_{n=1}^{N} z_{nk}!} \end{aligned}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$. Integrating out λ gives

$$p(Z|\alpha) = \prod_{k=1}^{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right)}{\Gamma\left(\frac{\alpha}{K}\right)(N+1)^{m_k+\frac{\alpha}{K}} \prod_{n=1}^{N} z_{nk}!}.$$

This is exchangeable and the columns are independent.

By associating each matrix Z with its equivalence class $[Z]$ as is done in the IBP, we see that each equivalence class consists of $\frac{K!}{\prod_{h=0}^{\infty} K_h!}$ matrices where K_h is the

number of columns that have a particular value[3]. Therefore

$$p([Z]|\alpha) = \frac{K!}{\prod_{h=0}^{\infty} K_h!} \prod_{k=1}^{K} \frac{\Gamma\left(m_k + \frac{\alpha}{K}\right)}{\Gamma\left(\frac{\alpha}{K}\right)(N+1)^{m_k+\frac{\alpha}{K}} \prod_{n=1}^{N} z_{nk}!}. \tag{2.16}$$

By using the exact same limits as in Griffiths and Ghahramani (2005) when $K \to \infty$ in Equation (2.16), we get that if K^+ is the number of non-empty columns, then

$$p([Z]|\alpha) = \frac{1}{\prod_{h=1}^{\infty} K_h!} \frac{\alpha^{K^+}}{(N+1)^{m+\alpha}} \frac{\prod_{k=1}^{K^+}(m_k-1)!}{\prod_{k=1}^{K^+}\prod_{n=1}^{N} z_{nk}!} \tag{2.17}$$

where $m = \sum_{k=1}^{K^+} m_k$.

Titsias (2008) shows that this is the same distribution as the IGPFM (up to equivalence classes). We should also note that the distribution of the $\{\lambda_k\}_{k=1}^{K}$ converges to the distribution of $\{p_k\}_{k=1}^{\infty}$ of B drawn from the gamma process as $K \to \infty$. Therefore, to get a finite approximation to B, we can now either get draws from a truncated stick breaking process or we can draw the set $\{\lambda_k\}_{k=1}^{K}$ independently from Gamma $\left(\frac{\alpha}{K}, 1\right)$.

Derivation summary We have derived the IGPFM in two ways, one based on the gamma process and one based on defining a parametric prior on $N \times K$ matrices and letting $K \to \infty$. Both of these define equivalent priors and can be used to generalize the IGPFM.

2.4.4 Extensions

The GP has been much less used than the BP in Bayesian nonparametric latent feature models, so there are few extensions. The main extension is the hierarchical GP by Thibaux (2008) which is the logical hierarchical extension to the GP applicable to simultaneous inference of feature counts across groups of data. There exist other variations of the GP, but most have been used for applications besides latent feature models. For example, taking advantage of the relationship between the GP and the DP, both Rao and Teh (2009) and Lin et al. (2010) have worked with variations on the GP for more flexible DP priors.

[3]We use some unique mapping from columns to integers such that the all zero column corresponds to K_0. In Titsias (2008), he assumes that there exists a c such that $z_{nk} < c$, but this isn't true. In our notation, an infinite number have $K_h = 0$, which is fine. Note also that in Titsias (2008), he accidentally uses a $\sum$ instead of $\prod$.

2.5 Summary

In this chapter, we have introduced the ideas behind Bayesian nonparametric latent feature models and discussed two concrete latent feature priors, the BP/IBP and the GP/IGPFM. Both of these classes of priors are exchangeable and can be derived either by constructing a parametric prior for finite $N \times K$ matrices and letting $K \to \infty$ or by identifying and working with their De Finetti mixing distributions, which themselves are Lévy processes/completely random measures. Both of these derivation techniques are extremely helpful and allow us to construct generalizations of these priors.

While we have motivated the use of these models and priors, we have not told how to use them in concrete applications, nor how to perform inference in them. In the next chapter, Chapter 3, we discuss how to inference algorithms for these models. The next two chapters, Chapters 4 and 5 discuss extensions to the basic priors as well as inference algorithms for these extensions. Finally, Chapter 6 ties this all together by discussing concrete applications.

Chapter 3

Bayesian Nonparametric Latent Feature Model Inference Algorithms

Now that we have established two different priors useful in Bayesian nonparametric latent feature models, we must be able to apply them to various applications and perform posterior inference.

Given an observation X, we wish to follow the outline from Section 2.1.3. First, we must determine if our data X is suitable for a latent feature model and if so, if one of our priors $p(Z)$ seems suitable. Assuming we are in a context where using one of these models makes sense, we must then define the likelihood $p(X|Z,\theta)$ and the prior $p(\theta)$. Inference algorithms, which we discuss in this chapter, involves computing our posterior beliefs about Z and θ. In particular, we must use Bayes's rule to compute

$$p(Z,\theta|X) \quad \propto \quad p(X|Z,\theta)p(\theta)p(Z).$$

Unfortunately, performing this inference exactly is intractable since the distribution $p(Z,\theta|X)$ has no simple closed form. Even if we ignored θ, we would need to compute the right hand side for every possible Z and then normalize to get the posterior distribution on the left hand side. Given that the domain of Z is countably infinite, this is not possible.

We must therefore resort to some kind of approximate inference algorithm. There are two main classes of approximate inference algorithms: sample-based techniques, of which Markov Chain Monte Carlo (MCMC) is the most popular for Bayesian nonparametric models, and variational techniques, for which naïve mean field is the most prevalent. For appropriate background on MCMC, see Robert and Casella (2004) and for appropriate background on variational methods, see Wainwright and Jordan (2008).

While we reviewed several different representations of the priors for Bayesian

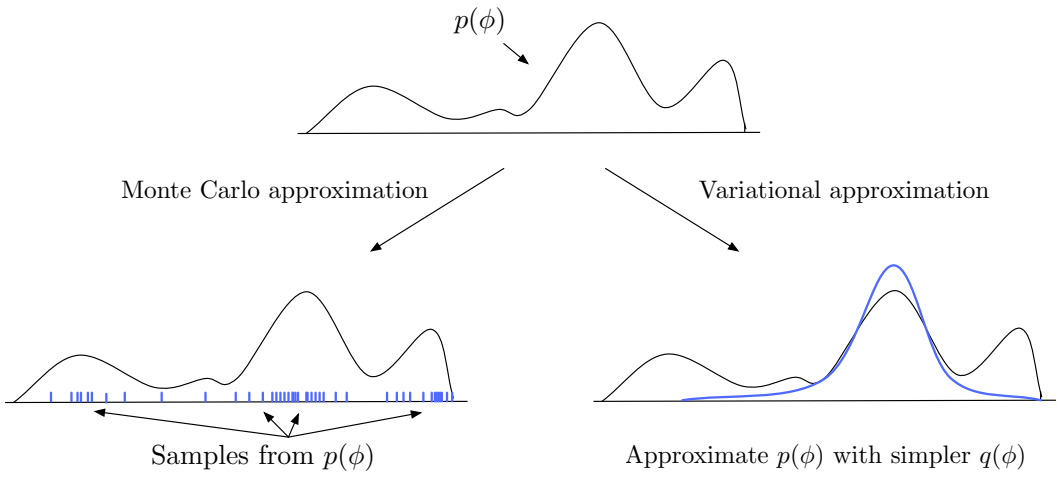


Figure 3.1: A comparison of the two main classes of approximate inference algorithms. Sample-based approaches such as Monte Carlo methods approximate the true distribution $p(\phi)$ by getting samples from it and using them to approximate the whole distribution. Variational approaches attempt to find a simpler distribution $q(\phi)$ that approximates $p(\phi)$ well.

nonparametric latent feature models in Chapter 2, the majority of MCMC techniques have been developed for their marginalized representations, the IBP and IGPFM, though there are a few that work with the Lévy process stick breaking constructions. Variational inference algorithms have so far focused on the Lévy process stick breaking constructions, but are potentially applicable to the marginalized representations.

3.1 Overview

The main difference between the sample-based MCMC approximations and variational approximations lies in how they both approximate a distribution. Suppose we wish to approximate some distribution $p(\phi)$ such as the distribution in Figure 3.1. If this distribution does not have a nice form that we can easily work with, then we can use sample-based and variational approaches to approximate it.

In sample-based approaches, we wish to get samples from $p(\phi)$ and approximate it with those samples. If we can either directly sample from $p(\phi)$ or sample from a very close distribution and use standard rejection sampling, that can work very well. However, it is often very hard to do this, especially if ϕ has multiple components. MCMC gives us a way to construct a Markov chain such that after the chain has burned-in (run long enough), the samples we get from it are from $p(\phi)$. The longer

we run the chain and the more samples we get, the better an approximation. This is a kind of Monte Carlo approximation shown in the left portion of Figure 3.1. We discuss how to use MCMC to approximate posterior distributions of the beta process and gamma process in Section 3.2.

In variational approximations, instead of approximating $p(\phi)$ with samples, we find a simpler distribution $q(\phi)$ from a parameterized family of distributions Q that is as close as possible to $p(\phi)$. In mean field approximations, one particular kind of variational approximation, we optimize the parameters of q so that it is as close as possible to p as measured by the KullbackLeibler divergence (KL divergence) $D(q||p)$. This is shown in the right portion of Figure 3.1. In that figure, Q is the family of normal distributions, so the closest q is still not a great fit for p, but the optimal q is easy to compute. Whereas in MCMC in which we can run the chain longer to get a better approximation of $p(\phi)$, once the optimization in the variational inference algorithm is done for a particular family Q, we have as close an approximation as we are going to get while staying in the family Q. However, we can always make Q a more complex family to get a better approximate of p, but the bigger Q gets, the harder the optimization. If Q is the family of all distributions, the minimizer of $D(q||p)$ is p itself, which we assume is hard to work with (else we would not be trying to approximate it). We discuss variational inference algorithms for our nonparametric priors in Section 3.3.

Therefore, the two kind of approximations have two different ways of finding closer approximations. With MCMC, the longer we run the chain, the better an approximation we get. With variational approximations, the bigger the family Q we allow, the better an approximation. In both cases, better approximations result in increased running time. However, there are instances when one works better than the other and it is therefore important to understand both and know when each one is the better algorithm to use.

Both MCMC and variational approximations have been developed for the BP/IBP, but to date, only MCMC approaches have been developed for the GP/IGPFM. We do not go into inference algorithms for any of the extensions of the nonparametric algorithms mentioned in Chapter 2.

After reviewing the MCMC approaches for both the BP and GP in Section 3.2, we discuss our variational approximation for the BP in Section 3.3. Then in Section 3.4, we compare the performance of MCMC and the variational approximation for the BP using the linear-Gaussian likelihood function discussed in Sections 3.2.1 and 3.3.

3.2 Markov Chain Monte Carlo

In this section, we introduce MCMC approximate inference algorithms for Bayesian nonparametric latent feature models that use both the beta process and the gamma process.

In general, the idea of MCMC is to start off with an arbitrary guess of what Z and θ are and construct a Markov Chain that eventually generates samples from the true posterior $p(Z, \theta|X)$. Assume that $\theta = (\theta_1, \ldots, \theta_d)$ is a d-dimensional vector and that Z stores only the finitely many non-zero columns of the latent feature matrix (thereby implicitly acknowledging the infinitely many all-zero features). Often, there is a parameter θ_k for each non-zero feature along with several other parameters independent of the features. We will construct a Markov chain to generate samples $(Z^{(1)}, \theta^{(1)}), (Z^{(2)}, \theta^{(2)}), \ldots$, where $(Z^{(m)}, \theta^{(m)})$ represents the m^{th} sample from the chain. To obtain these samples, we construct a transition kernel to go from the $(m-1)^{\text{th}}$ sample to the m^{th} sample. We will use a Gibbs sampler below, a particularly simple form of MCMC sampler, although these can also be augmented with Metropolis-Hastings proposals.

To sample $(Z^{(m)}, \theta^{(m)})$ from $(Z^{(m-1)}, \theta^{(m-1)})$, we iteratively update each single component of Z and θ. Assume below that whatever components of Z and θ that are being condition on are the most recent samples (of which some may be from sample $m-1$ and some from m). Also, let Z_{-ik} be all of Z except the (i,k) entry and θ_{-j} be all of θ except the j^{th} entry. A single iteration updates each element in turn until we have updated all components of Z and θ:

1. For $i = 1, \ldots, N$, if there are K^+ non-zero features in Z, then

   (a) For $k = 1, \ldots, K^+$, sample:

   $$\begin{aligned} z_{ik} &\sim p(z_{ik}|Z_{-ik}, \theta, X) \\ &\propto p(X|Z_{-ik}, z_{ik}, \theta)p(z_{ik}|Z_{-ik}). \end{aligned}$$

   (b) We must also consider the possibility that the i^{th} row has features unique to it that are not already present in Z. This is how the sampler can explore the whole space of potential latent feature matrices. We therefore sample the number of new features (and their configuration for the IGPFM) unique to the i^{th} row.

2. For $j = 1, \ldots, d$, sample

   $$\begin{aligned} \theta_j &\sim p(\theta_j|Z, \theta_{-j}, X) \\ &\propto p(X|Z, \theta_{-j}, \theta_j)p(\theta_j|\theta_{-j}). \end{aligned}$$

This algorithm is still correct if we vary the order of the updates from those above as long as we eventually update all components. The details for each of these steps are provided for each of the models in the following sections. As we will see, all the hard details were worked out in descriptions of the prior because the rows of Z are exchangeable, so these updates will be straightforward. We then ignore the first many samples until after the burn-in period of the Markov chain, at which point, $(Z^{(m)}, \theta^{(m)})$ will be samples from the desired posterior distribution. For details on why this works as well as convergence diagnostics, see Robert and Casella (2004) or Gelman et al. (2003).

3.2.1 MCMC for the Beta Process

There are several sample-based approximation algorithms for the BP/IBP. These include Gibbs samplers (Griffiths and Ghahramani, 2006), particle filters (Wood and Griffiths, 2006), slice sampling (Teh et al., 2007), and improved Gibbs samplers (Doshi-Velez and Ghahramani, 2009b; Doshi-Velez et al., 2009a). Of these samplers, all except the slice sampler work with the marginalized IBP representation of the beta process. We will only discuss the basic Gibbs sampler here.

The Gibbs sampler works with the IBP representation. We first present the Gibbs sampler assuming a generic likelihood and then present a concrete version using a linear-Gaussian likelihood. The purpose of the concrete model will be twofold. First, it will show that the steps needed for Gibbs samplers are not that challenging to compute and second, we will use this exact model to compare MCMC against the variational approximations in Section 3.4.

Generic Likelihood Earlier in Section 3.2, we presented an overview of all the steps we need to derive the Gibbs sampler. Step (2) of sampling θ is application specific and must be addressed individually for each application, so to fill in a generic inference algorithm, we only need to address the two parts of Step (1) of the Gibbs sampler. These two steps are:

1. For non-zero columns, we must first be able to sample

$$z_{ik} \quad \propto \quad p(X|Z_{-ik}, z_{ik}, \theta)p(z_{ik}|Z_{-ik}).$$

2. For each row we must also be able to sample the number of new features unique to the i^{th} row.

For the first step, we only need to consider $z_{ik} \in \{0, 1\}$, so we evaluate the right-hand side for $z_{ik} = 0$ and $z_{ik} = 1$, normalize, and sample z_{ik} from the resulting Bernoulli distribution. The term $p(X|Z_{-ik}, z_{ik}, \theta)$ is the likelihood assuming we knew

all of Z and θ. The term $p(z_{ik}|Z_{-ik})$ is the prior probability of $z_{ik} = 0$ or $z_{ik} = 1$ given the rest of the matrix. To evaluate this term, we use exchangeability of the rows of Z to assume the i^{th} row is the last row to be added to Z after the $N-1$ other rows have been added. Then the prior probability of z_{ik} being non-zero is the same as the probability that the last customer to enter the Indian buffet samples the k^{th} dish as discussed in Section 2.3.3. This is just

$$\begin{aligned} p(z_{ik} = 0|Z_{-ik}) &= \frac{N - m_{-ik}}{N} \\ p(z_{ik} = 1|Z_{-ik}) &= \frac{m_{-ik}}{N}, \end{aligned}$$

where m_{-ik} is the number of times feature k is present in Z excluding the i^{th} row.

For the second step, we must sample the number of features that are unique to the i^{th} row. This step can be tricky if the likelihood $p(X|Z,\theta)$ is not conjugate to the prior $p(\theta)$ for any feature specific parameters. For a review of conjugacy, see (Robert, 2007). This step is tricky because we only store the feature-specific θ_k for each non-zero column of Z. Therefore, when we sample new columns of Z, we have not yet sampled any corresponding θ_k and we must integrate these unknowns out. If we cannot, we can adapt techniques developed for the DP by Neal (1998) to the IBP in order to sample with non-conjugate priors. Here we will assume that θ does have a conjugate prior so that we do not need to worry about this.

To sample the number of features unique to row i, we sample K_i^{new}, the number of new features and since they are binary features, add the corresponding number of ones to the i^{th} row of Z. Fortunately, by exchangeability again, we know from Section 2.3.3 that $K_i^{\text{new}} \sim \text{Poisson}(\alpha/N)$. So we can therefore sample

$$p(K_i^{\text{new}}|X, Z, \theta, \alpha) \propto p(X|Z, K_i^{\text{new}}, \theta)p(K_i^{\text{new}}|\alpha)$$

where $p(X|Z, K_i^{\text{new}}, \theta)$ is the probability of our observations if we have the old part of Z augmented with K_i^{new} additional features in the i^{th} row and θ contains all the parameters associated with the old features and any likelihood specific features, but does not contain any parameters associated with the new columns. This is the term that is problematic in non-conjugate models. $p(K_i^{\text{new}}|\alpha)$ is $\text{Poisson}(\alpha/N)$.

We must evaluate the right-hand side for $K_i^{\text{new}} \in \mathbb{Z}^*$ where $\mathbb{Z}^* = \{0\} \cup \mathbb{Z}^+$. We cannot do this exactly, so we often sample K_i^{new} from a large set $\{0, 1, \ldots, c\}$ for some large c since the prior probability of $K_i^{\text{new}} > c$ gets arbitrarily close to zero as c increases. This is an approximation that can be removed by the use of a slice sampler as described by Teh et al. (2007).

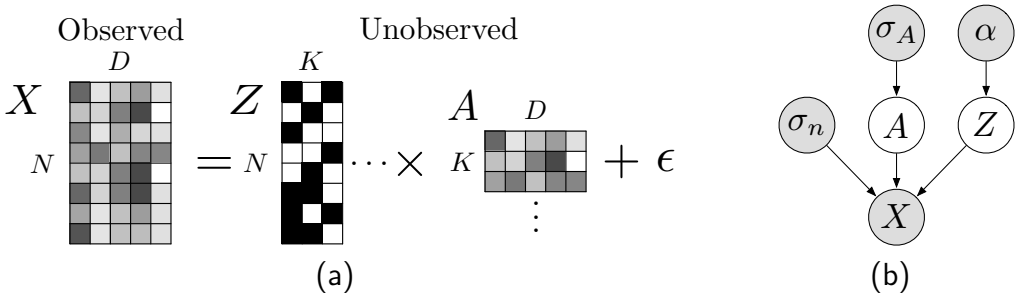


Figure 3.2: The Linear Gaussian model. (a) A visual display of what the data and unknowns of the linear-Gaussian might look like. (b) The corresponding graphical model.

Linear-Gaussian Likelihood We now demonstrate how this full process works using a specific likelihood. This likelihood is the linear-Gaussian model presented by Griffiths and Ghahramani (2006) as a means of discovering features shared across images. By examine a specific likelihood model, we will be able to make a concrete empirical comparison between the Gibbs sampler and variational inference techniques on real data with a real likelihood. We save this comparison for Section 3.4.

In the linear-Gaussian model, we assume that X is an $N \times D$ real-valued matrix where there are N-row observations each having D dimensions. We believe that there is some underlying latent feature matrix Z and that for each feature k, there is a normally distributed D-dimensional vector A_k that is the associated with it. We let A be the matrix where $\{A_k\}$ are arranged as the columns. A is the only parameter and is therefore the generic θ we discussed before. This can be seen pictorially in Figure 3.2(a). Our generative model is the graphical model seen in Figure 3.2(b) and takes the form

$$\begin{aligned} p(Z|\alpha) &= \text{IBP}(\alpha) \\ p(A_k|\sigma_A^2) &= \mathcal{N}(0, \sigma_A^2 I) \\ p(X_i|Z, A, \sigma_n^2) &= \mathcal{N}(z_i A, \sigma_n^2 I). \end{aligned}$$

We assume σ_A, σ_n, and α are all known, but in practice, we must also infer them.

In this simple model, the prior on A is conjugate to the likelihood, so we can

integrate A out of the model entirely. Therefore, we are left a likelihood of

$$p(X|Z,\sigma_n^2,\sigma_A^2) = \frac{1}{(2\pi)^{ND/2}\sigma_n^{(N-K)D}\sigma_A^{KD}\left|Z^\top Z+\frac{\sigma_n^2}{\sigma_A^2}I\right|^{D/2}} \exp\left\{-\frac{1}{2\sigma_n^2}\mathrm{tr}\left(X^\top\left(I-Z\left(Z^\top Z+\frac{\sigma_n^2}{\sigma_A^2}I\right)^{-1}Z^\top\right)X\right)\right\}$$

For details on how to derive this, see Griffiths and Ghahramani (2006).

Therefore, the posterior distribution is $p(Z|X)$ and we can plug this likelihood into our generic discussion and ignore all mentions of θ. In Chapter 6, we will see examples of non-conjugate priors in which we do need to worry about the parameters θ.

3.2.2 MCMC for the Gamma Process

For the gamma process, only a Gibbs sampler has been developed. We review it here in the context of using a generic likelihood. For details on an application specific implementation see Titsias (2008).

Earlier in Section 3.2, we presented an overview of all the steps we need to derive the Gibbs sampler. Step (2) of sampling θ is application specific and must be addressed individually for each application, so to fill in a generic inference algorithm, we only need to address the two parts of Step (1) of the Gibbs sampler. These two steps are:

1. For non-zero columns, we must first be able to sample

$$z_{ik} \propto p(X|Z_{-ik},z_{ik},\theta)p(z_{ik}|Z_{-ik}).$$

2. For each row we must also be able to sample the number and configuration of new features unique to the i^{th} row.

For the first step, we need to consider $z_{ik} \in \mathbb{Z}^*$, so we should evaluate the right-hand side for each of these infinite number of possibilities and then normalize. The term $p(X|Z_{-ik},z_{ik},\theta)$ is the likelihood assuming we knew all of Z and θ. The term $p(z_{ik}|Z_{-ik})$ is the prior probability of z_{ik} taking on any particular value. We again can use exchangeability and compute this as if the i^{th} row were the last row to be added to Z as described in Section 2.4.3. This distribution is $z_{ik} \sim \mathrm{NB}\left(m_{-ik}, \frac{N}{N+1}\right)$ where m_{-ik} is the total count of feature k in Z excluding the i^{th} row.

To be exact, a Metropolis-Hastings step would be appropriate to sample the new value of z_{ik} since we cannot evaluate the above for all $\mathbb{Z}^*$. A close approximation

often done in practice is to sample z_{ik} from a finite number of values, for example $\{0, 1, \ldots, c\}$ for some large c since the prior probability of $z_{ik} > c$ gets arbitrarily close to zero as c increases.

For the second step, we must sample the number and configuration of features that are unique to the i^{th} row. As is the case with the IBP, this step can be tricky if the likelihood $p(X|Z, \theta)$ is not conjugate to the prior $p(\theta)$ for any feature specific parameters. We will assume that it is conjugate here but can again adapt techniques from Neal (1998) to handle non-conjugate cases.

We must directly sample the number of new features in row i and their allocation. Designate the new non-zero columns containing the allocation of the new features by z_i^{new}. Let g_i be the number of new features in z_i^{new}. Then, since given z_i^{new}, g_i is deterministic, we wish to sample z_i^{new} from

$$p(z_i^{\text{new}}|X, Z, \theta, \alpha) \quad \propto \quad p(X|Z, z_i^{\text{new}}, \theta)p(z_i^{\text{new}}|g_i)p(g_i|Z, \alpha), \tag{3.1}$$

where $p(X|Z, z_i^{\text{new}}, \theta)$ is the likelihood of X given the old features Z, the new allocation of features z_i^{new} to the i^{th} row, and the old θ; $p(z_i^{\text{new}}|g_i)$ is the probability of the allocation of features given their count; and $p(g_i|Z, \alpha)$ is the probability that g_i new features are present in row i. By exchangeability, we know from Section 2.4.3 that $g_i \sim \text{NB}\left(\alpha, \frac{N}{N+1}\right)$ and that given g_i, z_i^{new} is distributed according to the CRP.

To be completely correct, we must evaluate Equation (3.1) for all possible counts g_i and all allocations of those counts z_i^{new} and then sample from the resulting posterior. Since we cannot do this in practice, we can either sample using an Metropolis-Hastings step or, given that the distribution on g_i makes it unlikely to have g_i large, we can often evaluate g_i for only a moderate range $\{0, 1, \ldots, g_{\max}\}$ for some $g_{\max}$ and find a reasonable approximation. There are also some likelihoods that allow us to first sample g_i and then sample the configuration z_i^{new} given g_i, allowing each step to be from a narrower sample space, increasing the likelihood of getting a decent sample.

3.3 Variational Inference Algorithms

In this section, we discuss our work on variational inference algorithms for the beta process. This was the first variational approximation for a nonparametric latent feature model and was originally published as Doshi-Velez et al. (2009b). A later variational approximation for the BP was introduced by Paisley and Carin (2009). There has been no work on variational inference algorithms for the GP when used in latent feature models, but for one example of a variational approximation for a related model using the GP, see (Hoffman et al., 2010).

3.3.1 Variational Inference Algorithms for the Beta Process Overview

We derive our variational inference procedures using the IBP representation with the linear-Gaussian likelihood model introduced in Section 3.2.1, in which A and ϵ are zero mean Gaussians with variances σ_A^2 and σ_n^2 respectively. The updates can also be adapted to other exponential family likelihood models such as the infinite ICA model (Knowles and Ghahramani, 2007), but we do not review that here. Details on such a variational inference algorithm for ICA can be found in (Doshi-Velez et al., 2009c).

We denote the set of hidden variables in the IBP by $\boldsymbol{W} = \{\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{A}\}$ and the set of parameters by $\boldsymbol{\theta} = \{\alpha, \sigma_A^2, \sigma_n^2\}$. Note that here $\boldsymbol{\pi}$ is equivalent to the set of p, the weights of the Beta process introduced in Section 2.3.1, renamed to match the original naming convention of Griffiths and Ghahramani (2006). Computing the true log posterior $\ln p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}) = \ln p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta})$ is difficult due to the intractability of computing the log marginal probability $\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \int p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta}) d\boldsymbol{W}$.

Mean field variational methods approximate the true posterior with a *variational distribution* $q(\boldsymbol{W})$ from some tractable family of distributions Q (Beal, 2003; Wainwright and Jordan, 2008). Inference in this approach then reduces to finding the member $q \in Q$ that minimizes the KL divergence $D(q(\boldsymbol{W})||p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta}))$. Since the KL divergence $D(q||p)$ is non-negative and equal to zero iff $p = q$, the unrestricted solution to our problem is to set $q(\boldsymbol{W}) = p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{\theta})$. However, this general optimization problem is intractable. We therefore restrict Q to a parameterized family of distributions for which this optimization is tractable.

For the IBP, we will let Q be the factorized family

$$q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{\pi}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z}) \tag{3.2}$$

where $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, and $\boldsymbol{\nu}$ are the variational parameters that we optimize to minimize $D(q||p)$. See Figure 3.3 to visualize this approximation.

Inference then consists of optimizing the parameters of the approximating distribution to most closely match the true posterior.

This optimization is equivalent to maximizing a lower bound on the evidence:

$$\arg\max_{\tau,\phi,\nu} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - D(q||p) \quad = \quad \arg\max_{\tau,\phi,\nu} H[q] + \mathbb{E}_q[\ln(p(\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{\theta})]. \tag{3.3}$$

where $H[q]$ is the entropy of distribution q. Therefore, to minimize $D(q||p)$, we can iteratively update the variational parameters so as to maximize the right side of Equation (3.3).

We derive two mean field approximations, both of which apply a truncation level K to the maximum number of features in the variational distribution. The first minimizes the KL-divergence between the variational distribution and a finite ap-

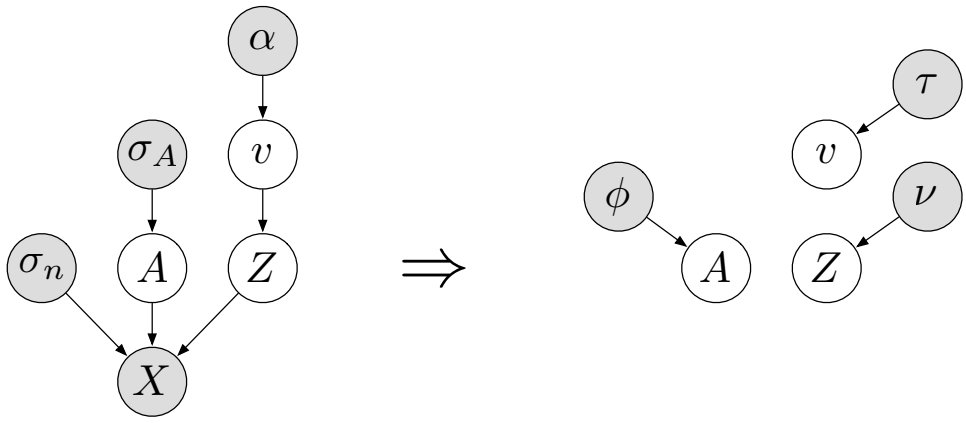


Figure 3.3: The variational approximation used for the IBP. On the left, we show the true graphical model. On the right, we show the fully factorized variational approximation along with the variational parameters to optimize.

proximation p_K to the IBP described below; we refer to this approach as the *finite variational* approach. The second approach minimizes the KL-divergence to the true IBP posterior. We call this approach the *infinite variational* method because, while our variational distribution is finite, its updates are based the true IBP posterior over an infinite number of features.

Most of the required expectations are straightforward to compute, and many of the parameter updates follow directly from standard update equations for variational inference in the exponential family (Beal, 2003; Wainwright and Jordan, 2008). We focus on the non-trivial computations. For details on the more trivial calculations, see (Doshi-Velez et al., 2009c).

3.3.2 Finite Variational Approach

The finite variational method uses a finite beta-Bernoulli approximation to the IBP as discussed in derivation two of Section 2.3.3. The finite beta-Bernoulli prior with K features first draws each feature's probability π_k independently from $\text{Beta}(\alpha/K, 1)$. Then, each z_{nk} is independently drawn from $\text{Bernoulli}(\pi_k)$ for all n.

Our finite variational approach approximates the true IBP prior $p(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})$ with $p_K(\boldsymbol{W}, \boldsymbol{X}|\boldsymbol{\theta})$ in Equation (3.3) where p_K uses the prior on Z defined by the finite beta-Bernoulli prior. While variational inference with the finite beta-Bernoulli prior is not the same as variational inference with the true IBP, the variational updates are significantly more straightforward and, in the limit of large K, the finite beta-Bernoulli approximation is equivalent to the IBP. We use a fully factor-

ized variational distribution $q_{\boldsymbol{\tau}_k}(\pi_k) = \text{Beta}(\pi_k; \tau_{k1}, \tau_{k2})$, $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathcal{N}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \boldsymbol{\Phi}_k)$, $q_{\nu_{nk}}(z_{nk}) = \text{Bernoulli}(z_{nk}; \nu_{nk})$.

3.3.3 Infinite Variational Approach

The second variational approach, similar to the one used by Blei and Jordan (2004) for the DP, uses a truncated version of the stick-breaking construction for the IBP as the approximating variational distribution q. Instead of directly approximating the distribution of π_k from the beta process in our variational algorithm, we will work with the distribution of the stick-breaking variables $\boldsymbol{v} = \{v_1, \ldots, v_K\}$. In our truncated distribution with truncation level K, the probability π_k of feature k is $\prod_{i=1}^{k} v_i$ for $k \leq K$ and zero otherwise. The advantage of using $\boldsymbol{v}$ as our hidden variable is that under the IBP prior, the $\{v_1 \ldots v_K\}$ are independent draws from the Beta distribution, whereas the $\{\pi_1 \ldots \pi_K\}$ are dependent. We therefore use the factorized variational distribution $q(\boldsymbol{W}) = q_{\boldsymbol{\tau}}(\boldsymbol{v}) q_{\boldsymbol{\phi}}(\boldsymbol{A}) q_{\boldsymbol{\nu}}(\boldsymbol{Z})$ where $q_{\boldsymbol{\tau}_k}(v_k) = \text{Beta}(v_k; \tau_{k1}, \tau_{k2})$, $q_{\boldsymbol{\phi}_k}(\boldsymbol{A}_{k\cdot}) = \mathcal{N}(\boldsymbol{A}_{k\cdot}; \bar{\boldsymbol{\phi}}_k, \Phi_k)$, and $q_{\nu_{nk}}(z_{nk}) = \text{Bernoulli}(Z_{nk}; \nu_{nk})$.

3.3.4 Variational Lower Bound

We split the expectation in Equation (3.3) into terms depending on each of the latent variables. Here, $\boldsymbol{v}$ are the stick-breaking parameters in the infinite approach; the expression for the finite Beta approximation is identical except with $\boldsymbol{\pi}$ substituted into the expectations.

$$\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) \;\geq\; & H[q] + \textstyle\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{v}} \left[\ln p(v_k|\alpha)\right] \\
& + \textstyle\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{A}} \left[\ln p(\boldsymbol{A}_{k\cdot}|\sigma_A^2)\right] \\
& + \textstyle\sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right] \\
& + \textstyle\sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{Z},\boldsymbol{A}} \left[\ln p(\boldsymbol{X}_{n\cdot}|\boldsymbol{Z}, \boldsymbol{A}, \sigma_n^2)\right]
\end{aligned}$$

In the finite Beta approximation, all of the expectations are straightforward exponential family calculations. In the infinite case, the key difficulty lies in computing the expectations $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right]$. We decompose this expectation as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk}|\boldsymbol{v})\right] &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}} \left[\ln p(z_{nk} = 1|\boldsymbol{v})^{\mathbb{I}(z_{nk}=1)} p(z_{nk} = 0|\boldsymbol{v})^{\mathbb{I}(z_{nk}=0)}\right] \\
&= \nu_{nk} \left(\textstyle\sum_{m=1}^{k} \psi(\tau_{k2}) - \psi(\tau_{k1} + \tau_{k2})\right) \\
&\quad + (1 - \nu_{nk}) \mathbb{E}_{\boldsymbol{v}} \left[\ln \left(1 - \textstyle\prod_{m=1}^{k} v_m\right)\right]
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function that its argument is true and $\psi(\cdot)$ is the digamma function. We are still left with the problem of evaluating the expectation $\mathbb{E}_{\boldsymbol{v}}[\ln(1 - \prod_{m=1}^{k} v_m)]$, or alternatively, computing a lower bound for the expression.

There are computationally intensive methods for finding arbitrarily good lower bounds for this term using a Taylor series expansion of $\ln(1-x)$. However, we present a more computationally efficient bound that is only slightly looser. We first introduce a multinomial distribution $q_k(y)$ that we will optimize to get as tight a lower bound as possible and use Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{v}}\left[\ln\left(1 - \prod_{m=1}^{k} v_m\right)\right] &= \mathbb{E}_{\boldsymbol{v}}\left[\ln\left(\sum_{y=1}^{k} q_k(y)\frac{(1-v_y)\prod_{m=1}^{y-1} v_m}{q_k(y)}\right)\right] \\
&\geq \mathbb{E}_{\boldsymbol{v}}\mathbb{E}_y\left[\ln\left((1-v_y)\prod_{m=1}^{y-1} v_m\right) - \ln q_k(y)\right] \\
&= \mathbb{E}_y\left[\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2})\right] + H(q_k).
\end{aligned}$$

These equations hold for any q_k. We take derivatives to find the q_k that maximizes the lower bound:

$$q_k(y) \propto \exp\left(\psi(\tau_{y2}) + \sum_{m=1}^{y-1}\psi(\tau_{m1}) - \sum_{m=1}^{y}\psi(\tau_{m1}+\tau_{m2})\right),$$

where the proportionality is required to make q_k a valid distribution. We can plug this multinomial lower bound for $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}\left[\ln p(z_{nk}|\boldsymbol{v})\right]$ back into the lower bound on $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ and then optimize this lower bound.

3.3.5 Parameter Updates

The parameter updates in the finite model are all straightforward updates from the exponential family (Wainwright and Jordan, 2008). In the infinite case, updates for the variational parameter for $\boldsymbol{A}$ remain standard exponential family updates. The update on $\boldsymbol{Z}$ is also relatively straightforward to compute

$$\begin{aligned}
q_{\nu_{nk}}(z_{nk}) &\propto \exp\left(\mathbb{E}_{\boldsymbol{v},\boldsymbol{A},\boldsymbol{Z}_{-nk}}\left[\ln p(\boldsymbol{W},\boldsymbol{X}|\boldsymbol{\theta})\right]\right) \\
&\propto \exp\left(\mathbb{E}_{\boldsymbol{A},\boldsymbol{Z}_{-nk}}(\ln p(X_{n\cdot}|\boldsymbol{Z}_{n\cdot},\boldsymbol{A},\sigma_n^2)) + \mathbb{E}_{\boldsymbol{v}}(\ln p(z_{nk}|\boldsymbol{v}))\right),
\end{aligned}$$

where we can again approximate $\mathbb{E}_{\boldsymbol{v}}(\ln p(z_{nk}|\boldsymbol{v}))$ with a Taylor series or the multinomial method presented in Section 3.3.4.

The update for the stick-breaking variables $\boldsymbol{v}$ is more complex because the variational updates no longer stay in the exponential family due to the terms $\mathbb{E}_{\boldsymbol{v}}(\ln p(z_{nk}|\boldsymbol{v}))$.

If we use a Taylor series approximation for this term, we no longer have closed form updates for $\boldsymbol{v}$ and must resort to numerical optimization. If we use the multinomial lower bound, then for fixed $q_k(y)$, terms decompose independently for each v_m and we get a closed form exponential family update. We will use the latter approach in our results section.

3.3.6 Truncation Error

Both of our variational inference approaches require us to choose a truncation level K for our variational distribution. Building on results from Thibaux and Jordan (2007) and Teh et al. (2007), we present bounds on how close the marginal distributions are when using a truncated stick-breaking prior and the true IBP stick-breaking prior. Our development parallels bounds for the Dirichlet Process by Ishwaran and James (2001) and presents the first such truncation bounds for the IBP.

Intuitively, the error in the truncation will depend on the probability that, given N observations, we observe features beyond the first K in the data (otherwise the truncation should have no effect). Let us denote the marginal distribution of observation X by $m_\infty(X)$ when we integrate over W drawn from the IBP. Let $m_K(X)$ be the marginal distribution when W are drawn from the truncated stick-breaking prior with truncation level K.

Using the Beta Process representation for the IBP (Thibaux and Jordan, 2007) and using an analysis similar to the one in Ishwaran and James (2001), we can show that the difference between these distributions is at most

$$\begin{aligned}
\frac{1}{4}\int |m_K(X) - m_\infty(X)| dX &\leq \Pr(\exists k > K, n \text{ with } z_{nk} = 1) \\
&= 1 - \Pr\left(\text{all } z_{ik} = 0,\ i \in \{1, \ldots, N\}, k > K\right) \\
&= 1 - \mathbb{E}\left[\left(\prod_{i=K+1}^{\infty} (1 - \pi_i)\right)^N\right]. \qquad (3.4)
\end{aligned}$$

We present here one formal bound for this difference. We have listed several similar bounds in Doshi-Velez et al. (2009c) that can be derived directly by applying Jensen's inequality to the expectation above as well as a heuristic bound which tends to be tighter in practice.

We begin the derivation of the truncation bound by applying Jensen's inequality to Equation (3.4):

$$-\mathbb{E}\left[\left(\prod_{i=K+1}^{\infty} (1 - \pi_i)\right)^N\right] \leq -\left(\mathbb{E}\left[\prod_{i=K+1}^{\infty} (1 - \pi_i)\right]\right)^N. \qquad (3.5)$$

The Beta Process construction for the IBP implies that the sequence $\pi_1, \pi_2, \ldots$ can be modeled as a Poisson process on the unit interval $(0, 1)$ with rate $\nu(x)dx = \alpha x^{-1}dx$. It follows that the unordered truncated sequence $\pi_{K+1}, \pi_{K+2}, \ldots$ may be modeled as a Poisson process on the interval $(0, \pi_K)$ with the same rate. The Levy-Khinchine formula states that the moment generating function of a Poisson process X with rate ν can be written as

$$\mathbb{E}[\exp(f(X))] = \exp\left(\int (\exp(f(x)) - 1)\nu(x)dx\right),$$

where $f(X) = \sum_{x \in X} f(x)$. We apply the Levy-Khinchine formula to simplify the inner expectation of Equation (3.5):

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=K+1}^{\infty}(1-\pi_i)\right] &= \mathbb{E}\left[\exp\left(\sum_{i=K+1}^{\infty}\ln(1-\pi_i)\right)\right] \\
&= \mathbb{E}_{\pi_K}\left[\exp\left(\int_0^{\pi_K}(\exp(\ln(1-x))-1)\,\nu(x)dx\right)\right] \\
&= \mathbb{E}_{\pi_K}\left[\exp\left(-\alpha\pi_K\right)\right].
\end{aligned}$$

Finally, we apply Jensen's inequality, using the fact that π_K is the product of independent Beta$(\alpha, 1)$ variables to get

$$\begin{aligned}
\mathbb{E}_{\pi_K}[\exp(-\alpha\pi_K)] &\geq \exp\left(\mathbb{E}_{\pi_k}[-\alpha\pi_K]\right) \\
&= \exp\left(-\alpha\left(\frac{\alpha}{1+\alpha}\right)^K\right).
\end{aligned}$$

Substituting the expression into Equation (3.5) gives

$$\frac{1}{4}\int |m_K(X) - m_\infty(X)|dX \leq 1 - \exp\left(-N\alpha\left(\frac{\alpha}{1+\alpha}\right)^K\right). \tag{3.6}$$

Similar to truncation bound for the Dirichlet Process, we see that for fixed K, the expected error increases with N and α— the factors that increase the expected number of features in a data set. However, the bound decreases exponentially quickly as K is increased.

Figure 3.4 shows our truncation bound and the true L_1 distance based on 1000 Monte Carlo simulations of an IBP matrix with $N = 30$ observations and $\alpha = 5$. As expected, the bound decreases exponentially fast with the truncation level K. However, the bound is fairly loose. In practice, we find that heuristic bound using

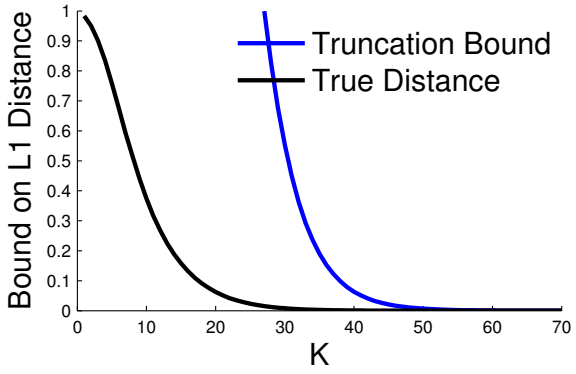


Figure 3.4: Truncation bound and true L_1 distance.

Taylor expansions (see extended version) provides much tighter estimates of the loss.

3.4 Comparison of MCMC and Variational Inference Algorithms for the Beta Process

We compared the variational approaches with both Gibbs sampling and particle filtering. Mean field variational algorithms are only guaranteed to converge to a *local* optimum, so we applied standard optimization tricks to improve performance. Each run was given a number of random restarts and the hyperparameters for the noise and feature variance were tempered to smooth the posterior. We also experimented with several other techniques such as gradually introducing data and merging correlated features that were less useful as the size and dimensionality of the data sets increased; they were not included in the final experiments.

The sampling methods we compared against were the collapsed Gibbs sampler described in Section 3.2.1 and a partially-uncollapsed alternative in which instantiated features are explicitly represented and new features are integrated out. In contrast to the variational methods, the number of features present in the IBP matrix will adaptively grow or shrink in the samplers. To provide a fair comparison with the variational approaches, we also tested finite variants of the collapsed and uncollapsed Gibbs samplers. Finally, we also tested against the particle filter of Wood and Griffiths (2006). All sampling methods were annealed and given an equal number of restarts as the variational methods.

Both the variational and Gibbs sampling algorithms were heavily optimized for efficient matrix computation so we could evaluate the algorithms both on their running times and the quality of the inference. For the particle filter, we used the imple-

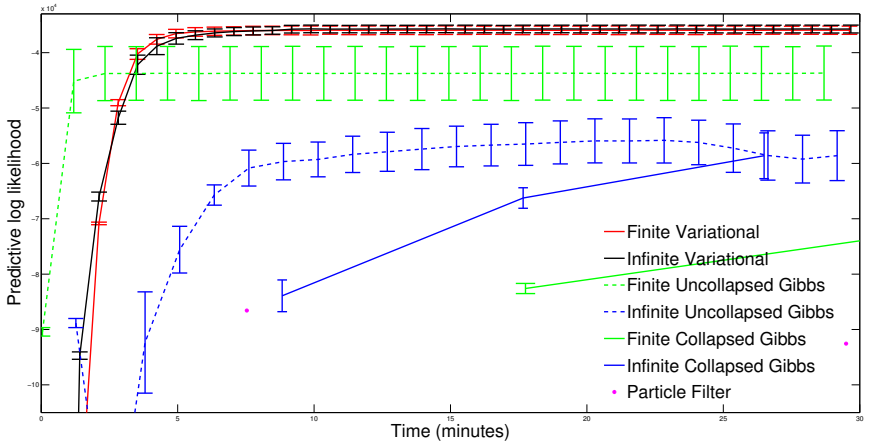


Figure 3.5: Evolution of test log-likelihoods over a thirty-minute interval for $N = 500$, $D = 500$, and $K = 20$. The finite uncollapsed Gibbs sampler has the fastest rise but gets caught in a lower optima than the variational approach.

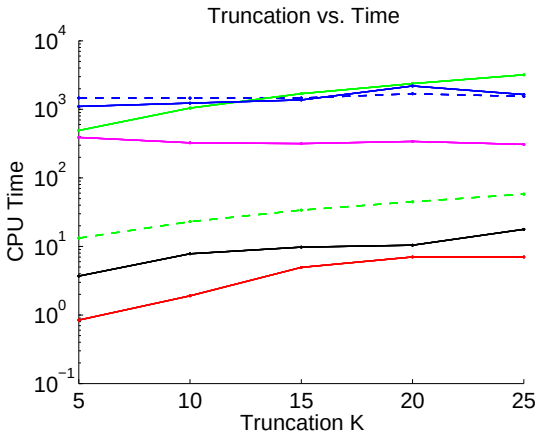
mentation provided by Wood and Griffiths (2006). To measure the quality of these methods, we held out one third of the observations on the last half of the data set. Once the inference was complete, we computed the predictive likelihood of the held out data and averaged over restarts.

3.4.1 Synthetic Data

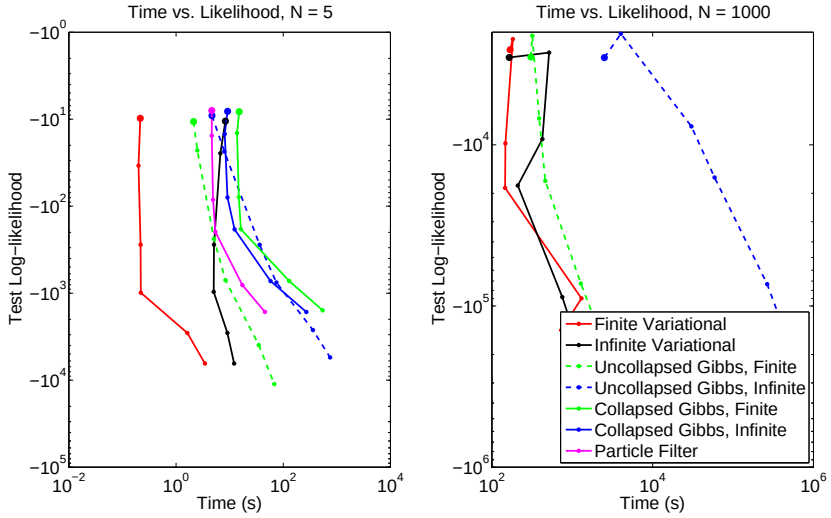
The synthetic data sets consisted of Z and A matrices randomly generated from the truncated stick-breaking prior. Figure 3.5 shows the evolution of the test-likelihood over a thirty minute interval for a data set with 500 observations of 500 dimensions each generated with 20 latent features.[1] The error bars indicate the variation over the 5 random starts. The finite uncollapsed Gibbs sampler (dotted green) rises quickly but consistently gets caught in a lower optima and has higher variance. This variance is not due to the samplers mixing, but instead due to each sampler getting stuck in widely varying local optima. The variational methods are slightly slower per iteration but soon find regions of higher predictive likelihoods. The remaining samplers are much slower per iteration, often failing to mix within the allotted interval.

Figures 3.6(a) and 3.6(b) show results from a systematic series of tests in which we

[1]The particle filter must be run to completion before making prediction, so we cannot test its predictive performance over time. We instead plot the test likelihood only at the end of the inference for particle filters with 10 and 50 particles (the two magenta points).



(a)



(b)

Figure 3.6: Inference algorithm comparison. (a) Time versus truncation (K). The variational approaches are generally orders of magnitude faster than the samplers (note log scale on the time axis). (b) Time versus log-likelihood plot for $K = 20$. The larger dots correspond to $D = 5$ the smaller dots to $D = 10, 50, 100, 500, 1000$.

tested all combinations of observation count $N = \{5, 10, 50, 100, 500, 1000\}$, dimensionality $D = \{5, 10, 50, 100, 500, 1000\}$, and truncation level $K = \{5, 10, 15, 20, 25\}$. Each of the samplers was run for 1000 iterations on three chains and the particle filter was run with 500 particles. For the variational methods, we used a stopping criterion that halted the optimization when the variational lower bound between the current and previous iterations changed by a multiplicative factor of less than 10^{-4} and the tempering process had completed.

Results are shown in Figure 3.6. Figure 3.6(a) shows how the computation time scales with the truncation level. The variational approaches and the uncollapsed Gibbs are consistently an order of magnitude faster than other algorithms. Figure 3.6(b) shows the interplay between dimensionality, computation time, and test log-likelihood for data sets of size $N = 5$ and $N = 1000$ respectively. For $N = 1000$, the collapsed Gibbs samplers and particle filter did not finish, so they do not appear on the plot. We chose $K = 20$ as a representative truncation level. Each line represents increasing dimensionality for a particular method (the large dot indicates $D = 5$, the subsequent dots correspond to $D = 10$, 50, etc.). The nearly vertical lines of the variational methods show that they are quite robust to increasing dimension. As dimensionality and data set size increase, the variational methods become increasingly faster than the samplers. By comparing the lines across the likelihood dimension, we see that for the very small data set, the variational method often has a lower test log-likelihood than the samplers. In this regime, the samplers are fast to mix and explore the posterior. However, the test log-likelihoods are comparable for the larger data set.

3.4.2 Real Data

We next tested two real-world data sets to show how our approach fared with complex, noisy data not drawn from the IBP prior (our main goal was not to demonstrate low-rank approximations). The Yale Faces (Georghiades et al., 2001) data set consisted of 721 32x32 pixel frontal-face images of 14 people with varying expressions and lighting conditions. We set σ_a and σ_n based on the variance of the data. The speech data set consisted of 245 observations sampled from a 10-microphone audio recording of 5 different speakers. We applied the ICA version of our inference algorithm, where the mixing matrix S modulated the effect of each speaker on the audio signals. The feature and noise variances were taken from an initial run of the Gibbs sampler where σ_n and σ_a were also sampled.

Tables 3.1 and 3.2 show the results for each of the data sets. All Gibbs samplers were uncollapsed and run for 200 iterations.[2] In the higher dimensional Yale data set,

[2]On the Yale data set, we did not test the collapsed samplers because the finite collapsed Gibbs

the variational methods outperformed the uncollapsed Gibbs sampler. When started from a random position, the uncollapsed Gibbs sampler quickly became stuck in a local optima. The variational method was able to find better local optima because it was initially very uncertain about which features were present in which data points; expressing this uncertainty explicitly through the variational parameters (instead of through a sequence of samples) allowed it the flexibility to improve upon its bad initial starting point.

The story for the speech data set, however, is quite different. Here, the variational methods were not only slower than the samplers, but they also achieved lower test-likelihoods. The evaluation on the synthetic data sets points to a potential reason for the difference: the speech data set is much simpler than the Yale data set, consisting of 10 dimensions (vs. 1032 in the Yale data set). In this regime, the Gibbs samplers perform well and the approximations made by the variational method become apparent. As the dimensionality grows, the samplers have more trouble mixing, but the variational methods are still able to find regions of high probability mass.

Table 3.1: Running times in seconds and test log-likelihoods for the Yale Faces data set.

Algorithm	**K**	**Time**	**Test Log-Likelihood** ($\times 10^6$)
Finite Gibbs	5	464.19	-2.250
	10	940.47	-2.246
	25	2973.7	-2.247
Finite Variational	5	163.24	-1.066
	10	767.1	-0.908
	25	10072	-0.746
Infinite Variational	5	176.62	-1.051
	10	632.53	-0.914
	25	19061	-0.750

3.4.3 Summary

The combinatorial nature of the BP/IBP poses specific challenges for sampling-based inference procedures. Whereas sampling methods work in the discrete space of binary matrices, the variational method allows for soft assignments of features because

sampler required one hour per iteration with $K = 5$ and the infinite collapsed Gibbs sampler generated one sample every 50 hours. In the iICA model, the features $\boldsymbol{A}$ cannot be marginalized.

Table 3.2: Running times in seconds and test log-likelihoods for the speech data set.

Algorithm	**K**	**Time**	**Test Log-Likelihood**
Finite Gibbs	2	56	-0.7444
	5	120	-0.4220
	9	201	-0.4205
Infinite Gibbs	na	186	-0.4257
Finite Variational	2	2477	-0.8455
	5	8129	-0.5082
	9	8539	-0.4551
Infinite Variational	2	2702	-0.8810
	5	6065	-0.5000
	9	8491	-0.5486

it approaches the inference problem as a continuous optimization. We showed experimentally that, especially for high dimensional problems, the soft assignments allow the variational methods to explore the posterior space faster than sampling-based approaches.

Chapter 4

Priors for Non-exchangeable Bayesian Nonparametric Latent Feature Models

As was mentioned in Chapter 2, there are several ways people can extend or generalize the priors for the beta-Bernoulli and gamma-Poisson nonparametric latent feature models. In this chapter, we focus on relaxing one assumption of these priors, the assumption of exchangeability.

While exchangeability is appropriate in some applications (e.g., bag-of-words models for documents), exchangeability is sometimes assumed simply for computational reasons; non-exchangeable models might be a better choice for applications based on subject matter. Drawing on ideas from graphical models, in this chapter, we present a framework for deriving non-exchangeable generalizations and provide two concrete examples for both the beta and gamma processes for which reasonable approximate posterior inference algorithms still exist. Our priors are applicable to the general setting in which the known dependencies between objects can be expressed using a tree, in which edge lengths indicate the strength of relationships, or using a Markov chain.

This chapter introduces the framework for working with these priors as well as the priors themselves. The main idea behind this work is that previously, we defined rich completely random measures as our De Finetti mixing distributions and sampled z_i conditionally independently given these measures. This is the simplest way we could sample z_i given B. Why not impose a richer way to sample the z_i given B, so that there can be dependence on the different draws? In our work, we assume we have some prior knowledge about the relationships between the observations or entities in our models and we wish to use that prior knowledge to induce richer dependencies amongst the z_i. In the next chapter, Chapter 5, we discuss inference algorithms for

models using these priors. This work has been previously presented at conferences and workshops in Miller et al. (2008a,b, 2010).

4.1 Alternate Views of the Exchangeable Priors

We begin by defining an alternate, but equivalent, view of the way matrices are generated in the exchangeable versions of the BP and GP. Though this view is more complicated than the view presented in Chapter 2, it will allow us to generalize our techniques to non-exchangeable priors. We then discuss our desiderata for how to generalize these priors in Section 4.2 and then present our nonexchangeable framework and each of our two non-exchangeable variations.

As before, we first draw B from either the BP or GP to get $\{(p_k, \omega_k)\}_{k=1}^{\infty}$. What is different is how we generate z_i. In this section, we construct stochastic processes that generate z_i in a way that is equivalent to the conditionally independent draws of the exchangeable prior, but that allows us to generalize into richer stochastic processes for generating z_i without drastically increasing the cost of posterior inference algorithms. We define these priors differently for the BP and GP, but the similarities between the derivations are quite obvious.

4.1.1 Alternate Views of the Beta Process

Once we have generated

$$B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k},$$

from the BP, then in the Bernoulli process, we sample each z_{ik} conditionally independently with probability p_k. We wish to develop an alternate view of this conditionally independent Bernoulli sampling.

Let $\gamma_k = -\log(1 - p_k)$. Now we generate each column based on either the tree or independent chains in Figure 4.1 with a zero at the root that mutates into a one along each edge with exponential rate γ_k or equivalently with independent zeros mutating to a one. Each entry is still a conditionally independent Bernoulli(p_k) draw.

While this might seems like an odd way to sample from a Bernoulli(p_k) distribution, the idea of mutations on a tree or chain will help us derive our non-exchangeable generalizations. This prior is obviously still equivalent to the beta-Bernoulli prior we introduced in Chapter 2.

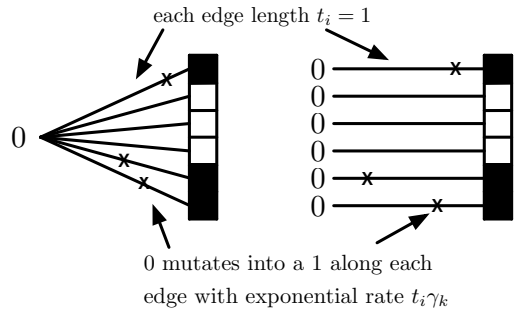


Figure 4.1: Alternate view on how to generate columns from the beta-Bernoulli prior.

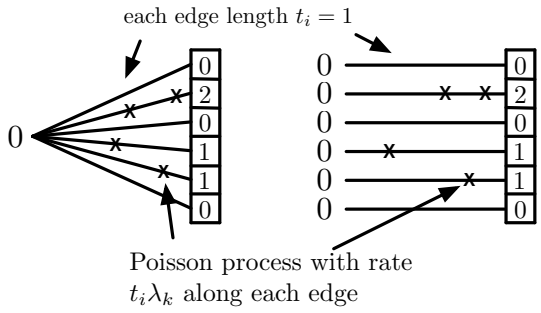


Figure 4.2: Alternate view on how to generate columns from the gamma-Poisson prior.

4.1.2 Alternate Views of the Gamma Process

Once we have generated

$$B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k},$$

from the GP, then we sample each z_{ik} conditionally independently from a Poisson(p_k). We wish to develop an alternate view of this conditionally independent Poisson sampling.

We now generate each column by letting there be an independent Poisson process with rate p_k along each edge in either the tree or independent chains in Figure 4.2 with each entry z_{ik} being the number of events that have occurred from the root to the leaf. Each entry is still a conditionally independent Poisson(p_k) draw.

While this might seems like an odd way to sample from a Poisson(p_k) distribution, the idea of counting events on a tree or chain will help us derive our non-exchangeable generalizations. This prior is obviously still equivalent to the gamma-Poisson prior we introduced in Chapter 2.

4.2 Desiderata for Non-Exchangeable Generalizations

We will show how these alternate views can be generalized when object relationships can be captured through a known tree or chain. There are several different ways to generalize these priors, so we lay out some desiderata motivating our particular generalizations.

Our desiderata are:

1. Each row z_i should be marginally BeP(B) or PP(B). In other words, marginally each element in column k is Beta(p_k) or Poisson(p_k) as in the BP and GP priors, respectively.

2. The prior should be consistent if we integrate out z_i for one or more i.

3. By changing the structure of our tree or chain, we can smoothly interpolate between the BeP or PP where all elements in a column are conditionally independent (exchangeable) and priors in which entries in columns are fully dependent. This allows us to vary how much prior information we include in our prior distributions.

We will show how to generalize our tree and chain-based framework in such a way that satisfies these desiderata. Before we present these two variations for both the BP and GP priors, we should note that there exists one other chain-based BP prior, the Markov IBP (mIBP) introduced by Van Gael et al. (2009). Despite still being useful, the mIBP was developed using different desiderata and does not satisfy any of our desiderata.

4.3 Non-Exchangeable Generalizations

We are now ready to present our non-exchangeable generalizations. In Section 4.4, we present the tree-based generalizations of both the BP and the GP. In Section 4.5, we present the chain-based generalizations of both the BP and the GP. These are just two of the non-exchangeable variations that can be developed by having a more complex stochastic process built on top of completely random measures instead of a simple exchangeable prior. Other variations are possible and should be pursued as applications require them.

When developing a non-exchangeable variation, there are two things that must be kept in mind. First is that the non-exchangeable structure must be amenable to tractable inference algorithms. Our two examples here are trees and chains, which are well known to have tractable inference algorithms (Pearl, 1988). Other variations might include more complex graphical models with low tree-width or other stochastic

processes on trees or chains. We will see how performing inference in these structures is a required part of our posterior inference algorithms.

The second thing that must be kept in mind is that we must define our priors in such a way that we can compute the full posterior distributions and still have a valid nonparametric model, meaning that there are always an infinite number of unobserved features that we know how to sample from. This means that in the case of the BP, the probability of a particular feature being unobserved in N observations, no matter if we are using trees, chains, or any other structure, must be proportional to the weighted sum of at least one improper (and potentially some proper) beta distributions and in the GP, the probability of a feature being unobserved must be proportional to the weighted sum of at least one improper (and potentially some proper) gamma distributions. Then we can use the ideas of Kim (1999) and Wolpert and Ickstadt (1998a) to compute the posterior completely random measures. We must also define our prior in such a way that whenever we have a non-zero feature, the posterior distribution of the atom in B corresponding to that feature becomes proper. This falls out of the fact that we will define these priors such that each z_{ik} is marginally Bernoulli(p_k) or Poisson(p_k) in order to satisfy the second of our desiderata.

These two aspects are key to our development of these priors. We now present our non-exchangeable priors for Z. In each generalization, we start by defining the non-exchangeable stochastic process for each feature, discuss the relevant posteriors, and then present a generative process for Z with the underlying Lévy process partially marginalized out. One key difference in the marginalized representations (the IBP and IGPFM forms) is that we cannot fully integrate out B and remain tractable. This is because we need to condition on the atoms and weights of B corresponding to the previously observed features of Z in order to take advantage of tractable inference algorithms on chains and trees. Without sampling this part of B, we cannot remain tractable. We will still marginalize out the parts of B corresponding to unobserved features, though. We do not discuss any stick-breaking processes because the prior completely random measures are identical to those of the exchangeable priors.

4.4 Tree-based Generalizations

Why should restrict ourselves to flat trees in Figures 4.1 and 4.2? Below we discuss how to generalize both the BP and the GP to use trees representing different assumptions about the relationships amongst objects in order to use prior knowledge in these prior distributions as indicated in Figure 4.3.

In the exchangeable BP and GP, for each k, entries $\{z_{ik}\}_{i=1}^{N}$ are independent, but in the tree-based generalizations, the entries will be dependent with the pattern of

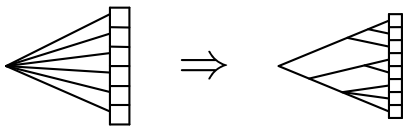


Figure 4.3: Framework for our tree-based generalizations.

dependence captured by a stochastic process on a rooted tree similar to models used in phylogenetics. In this tree, the N objects being modeled are at the leaves, and lengths are assigned to edges in such a way that the total edge length from the root to any leaf is equal to one. By defining a stochastic process on this tree to jointly generate the entries $\{z_{ik}\}_{i=1}^{N}$, we develop priors such that objects more closely related in the tree are more likely to have similar features than objects farther way in the tree. By using the sum-product algorithm for efficient inference on trees, we are then able to define a nonparametric prior for which posterior inference is nearly as efficient as inference developed for the exchangeable special case.

What can these priors be used for? They can be used when objects have a known relationship captured by a tree, such as in phylogenetics; when objects are grouped according to partial or even full exchangeability; and more generally, whenever we have similarity data about objects.

4.4.1 Tree-based BP

In this section, we present a non-exchangeable version of the BP in which relationships between objects are expressed via a known, fixed tree. This prior was introduced as the *phylogenetic Indian Buffet Process* (pIBP) in (Miller et al., 2008a).

4.4.1.1 Tree-based BP Stochastic Process

As in the exchangeable BP, we first sample

$$B \;\sim\; \mathrm{BP}(c, B_0).$$

Now instead of sampling $z_i \overset{\text{i.i.d}}{\sim} \mathrm{BeP}(B)$ or equivalently, sampling $z_{ik} \overset{\text{i.i.d}}{\sim} \mathrm{Bernoulli}(p_i)$, we sample $\{z_{ik}\}_{i=1}^{N}$ jointly while satisfying all the desiderata from Section 4.2. As in the alternate view of the BP in Section 4.1.1, we now define

$$\gamma_k \;=\; -\log(1-p_k).$$

We assume we have a known tree expressing the object relationships in which the objects appear at the leaves as shown in Figure 4.4(a). For each k, we then indepen-

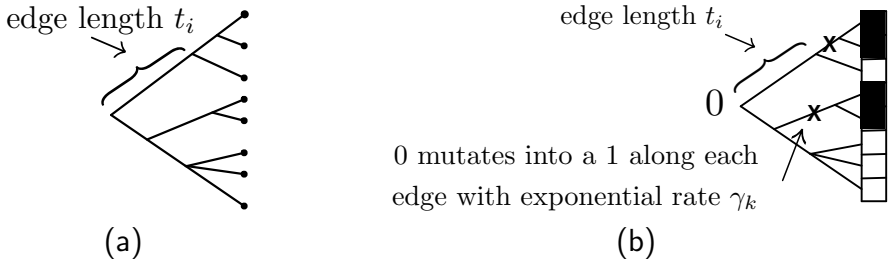


Figure 4.4: Tree-based BP per-feature stochastic process. (a) The tree representing the known object relationships for our tree-based generalizations. (b) The joint stochastic process for $\{z_{ik}\}_{i=1}^{N}$ in the tree-based BP generalization in which an $\times$ marks a mutation event, a black box indicates the corresponding z_{ik} is one, and a white box indicates it is zero.

dently generate each set of features $\{z_{ik}\}_{i=1}^{N}$ jointly by defining the stochastic process shown in Figure 4.4(b). To generate the entries of the k^{th} column, we proceed as follows. Assign the value zero to the root node of the tree. Along any path from the root to a leaf, let this value change to a one along each edge with exponential rate γ_k. That is, along an edge of length t, let the probability of changing from a zero to a one be $1 - \exp(-\gamma_k t)$. Once the value has changed to a one along any path from the root, all leaves below that point are assigned the value one as shown in Figure 4.4(b).

The parameterization $\gamma_k = -\log(1 - p_k)$ is convenient because it ensures that p_k remains the marginal probability that any single feature is equal to one. To see this, note that since every leaf node is at distance one from the root, for any entry in the matrix,

$$p(z_{ik} = 1|p_k) = 1 - \exp(-(-\log(1 - p_k))) = p_k$$

which also guarantees that we recover the beta-Bernoulli prior in the special case where all branches join at the root, as in the left part of Figure 4.3. By varying the tree structure, we can smoothly move from a fully exchangeable prior to a prior in which all object features must be identical. It is a simple corollary that any set of objects characterized by a set of branches that meet at a single point will be exchangeable within that set, meaning that the tree can be used to capture notions of partial exchangeability. Also, since this has been defined as a directed graphical model, it trivially consistent under marginalization. From this, we can see that we satisfy all our desiderata.

4.4.1.2 Tree-based BP Conditional Distributions

Now that we have defined the stochastic process, we show how to evaluate conditional probabilities in this prior. We treat the tree as a directed graph with variables at each of the interior nodes and z_{ik} at each leaf i. Then, given p_k, or equivalently γ_k, if there is a length t edge from a parent node x to a child node y, we have

$$\begin{aligned} p(y=0|x=0,\gamma_k) &= \exp(-\gamma_k t) \\ p(y=1|x=0,\gamma_k) &= 1-\exp(-\gamma_k t) \\ p(y=0|x=1,\gamma_k) &= 0 \\ p(y=1|x=1,\gamma_k) &= 1 \end{aligned}$$

as the conditional probabilities that define our tree-structured graphical model.

Expressing this process as a graphical model makes it possible to efficiently compute various conditional probabilities that are relevant for posterior inference. Specifically, we will need to evaluate

$$p(z_{ik}|z_{(-i)k},p_k) \tag{4.1}$$

for $z_{ik} \in \{0,1\}$ and

$$p\left(\{z_{ik}\}_{i=1}^N|p_k\right), \tag{4.2}$$

which are trivial in the beta-Bernoulli prior due to the conditional independence of z_{ik}, but more challenging in the tree-based BP where z_{ik} are no longer conditionally independent. To compute Equation (4.1), we use the sum-product algorithm (Pearl, 1988). In order to calculate Equation (4.2), we use the chain rule of probability to get a set of terms similar to Equation (4.1), the difference being that the posterior in each term is conditional only on a subset of the other variables. Each term can be reduced to a simple sum-product calculation by marginalizing over all variables that do not appear in that term, which can be done easily since all variables appear at the leaves of the tree. Both Equation (4.1) and Equation (4.2) can be calculated in $O(N)$ time by a dynamic program.

We can also compute the posterior of B given observations $z_1,\ldots,z_N$. Recall Equation (2.13) for the posterior of B in case of the exchangeable BP.

$$\begin{aligned} B|z_1,\ldots,z_N &\sim \mathrm{BP}(c+N,B_N) \\ \text{where } B_N &= \frac{c}{c+N}B_0+\frac{1}{c+N}\sum_{i=1}^N z_i. \end{aligned}$$

We will not have as simple a posterior for any of the non-exchangeable priors, but we can still reason about the posterior efficiently. We can again use Theorem 3.3 of Kim (1999) to reason about the continuous and discrete parts of the posterior of B independently. First note that if $\sum t$ is the sum of all edge lengths in the tree, then the probability of all entries corresponding to p_k being zero is $\exp\left(-\gamma_k \sum t\right) = (1-p_k)^{\sum t}$. This gives us that the continuous part of B has posterior

$$\begin{aligned} B_{\text{continuous}}|z_1,\ldots,z_N &\sim \text{BP}\left(c+\sum t, B_{N,\text{continuous}}\right) \\ \text{where } B_{N,\text{continuous}} &= \frac{c}{c+\sum t} B_0. \end{aligned} \tag{4.3}$$

For the discrete part, the atoms still appear only at locations where we have a non-zero z_{ik}, but there is no closed form that applies to all trees. However, for any particular tree, we can compute the posterior of p_k for all k with a non-zero feature. In the exchangeable case, the probability of seeing $\{z_{ik}\}_{i=1}^N$ is $p^{\sum_{i=1}^N z_{ik}}(1-p)^{N-\sum_{i=1}^N z_{ik}}$, which immediately gives us that the posterior of p_k is exactly $\text{Beta}\left(\sum_{i=1}^N z_{ik}, c+N-\sum_{i=1}^N\right)$. In the non-exchangeable case, the probability of seeing $\{z_{ik}\}_{i=1}^N$ given p_k is discussed above in Equation 4.2. Since there is at least one non-zero entry for the discrete part of B, the posterior when combined with the improper $cp^{-1}(1-p)^{c-1}$ prior will be proper.

4.4.1.3 Tree-based IBP

Just as we can marginalize out B in the exchangeable BP to directly sample Z via the IBP, we can again marginalize out B in the tree-based BP to sample Z via a process called the *phylogenetic Indian Buffet Process* (pIBP) (Miller et al., 2008a). As in the IBP in Section 2.3.3, the pIBP can be derived either by working with a marginalized BP or by defining the limit of a finite beta-Bernoulli model. The latter approach was taken in Miller et al. (2008a).

The pIBP can again be understood in terms of a culinary metaphor, in which each row of Z is viewed as the choices made by a diner in a buffet line, and in which we specify how each diner chooses their dishes based on the dishes chosen by previous diners. We present this process here, leaving derivations for Appendix 4.A.1. Consider a large extended family that is about to choose dishes at a buffet. Assume that we are given a tree describing the genealogical relationships of the family members and assume that dining preferences are related to genealogy. In particular, family members who are more closely related have more similar preferences. Therefore, as each diner moves through the buffet line, their choice of dishes will be more dependent on the selections of previous diners who are closely related to them and less dependent on the selections of other diners.

The pIBP generative process is specified as follows.

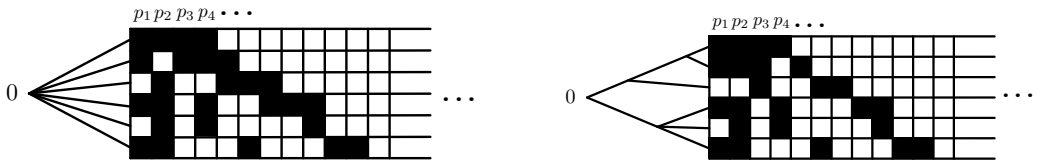
- The first diner (arbitrarily chosen) starts at the head of a buffet line that has infinitely many dishes. This person tries Poisson(α) dishes and also adds a brief annotation to each of these dishes, p_k, drawn uniformly from $[0, 1]$. This note, through its previously described equivalent representation, $\gamma_k = -\log(1 - p_k)$, will allow us to efficiently compute the probability that subsequent diners choose the k^{th} dish using the sum-product algorithm.
- Each subsequent diner enters the buffet line and samples some previously tasted dishes as well as some new ones.
  - Based on the annotations attached to the previously sampled dishes as well as the identity of previous diners, the i^{th} diner samples the k^{th} dish according to the probability in Equation (4.1) where $z_{(-i)k}$ indicates which of the previous diners have chosen the k^{th} dish. Through the stochastic process on the tree, if closely related diners have tried a dish, the current diner is more likely to also sample it. The preferences of all diners who have not entered the buffet line are ignored, which can be done by only performing the sum-product algorithm on the minimal subtree from the root containing the current diner and all previous diners.
  - Each diner also samples a number of new dishes. If t_i is the length of the branch connecting diner i to the rest of the minimal subtree just described, $\sum t$ is the total length of the rest of this subtree, and $\psi(\cdot)$ is the digamma function, then diner i tries

$$\text{Poisson}\left(\alpha\left(\psi\left(\sum t + t_i + 1\right) - \psi\left(\sum t + 1\right)\right)\right) \tag{4.4}$$

    new dishes. They also add an annotation, p_k, to each of the new dishes that will be used for future inferences, where the density of p_k is proportional to

$$\left(1 - (1 - p_k)^{t_i}\right)(1 - p_k)^{\sum t} p_k^{-1}. \tag{4.5}$$

This process repeats until all diners have gone through the buffet line, defining a matrix Z as in the IBP as seen in Figure 4.5(b). Though this process is not exchangeable, we can let any family member go first and get the same marginal distribution. This means that each row of Z has a Poisson(α) number of non-zero columns, yielding a sparse matrix as in the IBP. The IBP is the special case of the pIBP corresponding to the tree shown in Figure 4.5(a); this fact can be derived from the pIBP presented here by using identities of the digamma function on the integers,



(a) The IBP (b) The pIBP

Figure 4.5: The tree-based IBP. (a) The exchangeable IBP is a special case of the tree-based IBP (pIBP) where all branches meet at the root. (b) Different structure trees capture dependencies among featural representations of objects.

that is $\psi(N) = H_{N-1} - \gamma$, where H_i is the i^{th} harmonic number and γ is the Euler constant.

4.4.2 Tree-based GP

In this section, we present a non-exchangeable version of the GP in which relationships between objects are expressed via a known, fixed tree.

4.4.2.1 Tree-based GP Stochastic Process

As in the exchangeable GP, we first sample

$$B \;\sim\; \mathrm{GP}(c, B_0).$$

Now instead of sampling $z_i \overset{\text{i.i.d}}{\sim} \mathrm{PP}(B)$ or equivalently, sampling $z_{ik} \overset{\text{i.i.d}}{\sim} \mathrm{Poisson}(p_i)$, we sample $\{z_{ik}\}_{i=1}^{N}$ jointly while satisfying all the desiderata from Section 4.2.

We assume we have a known tree expressing the object relationships in which the objects appear at the leaves as shown in Figure 4.6(a). For each k, we then independently generate each set of features $\{z_{ik}\}_{i=1}^{N}$ jointly by defining the stochastic process shown in Figure 4.6(b). To generate the entries of the k^{th} column, we proceed as follows. Assign the value zero to the root node of the tree. Let there be a Poisson process on the tree with rate p_k. This implies that for an edge of length t_i, there will be a Poisson$(p_k t_i)$ number of events. Now let the values z_{ik} be the total number of events on the path from the root to the node corresponding to i. Therefore, nodes that are closer in the tree are more highly correlated.

Since the total path from root to any node is of length one, each entry z_{ik} is marginally Poisson(p_k), which guarantees that we recover the gamma-Poisson prior in the special case where all branches join at the root, as in the left part of Figure 4.3. By varying the tree structure, we can smoothly move from a fully exchangeable

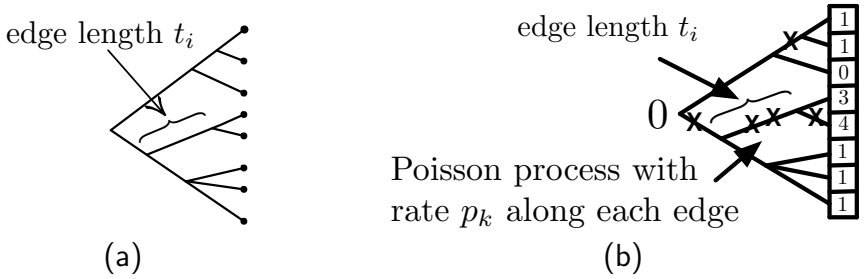


Figure 4.6: Tree-based GP per-feature stochastic process. (a) An example tree representing the known object relationships for our tree-based generalizations. (b) The joint stochastic process for $\{z_{ik}\}_{i=1}^{N}$ in the tree-based GP generalization.

prio to a prior in which all object features must be identical. It is a simple corollary that any set of objects characterized by a set of branches that meet at a single point will be exchangeable within that set, meaning that the tree can be used to capture notions of partial exchangeability. Also, since this has been defined as a directed graphical model, it trivially consistent under marginalization. From this, we can see that we satisfy all our desiderata.

4.4.2.2 Tree-based GP Conditional Distributions

Now that we have defined the stochastic process, we show how to evaluate conditional probabilities in this prior. We treat the tree as a directed graph with variables at each of the interior nodes and z_{ik} at each leaf i. Then, given p_k, if there is a length t edge from a parent node x to a child node y, we have

$$p(y - x|x, p_k) \quad \sim \quad \text{Poisson}(t_i p_k)$$

as the conditional probability that define our tree-structured graphical model.

Expressing this process as a graphical model makes it possible to efficiently compute various conditional probabilities that are relevant for posterior inference. Specifically, we will need to evaluate

$$p(z_{ik}|z_{(-i)k}, p_k) \tag{4.6}$$

for $z_{ik} \in \{0, 1\}$ and

$$p\left(\{z_{ik}\}_{i=1}^{N}|p_k\right), \tag{4.7}$$

which are trivial in the gamma-Poisson prior due to the conditional independence of

z_{ik}, but more challenging in the tree-based GP where z_{ik} are no longer conditionally independent. To compute Equation (4.6), we use the sum-product algorithm (Pearl, 1988). In order to calculate Equation (4.7), we use the chain rule of probability to get a set of terms similar to Equation (4.6), the difference being that the posterior in each term is conditional only on a subset of the other variables. Each term can be reduced to a simple sum-product calculation by marginalizing over all variables that do not appear in that term, which can be done easily since all variables appear at the leaves of the tree. Both Equation (4.6) and Equation (4.7) can be calculated in $O(N)$ time by a dynamic program.

We can also compute the posterior of B given observations $z_1, \ldots, z_N$. Recall Equation (2.15) for the posterior of B in the case of the exchangeable GP.

$$\begin{aligned} B|z_1, \ldots, z_N &\sim \text{GP}(c+N, B_N) \\ \text{where } B_N &= \frac{c}{c+N} B_0 + \frac{1}{c+N} \sum_{j=1}^{N} z_j. \end{aligned}$$

We will not have as simple a posterior, but we can still reason about the posterior efficiently. We will reason about the continuous and discrete parts of the posterior of B independently, using ideas from Wolpert and Ickstadt (1998a). First note that if $\sum t$ is the sum of all edge lengths in the tree, then the probability of all entries corresponding to p_k being zero is $e^{-p_k \sum t}$. This gives us that the continuous part of B has posterior

$$\begin{aligned} B_{\text{continuous}}|z_1, \ldots, z_N &\sim \text{GP}\left(c + \sum t, B_{N,\text{continuous}}\right) \\ \text{where } B_{N,\text{continuous}} &= \frac{c}{c + \sum t} B_0. \end{aligned} \tag{4.8}$$

For the discrete part, the atoms still appear only at locations where we have a non-zero z_{ik}, but there is no closed form that applies to all trees. However, for any particular tree, we can compute the posterior of p_k for all k with a non-zero feature. The likelihood of $\{z_{ik}\}_{i=1}^{N}$ given p_k is discussed above in Equation 4.7 and since there is at least one non-zero entry for the discrete part of B, the posterior when combined with the improper $cp^{-1}e^{-cp}$ prior will be proper.

4.4.2.3 Tree-based IGPFM

Just as we can marginalize out B in the exchangeable GP to directly sample Z via the IGPFM, we can again marginalize out B in the tree-based GP to sample Z, which we will call the *phylogenetic IGPFM* (pIGPFM) to be consistent with the naming of the pIBP. As in the IGPFM in Section 2.4.3, the pIGPFM can be derived either by

working with a marginalized GP or by defining the limit of a finite gamma-Poisson model.

The pIGPFM can again be understood in terms of an incremental generative process for Z in which we generate each row one at a time. We present this process here, leaving derivations for Appendix 4.A.2.

As in the IGPFM, we start off with an all-zero matrix Z, with $c = 1$ and $\alpha = B_0(\Omega)$. We generate all rows of Z as follows:

- In the first row, we first decide the total count of features we will add and then we decide how to split this count up into individual features. We do this by sampling g_1, a Negative Binomial $\mathrm{NB}(\alpha, 1/2)$ number of features and then partition g_1 according to the CRP. The partitions become the new features and the counts in the partitions are the counts entered in the matrix.

  We then sample p_k for each of these new non-zero columns from a $\mathrm{Gamma}(z_{1k}, 2)$ distribution.

- Now assuming we have filled in the first $i-1$ rows, $\sum t$ is the total length of the minimal tree connecting the first $i-1$ nodes, and t_i is the length of the edge connecting i to this tree, we fill in the i^{th} row as follows:

  - First look at all features that are present in $z_1, \ldots, z_{i-1}$. We then sample z_{ik} for each non-zero feature k from Equation (4.6).
  - Now select the total count of features g_i that will be unique to the i^{th} row from

$$g_i \sim \mathrm{NB}\left(\alpha, \frac{\sum t + 1}{\sum t + t_i + 1}\right), \tag{4.9}$$

    and distribute this into unique features according to the CRP.

    We then sample p_k for each of these newly non-zero columns from

$$p_k \sim \mathrm{Gamma}\left(z_{ik}, \sum t + t_i + 1\right). \tag{4.10}$$

This process repeats until all rows have been filled in, defining a matrix Z as in the IGPFM. Though this process is not exchangeable, we can let any node go first and get the same marginal distribution. This means that each row of Z has a $\mathrm{NB}(\alpha, 1/2)$ number of features distributed according to the CRP, yielding a sparse matrix as in the exchangeable IGPFM. The IGPFM is the special case of the pIGPFM corresponding to the flat tree shown in the left part of Figure 4.3. This fact can easily be seen from the above equations if every t_i is one.

4.5 Chain-based Generalizations

Now we move on to our second set of generalizations. Why should we restrict ourselves to independent chains in Figures 4.1 and 4.2? Below we discuss how to generalize both the BP and the GP to the case when objects have a known linear (for example temporal or spatial in one dimension) relationship that can be captured by a chain.

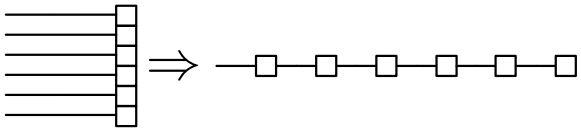


Figure 4.7: Framework for our chain-based generalizations.

By using the conditional independencies in the Markov chain, we are able to define a nonparametric prior in which inference is nearly as efficient as inference developed for the exchangeable special case.

This prior is applicable when objects have a known linear ordering such as those captured by temporal relationships or linear spatial relationships.

4.5.1 Chain-based BP

In this section, we present a non-exchangeable version of the BP in which relationships between objects are expressed via a known, fixed chain.

4.5.1.1 Chain-based BP Stochastic Process

As in the exchangeable BP, we first sample

$$B \;\sim\; \mathrm{BP}(c, B_0).$$

Now instead of sampling $z_i \overset{\text{i.i.d}}{\sim} \mathrm{BeP}(B)$ or equivalently, sampling $z_{ik} \overset{\text{i.i.d}}{\sim} \mathrm{Bernoulli}(p_i)$, for each k, we sample $\{z_{ik}\}_{i=1}^N$ jointly while satisfying all the desiderata from Section 4.2.

We assume we have a known chain such as the one in Figure 4.8(a) expressing the object relationships. These objects do not have to be evenly spaced, so the edge lengths have meaning, with shorter edges indicating a stronger dependence. For each k, we then independently generate each set of features $\{z_{ik}\}_{i=1}^N$ jointly by defining the stochastic process

$$z_{0k} \overset{t_1}{\to} z_{1k} \overset{t_2}{\to} z_{2k} \overset{t_3}{\to} \cdots \overset{t_N}{\to} z_{Nk},$$

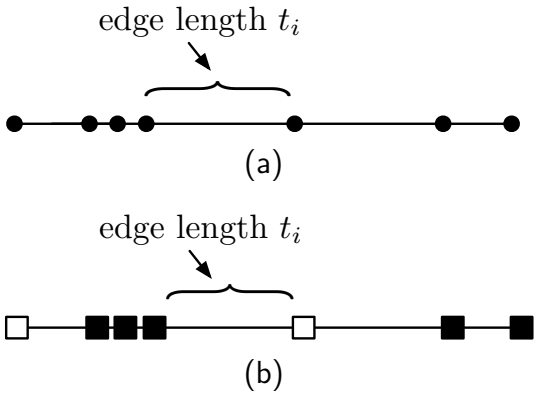


Figure 4.8: Chain-based BP per-feature stochastic process. (a) An example chain representing the known object relationships for our chain-based generalizations. (b) The joint stochastic process for $\{z_{ik}\}_{i=1}^N$ in the chain-based BP generalization in which a black box indicates the corresponding z_{ik} is one and a white box indicates it is zero.

where z_{0k} is a dummy variable deterministically set to zero and t_i is the edge length between object $i-1$ and object i where we set $t_1 = \infty$. We then define a binary-valued continuous time stochastic process such that z_{1k} is the value of the process at time zero, z_{2k} is the value of the process at time t_2, z_{3k} is the value of the process at time $t_2 + t_3$, z_{4k} is the value of the process at time $t_2 + t_3 + t_4$, etc. This is shown in Figure 4.8(b). We introduce a parameter κ which will be related to a continuous-time birth-death process.

Then to generate $\{z_{ik}\}_{i=1}^N$, we define the continuous-time binary-valued stochastic process via a birth-death process. For background on birth-death processes, see Feller (1968) or Cooper (1981). Let this process at time zero be Bernoulli(p_k). Then, at any later point in time, if the stochastic process is in state zero, there will be a birth process of rate κp_k and no death process. If it is in state one, then there will be a death process of rate $\kappa(1-p_k)$ and no birth process.

Equivalently, we show in Appendix 4.A.3.1, that from this continuous time stochastic process, we can define the transition probability from $z_{(i-1)k}$ to z_{ik}. For this stochastic process, define $c_i = 1 - e^{-\kappa t_i}$ so that κ controls the scaling of the time t_i from $[0, \infty]$ to $[0, 1]$. Now let the transition kernel be

		z_{ik}	
		0	1
$z_{(i-1)k}$	0	$1 - c_i p_k$	$c_i p_k$
	1	$c_i(1-p_k)$	$1 - c_i + c_i p_k$

(4.11)

From the transition kernel, we can see that the stationary distribution is Bernoulli(p_k)

and we have made $z_{1k} \sim \text{Bernoulli}(p_k)$, so all entries are marginally Bernoulli(p_k). From the continuous-time process, we know that this process is consistent under marginalization. Finally, from the transition kernel, if $t_i = \infty$, then $c_i = 1$ and $z_{(i-1)k}$ and z_{ik} are conditionally independent, and if $t_i = 0$, then $c_i = 0$ and $z_{(i-1)k} = z_{ik}$. As t_i varies between zero to ∞, we smoothly vary between these two extremes. Therefore in the special case of all $t_i = \infty$, we recover the exchangeable beta-Bernoulli prior which means we have independent chains as in the left part of Figure 4.7. It is a simple corollary that any groups of objects connected by infinite length edges are conditionally independent. From this, we can see that we satisfy all our desiderata.

4.5.1.2 Chain-based BP Conditional Distributions

The conditional distributions are easier for the chain-based BP than for the tree-based BP. We will be able to compute $p(z_{ik}|z_{(-i)k}, p_k)$ for $z_{ik} \in \{0, 1\}$ and $p\left(\{z_{ik}\}_{i=1}^N|p_k\right)$ directly using the Markov structure and Equation (4.11).

We can also compute the posterior of B given observations $z_1, \ldots, z_N$. Recall Equation (2.13) for the posterior of B in the case of the exchangeable BP.

$$\begin{aligned} B|z_1, \ldots, z_N &\sim \text{BP}(c + N, B_N) \\ \text{where } B_N &= \frac{c}{c+N} B_0 + \frac{1}{c+N} \sum_{i=1}^N z_i \end{aligned}$$

We will not have as simple a posterior, but we can still reason about the posterior efficiently. We can again use Theorem 3.3 of Kim (1999) to reason about the continuous and discrete parts of the posterior of B independently. First note that the probability of $\{z_{ik}\}_{i=1}^N$ all being zero is $\prod_{i=1}^N (1 - c_i p_k)$. Therefore, the posterior Lévy measure for the continuous part of B is

$$\nu(d\omega, dp) = cp^{-1}(1-p)^{c-1} \left[\prod_{i=1}^N (1 - c_i p_k)\right] dp B_o(d\omega). \tag{4.12}$$

This itself is not a beta process, but is a weighted sum of improper and proper beta measures.

For the discrete part, the atoms still appear only at locations where we have a non-zero z_{ik}. Since there is at least one non-zero entry for each of the atoms in the discrete part of B, the posterior when combined with the improper $cp^{-1}(1-p)^{c-1}$ prior for each atom will be proper. This posterior can be computed efficiently up to normalizing constant by combining the likelihood $p\left(\{z_{ik}\}_{i=1}^N|p_k\right)$ from Equation (4.11) with the prior.

4.5.1.3 Chain-based IBP

Just as we can marginalize out B in the exchangeable BP to directly sample Z via the IBP, we can again marginalize out B in the chain-based BP to sample Z via a process we will call the *chain Indian Buffet Process* (cIBP). As in the IBP in Section 2.3.3, the cIBP can be derived either by working with a marginalized BP or by defining the limit of a finite beta-Bernoulli model.

The cIBP can again be understood in terms of a sequential generative process. We present this process here, leaving derivations for Appendix 4.A.3. We will present the generative process of Z, skipping the culinary metaphor. We start off with an all-zero matrix Z, with $c = 1$ and $\alpha = B_0(\Omega)$. We generate all rows of Z as follows:

- In the first row, we sample a Poisson(α) number of features, sampling $p_k \sim$ Uniform$[0, 1]$ for each one.
- We fill in the i^{th} row as follows:
  - First look at all features that are present in $z_1, \ldots, z_{i-1}$. We then sample z_{ik} for each non-zero feature k from Equation (4.11).
  - Now sample the number of new features for the i^{th} row from

$$\text{Poisson}\left(\alpha\xi_i\right), \tag{4.13}$$

    where

$$\xi_i \;=\; c_i\left(1 - \frac{\sum_{j=1}^{i-1} c_j}{2} + \frac{\sum_{j<k}^{i-1} c_j c_k}{3} - \frac{\sum_{j<k<l}^{i-1} c_j c_k c_l}{4} + \cdots + (-1)^{i-1}\frac{\prod_{j=1}^{i-1} c_j}{i}\right).$$

    If all c_i are equal so that we have a homogeneous Markov chain, this can be computed in $O(i)$ time. Else, as we prove in Figure 4.11 of Appendix 4.A.3, we can compute ξ_i for all i in $O(N^2)$ time. If c_i take on a few distinct values, this can be greatly improved.

    We then sample p_k for each of these newly non-zero columns from the distribution proportional to

$$c_i \prod_{j=1}^{i-1}(1 - c_j p_k). \tag{4.14}$$

This process repeats until all rows have been filled in, defining a matrix Z as in the IBP. Since this process is consistent under marginalization, each row therefore has a Poisson(α) number of features. We show in Appendix 4.A.3 that in the case

when $c_i = 1$ for all i, then $\xi_i = 1/i$, again verifying that the independent chains are equivalent to the IBP.

4.5.2 Chain-based GP

In this section, we present a non-exchangeable version of the GP in which relationships between objects are expressed via a known, fixed chain.

4.5.2.1 Chain-based GP Stochastic Process

As in the exchangeable GP, we first sample

$$B \sim \mathrm{GP}(c, B_0).$$

Now instead of sampling $z_i \overset{\text{i.i.d}}{\sim} \mathrm{PP}(B)$ or equivalently, sampling $z_{ik} \overset{\text{i.i.d}}{\sim} \mathrm{Poisson}(p_i)$, we sample $\{z_{ik}\}_{i=1}^N$ jointly while satisfying all the desiderata from Section 4.2.

We assume we have a known chain such as the one in Figure 4.9(a) expressing the object relationships. These objects do not have to be evenly spaced, so the edge lengths have meaning, with shorter edges indicating stronger correlation. For each k, we then independently generate each set of features $\{z_{ik}\}_{i=1}^N$ jointly by defining the stochastic process

$$z_{0k} \overset{t_1}{\to} z_{1k} \overset{t_2}{\to} z_{2k} \overset{t_3}{\to} \cdots \overset{t_N}{\to} z_{Nk},$$

where z_{0k} is a dummy variable deterministically set to zero and t_i is the edge length between object $i-1$ and object i where we set $t_1 = \infty$. We then define a non-negative integer-valued continuous time stochastic process such that z_{1k} is the value of the process at time zero, z_{2k} is the value of the process at time t_2, z_{3k} is the value of the process at time $t_2 + t_3$, z_{4k} is the value of the process at time $t_2 + t_3 + t_4$, etc. This is shown in Figures 4.9(b) and 4.9(c). We introduce a parameter κ which will be related to a continuous-time birth-death process.

Then to generate $\{z_{ik}\}_{i=1}^N$, we define the continuous-time non-negative integer-valued stochastic process via a birth-death process. Let this process at time zero be $\mathrm{Poisson}(p_k)$. Then at any point in time, there will be a birth process with rate κp_k and, if there are j objects alive, there will be a death process of rate κj.

As we discuss in Appendix 4.A.4.1, this is equivalent to letting there be a Poisson process on $\mathbb{R}$ with rate κp_k. Each of the events from the Poisson process dies off with exponential rate κ. The value of the birth-death process at time t is the number of events alive at time t. This view of the process is shown in Figure 4.9(b). The red dots are the birth events and the blue function attached to each one is the probability it is still alive at a later point in time.

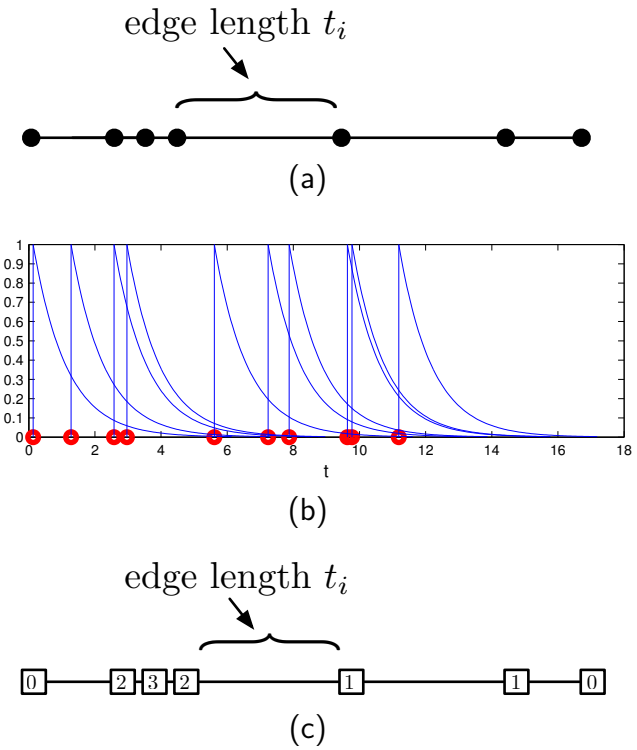


Figure 4.9: Chain-based GP per-feature stochastic process. (a) An example chain representing the known object relationships for our chain-based generalizations. (b) The joint stochastic process for $\{z_{ik}\}_{i=1}^{N}$ in the chain-based GP generalization. These measurements can be interpreted as the number of objects still alive at particular points in time of a continuous-time birth-death process. The x-axis is time and the red dots represent birth events. Each of these events dies with an exponential rate, so the blue function attached to each event is the probability is it still active at any point in time. (c) If z_{ik} is observed at time t, it is just the number of objects alive at that point.

From this view, we can define the transitions from $z_{(i-1)k}$ to z_{ik}. The equivalence between the following process and the birth-death process is shown in Appendix 4.A.4.1. We introduce variables y_{ik} as follows

$$z_{0k} \xrightarrow{t_1} y_{1k} \xrightarrow{t_1} z_{1k} \xrightarrow{t_2} y_{2k} \xrightarrow{t_2} z_{2k} \xrightarrow{t_3} \cdots \xrightarrow{t_N} y_{Nk} \xrightarrow{t_N} z_{Nk}.$$

These are variables that can be marginalized out, but are introduced to simplify the generative description. We again define $c_i = 1 - e^{-\kappa t_i}$ so that κ controls the scaling of the time t_i from $[0, \infty]$ to $[0, 1]$.

The transitions of this Markov chain are based on an alternating birth-death process. Recall that z_{0k} is a dummy variable set to zero and $t_1 = \infty$. Then if $z_{(i-1)k} = K$, each of the K objects (features) dies independently with parameter $c_i \in [0, 1]$ before y_{ik}. y_{ik} is the number of survivors, so

$$p(y_{ik}|z_{(i-1)k}, p_k) \quad \sim \quad \text{Binomial}(z_{(i-1)k}, 1 - c_i).$$

Given y_{ik}, an additional (independent) Poisson$(c_i p_k)$ number of objects are born. z_{ik} is the sum of y_{ik} and the number of new objects, so

$$p(z_{ik} - y_{ik}|y_{ik}, p_k) \quad \sim \quad \text{Poisson}(c_i p_k).$$

Therefore, if we are concerned with just the transition from $z_{(i-1)k}$ to z_{ik},

$$\begin{aligned}
&p(z_{ik}|z_{(i-1)k}, p_k) \\
&= \sum_{y_{ik}=0}^{z_{(i-1)k} \bigwedge z_{ik}} p(z_{ik}|y_{ik}, p_k) p(y_{ik}|z_{(i-1)k}, p_k) \\
&= \sum_{y_{ik}=0}^{z_{(i-1)k} \bigwedge z_{ik}} \text{Poisson}(z_{ik} - y_{ik}; c_i p_k) \text{Binomial}(y_{ik}; z_{(i-1)k}, 1 - c_i) \\
&= \sum_{y_{ik}=0}^{z_{(i-1)k} \bigwedge z_{ik}} \frac{(c_i p_k)^{z_{ik} - y_{ik}} e^{-c_i p_k}}{(z_{ik} - y_{ik})!} \binom{z_{(i-1)k}}{y_{ik}} (1 - c_i)^{y_{ik}} c_i^{z_{(i-1)k} - y_{ik}}. \qquad (4.15)
\end{aligned}$$

Since $t_1 = \infty$, then $c_1 = 1$ which means $y_{1k} = 0$ so therefore z_{1k} is Poisson(p_k). Furthermore, if $z_{(i-1)k}$ is Poisson(p_k), then since y_{ik} is just a thinning of $z_{(i-1)k}$, then y_{ik} is distributed Poisson$((1-c_i)p_k)$. Then since we add an independent Poisson$(c_i p_k)$ number of objects to get z_{ik}, then z_{ik} is Poisson(p_k), so all z_{ik} are marginally Poisson(p_k) for $i \geq 1$. From the continuous-time description, we also know that this process is consistent under marginalization.

If $t_i = 0$, then $z_{ik} = z_{(i-1)k}$. If $t_i = \infty$, then z_{ik} is conditionally independent

of $z_{(i-1)k}$. As t_i varies between zero and ∞, we smoothly vary between these two extremes. This means that in the special case of $t_i = \infty$ for all i, we recover the exchangeable gamma-Poisson prior shown with independent chains as in the left part of Figure 4.7. It is a simple corollary that any group of objects connected by infinite length edges are conditionally independent. From this, we can see that we satisfy all our desiderata.

4.5.2.2 Chain-based GP Conditional Distributions

The conditional distributions are easier for the chain-based GP than for the tree-based GP. We will be able to compute $p(z_{ik}|z_{(-i)k}, p_k)$ for $z_{ik} \in \{0, 1\}$ and $p\left(\{z_{ik}\}_{i=1}^N|p_k\right)$ directly using the Markov structure and Equation (4.15).

We can also compute the posterior of B given observations $z_1, \ldots, z_N$. Recall Equation (2.15) for the posterior of B in the case of the exchangeable GP.

$$\begin{aligned} B|z_1, \ldots, z_N &\sim \mathrm{GP}(c+N, B_N) \\ \text{where } B_N &= \frac{c}{c+N}B_0 + \frac{1}{c+N}\sum_{j=1}^{N} z_j. \end{aligned}$$

We will not have as simple a posterior, but we can still reason about the posterior efficiently. We will reason about the continuous and discrete parts of the posterior of B independently, using ideas from Wolpert and Ickstadt (1998a). First note that the probability of $\{z_{ik}\}_{i=1}^N$ all being zero is $e^{-p_k \sum_{i=1}^N c_i}$. This gives us that the continuous part of B has posterior

$$\begin{aligned} B_{\text{continuous}}|z_1, \ldots, z_N &\sim \mathrm{GP}\left(c + \sum_{i=1}^{N} c_i, B_{N,\text{continuous}}\right) \\ \text{where } B_{N,\text{continuous}} &= \frac{c}{c + \sum_{i=1}^N c_i}B_0. \end{aligned} \tag{4.16}$$

Or in other words, the posterior Lévy measure for the continuous part of B is

$$\nu(d\omega, dp) = c\frac{e^{-\left(c+\sum_{i=1}^N c_i\right)p}}{p}dpB_0(d\omega).$$

For the discrete part, the atoms still appear only at locations where we have a non-zero z_{ik}. Since there is at least one non-zero entry for each of the atoms in the discrete part of B, the posterior when combined with the improper $cp^{-1}e^{-cp}$ prior for each atom will be proper. This posterior can be computed efficiently up to normalizing constant by combining the likelihood $p\left(\{z_{ik}\}_{i=1}^N|p_k\right)$ from Equation (4.15) with the

prior.

4.5.2.3 Chain-based IGPFM

Just as we can marginalize out B in the exchangeable GP to directly sample Z via the IGPFM, we can again marginalize out B in the chain-based GP to sample Z, which we will call the *chain IGPFM* (cIGPFM) to be consistent with the naming of the cIBP. As in the IGPFM in Section 2.4.3, the cIGPFM can be derived either by working with a marginalized GP or by defining the limit of a finite gamma-Poisson model.

The cIGPFM can again be understood in terms of an incremental generative process for Z in which we generate each row one at a time. We present this process here, leaving derivations for Appendix 4.A.4.

As in the IGPFM, we start off with an all-zero matrix Z, with $c = 1$ and $\alpha = B_0(\Omega)$. We generate all rows of Z as follows:

- In the first row, we first decide the total count of features we will add and then we decide how to split this count up into individual features. We do this by sampling g_1, a Negative Binomial $\mathrm{NB}(\alpha, 1/2)$ number of features and then partition g_1 according to the CRP. The partitions become the new features and the counts in the partitions are the counts entered in the matrix.

  We then sample p_k for each of these new non-zero columns from a $\mathrm{Gamma}(z_{1k}, 2)$ distribution.

- Now assuming we have filled in the first $i-1$ rows, we fill in the i^{th} row as follows:

  - First look at all features that are present in $z_1, \ldots, z_{i-1}$. We then sample z_{ik} for each non-zero feature k from Equation (4.15).
  - Now select the total count of features g_i that will be unique to the i^{th} row from

$$g_i \sim \mathrm{NB}\left(\alpha, \frac{\sum_{j=1}^{i-1} c_j + 1}{\sum_{j=1}^{i-1} c_j + c_i + 1}\right), \tag{4.17}$$

    and distribute this into unique features according to the CRP.

    We then sample p_k for each of these newly non-zero columns from

$$p_k \sim \mathrm{Gamma}\left(z_{ik}, \sum_{j=1}^{i-1} c_j + c_i + 1\right). \tag{4.18}$$

This process repeats until all rows have been filled in, defining a matrix Z as in the IGPFM. Though this process is not exchangeable, thanks to the consistency under marginalization, each row of Z has a $\text{NB}(\alpha, 1/2)$ number of features distributed according to the CRP, yielding a sparse matrix as in the exchangeable IGPFM. The IGPFM is the special case of the pIGPFM corresponding to the independent chains shown in the left part of Figure 4.7. This fact can easily be seen from the above equations if every c_i is one.

4.6 Further Power of These Priors

We have now show how to create non-exchangeable generalizations to infer latent features for a single data set. However, we are often presented with multiple structured data sets for which we might wish to simultaneously infer latent features. Using the priors defined in this chapter, it is easy to see that we can directly use them in models to simultaneously infer features shared across multiple data sets in which each individual data set might be related through a tree, chain, or even be exchangeable. We show how to do this in Figure 4.10. In Figure 4.10(a), we show an example of two groups of data in which objects in each group are related through a tree. If we believe that there is a common set of features present in each of the two groups, we can combine the two trees at the root so that conditioned on B, the groups are independent, but share the same set of features. In Figure 4.10(b), we show how to do the same thing with chains in which we connect the two conditionally independent chains with an infinite length edge t_i. An example in which this is very useful is in longitudinal data analysis. Each of these extensions also works with more than two groups of data.

4.7 Summary

In this chapter, we have presented two non-exchangeable generalizations for each of the BP and GP, one based on known relationships of objects captured by a tree and the other by a chain. These are just two of the examples of non-exchangeable, tractable structures that can be used instead of a simple exchangeable prior. We laid out desiderata for our priors, showed how to define them in such a way that they satisfy all desiderata, and derived various properties of them.

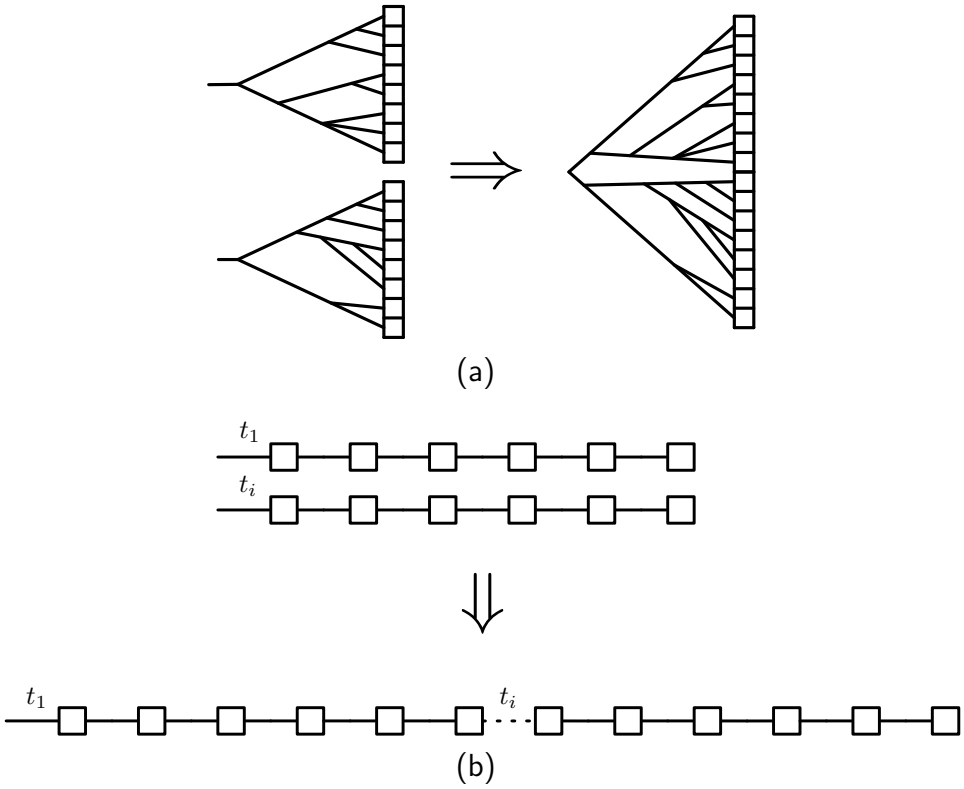


Figure 4.10: Further power of these priors. (a) We can simultaneously infer latent features for two independent sets of objects, each related by a particular tree. By connecting the trees at the root, we can infer features based on the joint tree. (b) We can infer latent features for two independent sets of data where each one has Markov structure. An example of this would be longitudinal data. By connecting the independent chains by an infinite length edge, we can infer features using this joint chain. Each of these extensions can also work with more than two groups of data.

Appendix 4.A Derivations

In this appendix, we discuss the derivations of the generative processes for each non-exchangeable variation as well as a other properties described in Chapter 4. For each prior, we will perform the derivation in two different ways, one directly using the Lévy measure and one using the infinite limit of a finite parametric prior. This shows that for deriving similar priors, we can work with whichever formulation is more convenient or that we are more comfortable or familiar with. Many of the steps of the two different kinds of derivations are similar so once one way has been done, it is not too much effort to switch it to the other kind of derivation. The majority of existing extensions have been derived using infinite limits of finite priors, but as shown here, the Lévy measure formulation, though relying on more sophisticated knowledge of stochastic processes, can often be simpler and more straightforward.

4.A.1 Tree-based BP

We discuss the derivation of the generative process for the tree-based BP discussed in Section 4.4.1.3.

As with the exchangeable BP, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite beta-Bernoulli prior with the stochastic process discussed in Section 4.4.1.1 for each column.

Derivation 1 We now derive how to get the pIBP from the beta-Bernoulli process construction of $p(Z)$. To correspond exactly to the pIBP, we set $c = 1$, but leave c as a variable in the below derivation.

The first person to enter the the restaurant corresponds to z_1. We wish to sample z_1 from the distribution

$$p(z_1) \quad = \quad \int p(z_1|B)dP(B).$$

In this case, the distribution of B is the prior $\text{BP}(c, B_0)$.

Since we have assumed that the total edge length to root for any entry is one, this is equivalent to the first entry in the IBP, so as before, we sample $\text{Poisson}(B_0(\Omega))$ feature.

Unlike in the IBP, in the pIBP, we now draw p_k for the entries sampled by the first customer uniform from $[0, 1]$. This is the posterior of p_k with only a single, non-zero observation as discussed in Section 4.4.1.2.

For the i^{th} person to enter the restaurant, we wish to sample z_i from the posterior

distribution

$$p(z_i|z_1,\ldots,z_{i-1}) \;=\; \int p(z_i|B)dP(B|z_1,\ldots,z_{i-1}).$$

Recalling Section 4.4.1.2, we must reason about the discrete and continuous parts of B independently. We first reason about the discrete part. Since we have sampled p_k, then by the sum product algorithm, we can directly sample z_{ik} from Equation (4.1).

We now reason about the continuous part of B for the i^{th} customer. By Equation (4.3), letting $\sum t$ and t_i be defined as in Section 4.A.1, the posterior Lévy measure of the continuous part of B given the first $i-1$ customers is

$$\nu(d\omega, dp) \;=\; cp^{-1}(1-p)^{\sum t+c-1}dpB_0(d\omega).$$

Now our derivation diverges from the IBP derivation. In the IBP, the probability of observing $z_{ik}=1$ given p_k and that $\{z_{jk}\}_{j=1}^{i-1}$ are all zero is just p_k. In the pIBP, the probability of observing $z_{ik}=1$ given p_k and that $\{z_{jk}\}_{j=1}^{i-1}$ are all zero is the probability that a zero mutates to a one with exponential rate γ_k along an edge of length t_i since we know it could not have mutated anywhere else in the tree. This probability is $\exp(-\gamma_k t_i) = 1-(1-p_k)^{t_i}$.

Therefore, the number of new dishes for the i^{th} customer is Poisson with parameter λ where

$$\lambda \;=\; \int\int_0^1 c\left(1-(1-p)^{t_i}\right)p^{-1}(1-p)^{\sum t+c-1}dpB_0(d\omega).$$

Define $p = 1-e^{-s}$ or equivalently $s = -\log(1-p)$, we get

$$\begin{aligned}
\lambda &= cB_0(\Omega)\int_0^\infty \frac{e^{-(\sum t+c)s}e^s - e^{-(\sum t+t_i+c)s}e^s}{1-e^{-s}}e^{-s}ds \\
&= cB_0(\Omega)\left[\int_0^\infty\left[\frac{e^{-s}}{s}-\frac{e^{-(\sum t+t_i+c)s}}{1-e^{-s}}\right]ds - \int_0^\infty\left[\frac{e^{-s}}{s}-\frac{e^{-(\sum t+c)s}}{1-e^{-s}}\right]ds\right] \\
&= cB_0(\Omega)\left[\psi\left(\sum t+t_i+c\right)-\psi\left(\sum_t +c\right)\right].
\end{aligned}$$

Letting $c=1$ and remembering $\alpha = B_0(\Omega)$, we see that the i^{th} customer must draw a

$$\text{Poisson}\left(\alpha\left(\psi\left(\sum t+t_i+1\right)-\psi\left(\sum t+1\right)\right)\right)$$

number of new dishes, as in Equation (4.4).

Finally, we draw p_k for the new dishes, which is drawing p_k from the posterior

distribution given that a zero mutated to a one along an edge of length t_i, but did not anywhere else on the tree of total edge length $\sum t$, which is proportional to

$$\left(1-(1-p_k)^{t_i}\right)(1-p_k)^{\sum t} p_k^{-1}$$

as in Equation (4.5).

Derivation 2 The alternate derivation of the pIBP is similar to the alternate derivation of the IBP. We start by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. As with the IBP derivation, we start by sampling π_k which will play the role of p_k in the infinite limit from

$$\pi_k \sim \text{Beta}(\alpha/K, 1) \qquad k \in \{1, \ldots, K\}$$

but now, the z_{ik} are drawn via the stochastic process from Section 4.4.1.1.

Instead of computing the closed form distribution of Z, we will compute how to sample the rows of Z incrementally while marginalizing out all π_k for any all-zero columns. For the first row, since we have marginalized out all π_k, then we sample each of the K entries with probability α/K. Remembering that

$$\text{Binomial}\left(K, \frac{\alpha}{K}\right) \overset{K\to\infty}{\to} \text{Poisson}(\alpha),$$

this shows that the first row will have a Poisson(α) number of non-zero entries. The posterior of π_k for these entries is Beta $\left(\frac{\alpha}{K}+1, 1\right)$ which goes to a Beta$(1,1)$, i.e. a Uniform$[0,1]$ distribution.

For all subsequent rows, as $K \to \infty$, we can sample z_{ik} for non-zero columns exactly from Equation (4.1).

For the all-zero columns of subsequent rows, we will show that for some constant ξ_i depending on the row,

$$p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) = \frac{\alpha\xi_i}{K} + o\left(\frac{1}{K}\right),$$

where $f = o(g)$ means $f/g \to 0$. Let K^+ be the number of already non-zero rows. The using the fact that as $K \to \infty$, K^+ will be finite and for any fixed, finite K^+, the number of newly sampled non-zero rows is

$$\text{Binomial}\left(K - K^+, \frac{\alpha\xi_i}{K}\right) \overset{K\to\infty}{\to} \text{Poisson}(\alpha\xi_i).$$

We therefore wish to find ξ_i such that $p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) = \alpha\xi_i/K + o(1/K)$.

For finite K, we can evaluate the probability that $z_{ik} = 1$ given that $z_{(-i)k} = 0$ in the pIBP. If t_i is the length of the edge that ends at the i^{th} object and $\sum t$ is the total length of all other edges in the tree, then we get

$$\begin{aligned} p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) &\propto p(z_{ik} = 1, z_{(-i)k} = 0 | \alpha) \\ &= \int_0^1 p(z_{ik} = 1 | z_{(-i)k} = 0, \pi_k) p(z_{(-i)k} = 0 | \pi_k) p(\pi_k | \alpha) d\pi_k \\ &= \frac{\alpha}{K} \left(\frac{\Gamma(\alpha/K)\,\Gamma(\sum t + 1)}{\Gamma(\sum t + \alpha/K + 1)} - \frac{\Gamma(\alpha/K)\,\Gamma(\sum t + t_i + 1)}{\Gamma(\sum t + t_i + \alpha/K + 1)} \right), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function and similarly

$$\begin{aligned} p(z_{ik} = 0 | z_{(-i)k} = 0, \alpha) &\propto p(z_{ik} = 0, z_{(-i)k} = 0 | \alpha) \\ &= \frac{\alpha}{K} \left(\frac{\Gamma(\alpha/K)\,\Gamma(\sum t + t_i + 1)}{\Gamma(\sum t + t_i + \alpha/K + 1)} \right), \end{aligned}$$

which combined means

$$p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) = 1 - \frac{\Gamma(\sum t + \alpha/K + 1)}{\Gamma(\sum t + 1)} \frac{\Gamma(\sum t + t_i + 1)}{\Gamma(\sum t + t_i + \alpha/K + 1)}.$$

Treating the value α/K as a variable in the equation $f(\alpha/K) = p(z_{ik} = 1 | z_{(-i)k} = 0)$ and doing a first-order Taylor expansion about zero, we get

$$\begin{aligned} p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) &= f\left(\frac{\alpha}{K}\right) \\ &= f(0) + \frac{\alpha}{K} f'(0) + o\left(\frac{1}{K}\right) \\ &= 0 + \frac{\alpha}{K} \left(\psi\left(\sum t + t_i + 1\right) - \psi\left(\sum t + 1\right) \right) + o\left(\frac{1}{K}\right) \end{aligned}$$

where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.

This gives us that

$$\xi_i = \psi\left(\sum t + t_i + 1\right) - \psi\left(\sum t + 1\right).$$

Therefore the number of new non-zero columns for row i is distributed Poisson$(\alpha\xi_i)$ under the prior, as in Equation (4.4).

Finally, we sample π_k for the newly non-zero rows. This is drawn for a distribution

proportional to

$$\left(1-(1-p_k)^{t_i}\right)(1-p_k)^{\sum t}\, p_k^{\frac{\alpha}{K}-1} \quad \overset{K\to\infty}{\to} \quad \left(1-(1-p_k)^{t_i}\right)(1-p_k)^{\sum t}\, p_k^{-1}$$

as desired.

4.A.2 Tree-based GP

We discuss the derivation of the generative process for the tree-based GP discussed in Section 4.4.2.3.

As with the exchangeable GP, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite gamma-Poisson prior with the stochastic process discussed in Section 4.4.2.1 for each column.

Derivation 1 We now derive how to get the pIGPFM from the gamma-Poisson process construction of $p(Z)$. To correspond exactly to the pIGPFM, we set $c = 1$, but leave c as a variable in the below derivation.

For the first row z_1, we wish to sample from the distribution

$$p(z_1) \;=\; \int p(z_1|B)dP(B).$$

In this case, the distribution of B is the prior $\mathrm{GP}(c, B_0)$.

Since we have assumed that the total edge length to root for any entry is one, this is equivalent to the first entry in the IGPFM, so as before, we sample a $\mathrm{NB}(\alpha, 1/2)$ number of points distributed according to the CRP.

Unlike in the IGPFM, in the pIGPFM, we now draw p_k for the newly non-zero entries. This is the posterior of p_k with only a single, non-zero observation at z_{1k} as discussed in Section 4.4.2.2. This posterior is proportional to the Poisson likelihood times the improper gamma prior, which is therefore $\mathrm{Gamma}(z_{1k}, 2)$, as desired.

For the i^{th} row, we wish to sample z_i from the posterior distribution

$$p(z_i|z_1,\ldots,z_{i-1}) \;=\; \int p(z_i|B,z_1,\ldots,z_{i-1})dP(B|z_1,\ldots,z_{i-1}).$$

Recalling Section 4.4.2.2, we must reason about the discrete and continuous parts of B independently. We first reason about the discrete part. Since we have sampled p_k, then by the sum product algorithm, we can directly sample z_{ik} from Equation (4.6).

We now reason about the continuous part of B for the i^{th} row. By Equation (4.8), letting $\sum t$ and t_i be defined as in Section 4.A.2, the posterior Lévy measure of the

continuous part of B given the first $i-1$ rows is

$$\nu(d\omega, dp) \;=\; cp^{-1}e^{-(\sum t+c)p}dpB_0(d\omega).$$

Similar to the IGPFM, using Equation (2.14), we get that

$$B(\Omega)|z_1,\ldots,z_{i-1} \sim \text{Gamma}\left(cB_0(\Omega), \sum t + c\right).$$

Now using the fact that if $x|\lambda \sim \text{Poisson}(t_i\lambda)$ and $\lambda \sim \text{Gamma}(a,b)$, then by marginalizing out λ, $x \sim \text{NB}\left(a, \frac{b}{b+t_i}\right)$, we get that the total count of unobserved features g_i in z_i is $\text{NB}\Big(cB_0(\Omega), \frac{\sum t+c}{\sum t+t_i+c}\Big)$, which substituting in $c=1$ and $\alpha = B_0(\Omega)$ gives us $\text{NB}\Big(\alpha, \frac{\sum t+1}{\sum t+t_i+1}\Big)$ as in Equation (4.9). We must now figure out how to distribute these g_i new features. The posterior of B is still a GP in this case and $t_i \leq 1$, so we can interpret t_i as a thinning parameter for a full Poisson process and we know that the allocation of all features for the full process is distributed according to the CRP. Thinning the CRP still gives a CRP just with fewer features, so the g_i are allocated according to the CRP.

Finally, we draw p_k for the new features, which is drawing p_k from the posterior distribution given that the only events on the tree for feature k occurred along an edge of length t_i with no events anywhere else on the tree of total edge length $\sum t$, which is

$$\text{Gamma}\left(z_{ik}, \sum t + t_i + 1\right)$$

as in Equation (4.10).

Derivation 2 The alternate derivation of the pIGPFM is similar to the alternate derivation of the IGPFM. We start by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. As with the IGPFM derivation, we start by sampling λ_k which will play the role of p_k in the infinite limit from

$$\lambda_k \;\sim\; \text{Gamma}\left(\frac{\alpha}{K}, 1\right) \qquad k \in \{1,\ldots,K\}$$

but now, the z_{ik} are drawn via the stochastic process from Section 4.4.2.1.

Instead of computing the closed form distribution of Z, we will compute how to sample the rows of Z incrementally while marginalizing out all λ_k for any all-zero columns. For the first row, since we have marginalized out all λ_k, then the distribution of each z_{1k} is $\text{NB}\big(\frac{\alpha}{K}, \frac{1}{2}\big)$. Since the z_{1k} are independent, then since the

sum of independent $\mathrm{NB}(a_i, b)$ variables is $\mathrm{NB}(\sum a_i, b)$, we have that

$$p\left(\sum_k z_{1k} \middle| \alpha\right) \sim \mathrm{NB}\left(\alpha, \frac{1}{2}\right).$$

We now must compute how to partition these features. This will take several steps. We will only be concerned with unique partitions of the data, but not the order that groups of features come in, so if we let b_h be the number of z_{1k} that take on value h, then we can see that there are $K!/\prod_{h=0}^{\infty} b_h!$ elements of the equivalence class of $z_{11}, \ldots, z_{1K}$, where by definition $b_0 = K - K^+$.

$$\begin{aligned}
&p(z_{11}, \ldots, z_{1K}|\alpha) \\
&= \frac{K!}{\prod_{h=0}^{\infty} b_h!} \prod_{k=1}^{K} \mathrm{NB}\left(z_{1k}; \frac{\alpha}{K}, \frac{1}{2}\right) \\
&= \frac{K!}{\prod_{h=0}^{\infty} b_h!} \prod_{k=1}^{K} \frac{\Gamma(\frac{\alpha}{K} + z_{1k})}{\Gamma(\frac{\alpha}{K}) z_{1k}!} \left(\frac{1}{2}\right)^{\frac{\alpha}{K}} \left(1 - \frac{1}{2}\right)^{z_{1k}} \\
&= \frac{K!}{\prod_{h=0}^{\infty} b_h!} \left(\frac{\alpha}{K}\right)^{K^+} \left(\prod_{k:z_{1k}>1} \frac{\prod_{j=1}^{z_{1k}-1}(j + \frac{\alpha}{K})}{z_{1k}!}\right) \left(\frac{1}{2}\right)^{\alpha} \left(1 - \frac{1}{2}\right)^{\sum_{k=1}^{K} z_{1k}}.
\end{aligned}$$

Using similar limits to Griffiths and Ghahramani (2006), the probability of the partition having values $z_{11}, \ldots, z_{1K}$ is therefore

$$\begin{aligned}
&p(z_{11}, \ldots, z_{1K}|\alpha) \\
&= \frac{\alpha^{K^+}}{\prod_{h=1}^{\infty} b_h!} \frac{K!}{b_0! K^{K^+}} \left(\prod_{k:z_{1k}>1} \frac{\prod_{j=1}^{z_{1k}-1}(j + \frac{\alpha}{K})}{z_{1k}!}\right) \left(\frac{1}{2}\right)^{\alpha} \left(1 - \frac{1}{2}\right)^{\sum_{k:z_{1k}>1} z_{1k}} \\
&\overset{K\to\infty}{\to} \frac{\alpha^{K^+}}{\prod_{h=1}^{\infty} b_h!} \times 1 \times \left(\prod_{k:z_{1k}>1} \frac{(z_{1k}-1)!}{z_{1k}!}\right) \left(\frac{1}{2}\right)^{\alpha} \left(1 - \frac{1}{2}\right)^{\sum_{k:z_{1k}>1} z_{1k}} \\
&= \frac{\alpha^{K^+}}{\prod_{h=1}^{\infty} b_h!} \frac{1}{\prod_{k:z_{1k}>1} z_{1k}} \left(\frac{1}{2}\right)^{\alpha} \left(1 - \frac{1}{2}\right)^{\sum_{k:z_{1k}>1} z_{1k}}.
\end{aligned}$$

We now note that if we sample $g_1 = \sum_k z_{1k}$, the total number of non-zero features, from a $\mathrm{NB}\big(\alpha, \frac{1}{2}\big)$ distribution and then partition it into K^+ groups according to Ewens

distribution (Ewens, 1972), we get that

$$\begin{aligned} & p(z_{1k}, \ldots, z_{1K}|\alpha) \\ &= \underbrace{\frac{\Gamma(\alpha+g_1)}{\Gamma(\alpha)g_1!}\left(\frac{1}{2}\right)^{\alpha}\left(1-\frac{1}{2}\right)^{g_1}}_{\text{Negative Binomial}} \times \underbrace{\frac{\alpha^{K^+}}{\prod_{k:z_{1k}>1} z_{1k}} \frac{\Gamma(\alpha)g_1!}{\Gamma(\alpha+g_1)} \prod_{h=1}^{\infty} \frac{1}{b_h!}}_{\text{Ewens}} \\ &= \frac{\alpha^{K^+}}{\prod_{h=1}^{\infty} b_h! \prod_{k:z_{1k}>1} z_{1k}} \left(\frac{1}{2}\right)^{\alpha}\left(1-\frac{1}{2}\right)^{\sum_{k:z_{1k}>0} z_{1k}}. \end{aligned} \tag{4.19}$$

This shows that the proper way to partition g_1 to get the new columns is to partition it according to Ewens distribution.

It is also straightforward to compute the posterior distribution of λ_k (which play the role of p_k) for non-zero columns. Since this is a conjugate prior, this posterior distribution is

$$\text{Gamma}\left(z_{ik} + \frac{\alpha}{K}, 2\right) \overset{K\to\infty}{\to} \text{Gamma}(z_{ik}, 2)$$

as desired.

For all subsequent rows, as $K \to \infty$, we can sample z_{ik} for non-zero columns exactly from Equation (4.6) using sum-product.

For the all-zero columns of row i, we will show that for some constant ξ_i,

$$p(z_{ik}|z_{(-i)k} = 0, \alpha) \sim \text{NB}\left(\frac{\alpha}{K}, \xi_i\right).$$

From this, if we assume the first K^+ columns are the non-zero columns so that columns $K^+ + 1$ through K are all-zero, then

$$\begin{aligned} p\left(\sum_{k=K^+ +1}^{K} z_{ik} \,\middle|\, z_{(-i)k} = 0 \;\forall k \in \{K^+ + 1, \ldots, K\}, \alpha\right) &\sim \text{NB}\left(\frac{\alpha(K-K^+)}{K}, \xi_i\right) \\ &\overset{K\to\infty}{\to} \text{NB}\left(\alpha, \xi_i\right). \end{aligned}$$

So the sum of all the non-zero z_{ik} for previously all-zero columns follows a negative binomial distribution. As we did for the first row, we'll need to figure out how to partition these elements. Using exactly the same argument as we did for the first

row, if we let K_i^{new} be the number of new non-zero columns in the i^{th} row, then

$$\begin{aligned}
&p(z_{i(K^+{+}1)}, \ldots, z_{iK} | z_{(-i)k} = 0 \ \forall k \in \{K^+ + 1, \ldots, K\}, \alpha) \\
&= \frac{\alpha^{K_i^{\text{new}}}}{\prod_{h=1}^{\infty} b_h!} \frac{(K - K^+)!}{b_0! K^{K_i^{\text{new}}}} \left(\prod_{k: z_{ik} > 0} \frac{\prod_{j=1}^{z_{ik}-1} (j + \frac{\alpha}{K})}{z_{ik}!} \right) \xi_i^{\alpha(K-K^+)/K} (1 - \xi_i)^{\sum_{k=K^+{+}1}^{K^+ + K_i^{\text{new}}} z_{ik}} \\
&\overset{K \to \infty}{\to} \frac{\alpha^{K_i^{\text{new}}}}{\prod_{h=1}^{\infty} b_h!} \times 1 \times \left(\prod_{k: z_{ik} > 0} \frac{(z_{ik} - 1)!}{z_{ik}!} \right) \xi_i^{\alpha} (1 - \xi_i)^{\sum_{k=K^+{+}1}^{K^+ + K_i^{\text{new}}} z_{ik}} \\
&= \frac{\alpha^{K_i^{\text{new}}}}{\prod_{h=1}^{\infty} b_h!} \frac{1}{\prod_{k: z_{ik} > 0} z_{ik}} \xi_i^{\alpha} (1 - \xi_i)^{\sum_{k=K^+{+}1}^{K^+ + K_i^{\text{new}}} z_{ik}}.
\end{aligned}$$

As before, we note that if we sample $g_i = \sum_{k=K^+{+}1}^{K^+ + K_i^{\text{new}}} z_{ik}$ from a $\text{NB}(\alpha, \xi_i)$ distribution and then partition it according to Ewens distribution (Ewens, 1972) into K_i^{new} groups, we get that

$$\begin{aligned}
&p(z_{i(K^+{+}1)}, \ldots, z_{iK} | z_{(-i)k} = 0 \ \forall k \in \{K^+ + 1, \ldots, K\}, \alpha) \\
&= \underbrace{\frac{\Gamma(\alpha + g_i)}{\Gamma(\alpha) g_i!} \xi_i^{\alpha} (1 - \xi_i)^{g_i}}_{\text{Negative Binomial}} \times \underbrace{\frac{\alpha^{K_i^{\text{new}}}}{\prod_{k: z_{ik} > 0} z_{ik}} \frac{\Gamma(\alpha) g_i!}{\Gamma(\alpha + g_i)} \prod_{h=1}^{\infty} \frac{1}{b_h!}}_{\text{Ewens}} \\
&= \frac{\alpha^{K_i^{\text{new}}}}{\prod_{h=1}^{\infty} b_h!} \frac{1}{\prod_{k: z_{ik} > 0} z_{ik}} \xi_i^{\alpha} (1 - \xi_i)^{\sum_{k=K^+{+}1}^{K^+ + K_i^{\text{new}}} z_{ik}}. \qquad (4.20)
\end{aligned}$$

This shows that the proper way to partition g_i to get the new columns is to partition it according to Ewens distribution.

We must now identify ξ_i and we will be done with our derivation. In this case, we can see that

$$\xi_i = \frac{\sum t + 1}{\sum t + t_i + 1}, \qquad (4.21)$$

which therefore means we sample $g_i \sim \text{NB}\left(\alpha, \frac{\sum t+1}{\sum t+t_i+1}\right)$ and partition it according to the CRP, agreeing with Equation (4.9).

Finally, for each new non-zero column, we sample λ_k from the posterior distribution which is

$$\text{Gamma}\left(z_{ik} + \frac{\alpha}{K}, \sum t + t_i + 1\right) \overset{K \to \infty}{\to} \text{Gamma}\left(z_{ik}, \sum t + t_i + 1\right)$$

agreeing with Equation (4.10).

4.A.3 Chain-based BP

Here we show four sets of derivations. The first is the equivalence between the continuous-time stochastic process and the transition kernel defined in Section 4.5.1.1. The second is the marginal distribution of the cIBP in Section 4.5.1.3. The third is the dynamic program to efficiently compute ξ_i in Equation (4.13). The last shows that Equation (4.13) is equivalent to the IBP when all $c_i = 1$.

4.A.3.1 Chain-based BP Stochastic Process

Here we show the equivalence between the continuous time stochastic process and the transition kernel defined in Section 4.5.1.1.

First we start with the transition kernel defined by Equation (4.11)

		z_{ik}	
		0	1
$z_{(i-1)k}$	0	$1 - c_i p_k$	$c_i p_k$
	1	$c_i(1 - p_k)$	$1 - c_i + c_i p_k$

where $c_i = 1 - e^{-\kappa t_i}$. Since we have defined z_{0k} to be a dummy variable deterministically set to zero and $t_1 = \infty$, we have that $z_{1k} \sim \text{Bernoulli}(p_k)$, so at time zero, these two processes are the same.

We note that for any transition matrix

$$\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

the stationary distribution is $\frac{a}{a+b}$, so the stationary distribution is Bernoulli(p_k), so all z_{ik} are Bernoulli(p_k).

We can also show that this is consistent under marginalization. That is, if we do not observe z_{ik}, then the conditional distribution of $z_{(i+1)k}$ given $z_{(i-1)k}$ is the same if we use directly use Equation (4.11) with edge length $t_i + t_{i+1}$ between $z_{(i-1)k}$ and $z_{(i+1)k}$ and if we marginalize out z_{ik}. In equations, if we define $c_j = 1 - e^{-\kappa(t_i + t_{i+1})}$, this just means that

$$\begin{pmatrix} 1 - c_j p_k & c_j p_k \\ c_j(1-p_k) & 1 - c_j + c_j p_k \end{pmatrix} = \begin{pmatrix} 1 - c_i p_k & c_i p_k \\ c_i(1-p_k) & 1 - c_i + c_i p_k \end{pmatrix} \begin{pmatrix} 1 - c_{i+1} p_k & c_{i+1} p_k \\ c_{i+1}(1-p_k) & 1 - c_{i+1} + c_{i+1} p_k \end{pmatrix}$$

So now we know that we have a stationary process consistent under marginalization, which is a good indication that there is an underlying continuous-time binary

valued stochastic process. Since it is consistent under marginalization, we can look at the transition kernel if we were to shrink t_i (remember that we assume known, fixed t_i, so this is a thought exercise for the sake of derivation). Then as $t_i \downarrow 0$, using the Taylor expansion of e^{-x} and the definition $c_i = 1 - e^{-\kappa t_i}$, the transition kernel becomes

$$\begin{pmatrix} 1 - c_i p_k & c_i p_k \\ c_i(1 - p_k) & 1 - c_i + c_i p_k \end{pmatrix} = \begin{pmatrix} 1 - \kappa p_k t_i + o(t_i) & \kappa p_k t_i + o(t_i) \\ \kappa(1 - p_k)t_i + o(t_i) & 1 - \kappa(1 - p_k)t_i + o(t_i) \end{pmatrix}$$

This means that in state zero, there is a birth process of rate κp_k and no death process, and in state one, there is a death process of rate $\kappa(1 - p_k)$ and no birth process. From this, we see that this therefore defines the continuous-time birth-death process from Section 4.5.1.1.

Directly from the definition of this process, we can also verify that the stationary distribution is Bernoulli(p_k) by checking the fixed point of the birth-death rates. Let $z_k(t)$ be the value of the continuous-time stochastic process at time t. Then following notation from (Cooper, 1981), define λ_i as the birth rate when $z_k(t) = i$, μ_0 as the death rate, and $p_i(t)$ is the probability $z_k(t) = i$. Then we can verify that constant probabilities $p_0(t) = 1 - p_k$ and $p_1(t) = p_k$ satisfy

$$0 = \frac{d}{dt} p_i(t) \quad = \quad \lambda_{i-1} p_{i-1}(t) + \mu_{i+1} p_{i+1}(t) - (\lambda_i + \mu_i) p_i(t)$$

for $i = 0$ and $i = 1$. Since $z_{k0} \sim \text{Bernoulli}(p_k)$, the stochastic process is already at its stationary distribution and we can see that our two definitions are equivalent.

4.A.3.2 Chain-based BP Derivation

We discuss the derivation of the generative process for the chain-based BP discussed in Section 4.5.1.3.

As with the exchangeable BP, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite beta-Bernoulli prior with the stochastic process discussed in Section 4.5.1.1 for each column.

Derivation 1 We now derive how to get the cIBP from the beta-Bernoulli process construction of $p(Z)$. To correspond exactly to the cIBP, we set $c = 1$, but leave c as a variable in the below derivation.

For the first row z_1, we wish to sample z_1 from the distribution

$$p(z_1) \quad = \quad \int p(z_1|B) dP(B).$$

In this case, the distribution of B is the prior $\text{BP}(c, B_0)$ and which is equivalent to the IBP, so we sample $\text{Poisson}(B_0(\Omega))$ features.

Unlike in the IBP, in the cIBP, we now draw p_k for the features sampled in the first row uniform from $[0, 1]$. This is the posterior of p_k with only a single, non-zero observation as discussed in Section 4.5.1.2.

For the i^{th} row, we wish to sample z_i from the posterior distribution

$$p(z_i|z_1, \ldots, z_{i-1}) \quad = \quad \int p(z_i|B, z_1, \ldots, z_{i-1}) dP(B|z_1, \ldots, z_{i-1}).$$

Recalling Section 4.5.1.2, we must reason about the discrete and continuous parts of B independently. We first reason about the discrete part. Since we have sampled p_k, then we can directly sample z_{ik} from Equation (4.11).

We now reason about the continuous part of B for the i^{th} row. By Equation (4.12), the posterior Lévy measure of the continuous part of B given the first $i-1$ rows is

$$\nu(d\omega, dp) \quad = \quad cp^{-1}(1-p)^{c-1} \left[\prod_{j=1}^{i-1}(1 - c_j p)\right] dp B_0(d\omega).$$

Now our derivation diverges from the IBP derivation. In the IBP, the probability of observing $z_{ik} = 1$ given p_k and that $\{z_{jk}\}_{j=1}^{i-1}$ are all zero is just p_k. In the cIBP, the probability of observing $z_{ik} = 1$ given p_k and that $\{z_{jk}\}_{j=1}^{i-1}$ are all zero is $c_i p_k$ as in Equation (4.11).

Therefore, the number of new features for the i^{th} row is Poisson with parameter λ where

$$\begin{aligned} \lambda \quad &= \quad \int \int_0^1 c_i p c p^{-1}(1-p)^{c-1} \left[\prod_{j=1}^{i-1}(1 - c_j p)\right] dp B_0(d\omega) \\ &= \quad c_i B_0(\Omega) \int_0^1 c(1-p)^{c-1} \left[\prod_{j=1}^{i-1}(1 - c_j p)\right] dp. \end{aligned}$$

Now we note that by expanding $\left[\prod_{j=1}^{i-1}(1 - c_j p)\right]$, we get a weighted sum of proper

beta distributions, so we can evaluate this integral.

$$\begin{aligned}
&\int_0^1 c(1-p)^{c-1}\left[\prod_{j=1}^{i-1}(1-c_j p)\right] dp \\
&= \int_0^1 c(1-p)^{c-1}\left[1 - p\sum_{j=1}^{i-1} c_j + p^2\sum_{j<k}^{i-1} c_j c_k - \cdots + (-1)^{i-1}p^{i-1}\prod_{j=1}^{i-1} c_j\right] dp \\
&= c\left[\frac{\Gamma(1)\Gamma(c)}{\Gamma(1+c)} - \frac{\Gamma(2)\Gamma(c)}{\Gamma(2+c)}\sum_{j=1}^{i-1} c_j + \frac{\Gamma(3)\Gamma(c)}{\Gamma(3+c)}\sum_{j<k}^{i-1} c_j c_k - \cdots + \frac{\Gamma(i)\Gamma(c)}{\Gamma(i+c)}(-1)^{i-1}\prod_{j=1}^{i-1} c_j\right].
\end{aligned}$$

Substituting $c = 1$, this simplifies to

$$\int_0^1 \left[\prod_{j=1}^{i-1}(1-c_j p)\right] dp = 1 - \frac{1}{2}\sum_{j=1}^{i-1} c_j + \frac{1}{3}\sum_{j<k}^{i-1} c_j c_k - \cdots + \frac{1}{i}(-1)^{i-1}\prod_{j=1}^{i-1} c_j.$$

Therefore

$$\lambda = \alpha c_i\left(1 - \frac{1}{2}\sum_{j=1}^{i-1} c_j + \frac{1}{3}\sum_{j<k}^{i-1} c_j c_k - \cdots + \frac{1}{i}(-1)^{i-1}\prod_{j=1}^{i-1} c_j\right),$$

which agrees with Equation (4.13) in which

$$\alpha\xi_i = \alpha c_i\left(1 - \frac{\sum_{j=1}^{i-1} c_j}{2} + \frac{\sum_{j<k}^{i-1} c_j c_k}{3} - \frac{\sum_{j<k<l}^{i-1} c_j c_k c_l}{4} + \cdots + (-1)^{i-1}\frac{\prod_{j=1}^{i-1} c_j}{i}\right).$$

So for the i^{th} row, we must draw a Poisson (λ) number of new features.

Finally, we draw p_k for the new features, which is drawing p_k from the posterior distribution given that only the i^{th} entry is non-zero, which is proportional to $c_i\left[\prod_{j=1}^{i-1}(1-c_j p_k)\right]$ as in Equation (4.14).

Derivation 2 The alternate derivation of the cIBP is similar to the alternate derivation of the IBP. We start by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. As with the IBP derivation, we start by sampling π_k which will play the role of p_k in the infinite limit from

$$\pi_k \sim \text{Beta}(\alpha/K, 1) \qquad k \in \{1, \ldots, K\}$$

but now, the z_{ik} are drawn via the stochastic process from Section 4.5.1.1.

Now since we are working with a proper prior distribution, computing all quantities is relatively straightforward. As for the pIBP, to figure out the number of new features sampled for both the first row as well as all subsequent rows i, we must show

$$p(z_{ik} = 1 | \{z_{jk}\}_{j=1}^{i} = 0, \alpha) \;\; = \;\; \frac{\alpha \xi_i}{K} + o\left(\frac{1}{K}\right),$$

where $f = o(g)$ means $f/g \to 0$. Then using the fact that

$$\text{Binomial}\left(K - K^+, \frac{\alpha \xi_i}{K}\right) \overset{K \to \infty}{\to} \text{Poisson}(\alpha \xi_i),$$

we can derive the desired Poisson distributions.

This approach is doable, but rather tedious. For example, we must show

$$\begin{aligned}
& p\left(z_{ik} = 1 | \{z_{jk}\}_{j=1}^{i} = 0, \alpha\right) \\
& \quad \propto \quad \int_0^1 p(z_{ik} = 1 | z_{(i-1)k} = 0, \pi_k) p(\pi_k | \alpha) \prod_{j=1}^{i-1} p(z_{jk} = 0 | z_{(j-1)k} = 0, \pi_k) d\pi_k \\
& \quad = \quad \int_0^1 (c_i \pi_k) \frac{\alpha}{K} \pi_k^{\alpha/K - 1} \prod_{j=1}^{i-1} (1 - c_j \pi_k) d\pi_k \\
& \quad = \quad \int_0^1 c_i \frac{\alpha}{K} \pi_k^{\alpha/K} \left(1 - \pi_k \sum_{j=1}^{i-1} c_j + \pi_k^2 \sum_{j<k}^{i-1} c_j c_k - \ldots + (-1)^{i-1} \pi_k^{i-1} \prod_{j=1}^{i-1} c_j\right) d\pi_k \\
& \quad = \quad c_i \frac{\alpha}{K} \left(\frac{1}{\alpha/K + 1} - \frac{1}{\alpha/K + 2} \sum_{j=1}^{i-1} c_j + \frac{1}{\alpha/K + 3} \sum_{j<k}^{i-1} c_j c_k - \ldots + (-1)^i \frac{1}{\alpha/K + i} \prod_{j=1}^{i-1} c_j\right)
\end{aligned}$$

and

$$\begin{aligned}
& p\left(z_{ik}=0|\left\{z_{jk}\right\}_{j=1}^{i}=0,\alpha\right) \\
& \propto \int_0^1 p(\pi_k)\prod_{j=1}^{i}p(z_{jk}=0|z_{(j-1)k}=0,\pi_k)d\pi_k \\
& = \int_0^1 \frac{\alpha}{K}\pi_k^{\alpha/K-1}\prod_{j=1}^{i}(1-c_j\pi_k)d\pi_k \\
& = \int_0^1 \frac{\alpha}{K}\pi_k^{\alpha/K-1}\left(1-\pi_k\sum_{j=1}^{i}c_j+\pi_k^2\sum_{j<l}^{i}c_jc_l-\ldots+(-1)^i\pi_k^i\prod_{j=1}^{i}c_j\right)d\pi_k \\
& = 1-\frac{\alpha/K}{\alpha/K+1}\sum_{j=1}^{i}c_j+\frac{\alpha/K}{\alpha/K+2}\sum_{j<l}^{i}c_jc_l-\ldots+(-1)^i\frac{\alpha/K}{\alpha/K+i}\prod_{j=1}^{i}c_j.
\end{aligned}$$

Normalizing to compute the value of $p\left(z_{ik}=1|\left\{z_{jk}\right\}_{j=1}^{i}=0,\alpha\right)$ and taking the first order Taylor series expansion, we can show that ξ_i must take the desired form.

To compute the posterior distributions of π_k for both the first and all subsequent rows, we take the limit of the resulting posterior distribution to again get the desired distributions.

4.A.3.3 Computation of ξ_i

For each i, we must compute

$$\xi_i = 1-\frac{1}{2}\sum_{j=1}^{i-1}c_j+\frac{1}{3}\sum_{j<k}^{i-1}c_jc_k-\cdots+\frac{1}{i}(-1)^{i-1}\prod_{j=1}^{i-1}c_j$$

If all c_i (except possibly c_1 which is always one) are homogeneous, we can compute each term directly in $O(1)$ time. If not, then a naïve computation of each term would take $O(2^n)$ time which is intractable. However, using a dynamic program, we can compute all terms for all i in $O(N^2)$ time. We show this dynamic program in Figure 4.11. If c_i take on a few distinct values, this can be greatly decreased.

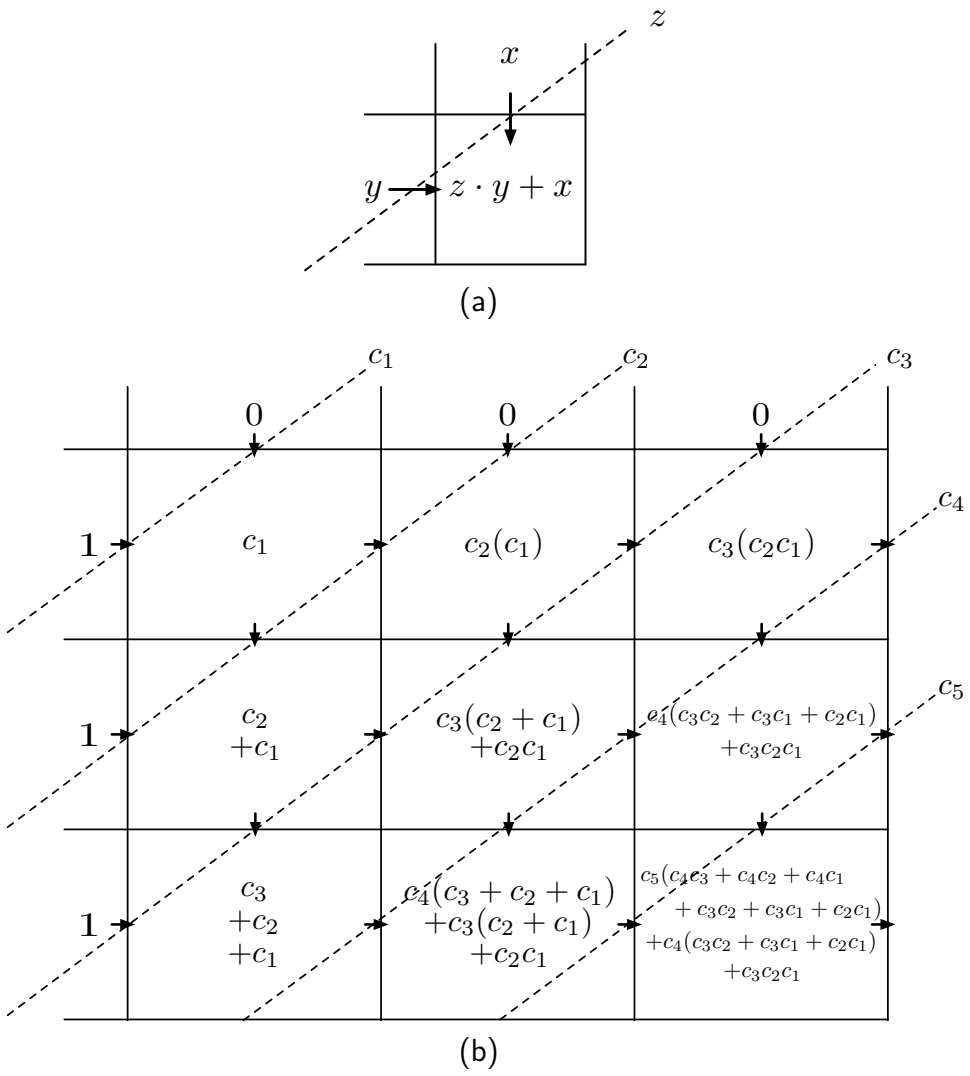


Figure 4.11: Dynamic program for computing coefficients of ξ_i in the cIBP. We initialize a matrix with zeros along the first row and ones along the first column. (a) By introducing an external variable along diagonals, we use the recursive step shown here to build up the computations. (b) Using the step from (a), we can fill in the matrix, introducing $c_1, c_2, \ldots$ along the diagonals. Then the bottom entry in the first column before the diagonal with c_i is $\sum_{j=1}^{i-1} c_j$, in the second column, it is $\sum_{j<k}^{i-1} c_j c_k$, and so on. For $i = 4$, we can have all required computations above. An inductive argument shows that this is correct.

4.A.3.4 Chain-based BP Equivalence

Here we show that Equation (4.13) is equivalent to the IBP when all $c_i = 1$. To do this, we must show that

$$\frac{1}{i} = 1 - \frac{\sum_{j=1}^{i-1} 1}{2} + \frac{\sum_{j<k}^{i-1} 1}{3} - \frac{\sum_{j<k<l}^{i-1} 1}{4} + \cdots + (-1)^{i-1}\frac{1}{i}$$

We rewrite the sum on the right hand side and express it as

$$\sum_{j=0}^{i-1} \frac{(-1)^j}{j+1}\binom{i-1}{j} = \sum_{j=0}^{i-1} a_j,$$

We can also define $r_0 = 1$ and for $j > 0$,

$$\begin{aligned} r_j &\equiv \frac{a_j}{a_{j-1}} \\ &= \frac{\frac{(-1)^j}{j+1}\binom{i-1}{j}}{\frac{(-1)^{j-1}}{j}\binom{i-1}{j-1}} \\ &= \frac{j-i}{j+1}. \end{aligned}$$

So therefore our original sum is

$$\begin{aligned} \sum_{j=0}^{i-1} a_j &= \sum_{j=0}^{i-1}\prod_{k=0}^{j} r_k \\ &= r_0 + r_0 r_1 + \cdots + \prod_{k=0}^{i-1} r_k \\ &= r_0(1 + r_1(1 + r_2(\cdots(1 + r_{i-2}(1 + r_{i-1}))\cdots))) \\ &= 1 + \frac{1-i}{2}\left(1 + \frac{2-i}{3}\left(\cdots\left(1 + \frac{-2}{i-1}\left(1 + \frac{-1}{i}\right)\right)\cdots\right)\right) \\ &= 1 + \frac{1-i}{2}\left(1 + \frac{2-i}{3}\left(\cdots\left(1 + \frac{-2}{i-1}\frac{i-1}{i}\right)\cdots\right)\right) \\ &= 1 + \frac{1-i}{2}\left(\frac{2}{i}\right) \\ &= \frac{1}{i} \end{aligned}$$

as desired.

4.A.4 Chain-based GP

Here we show two sets of derivations. The first is the equivalence between the continuous-time stochastic process and the transition kernel defined in Section 4.5.2.1. The second is the marginal distribution of the cIGPFM in Section 4.5.2.3.

4.A.4.1 Chain-based GP Stochastic Process

Here we show the equivalence between three descriptions of the continuous-time stochastic process 4.5.2.1. These three descriptions are

- The continuous-time birth-death process.
- The continuous-time Poisson process with each event dying independently with exponential rate.
- The transitions defined on the discrete-time birth-death process.

We will show that all three views are equivalent and therefore we can work with any of them as we see convenient to prove properties of the process.

Description 1 First we establish properties of the continuous-time birth death process. Let $z_k(t)$ be the value of this process at time t for feature k. Therefore z_{ik} is the value of $z_k(t)$ at time $\sum_{j=2}^{i} t_j$. Remember that at time $t = 0$, we set $z_k(0)$ to be initialized as Poisson(p_k). Again, we use the notation from (Cooper, 1981) that λ_i is the birth rate when $z_k(t) = i$, μ_i is the death rate when $z_k(t) = i$, and $p_i(t) = p(z_k(t) = i)$. For this process, $\lambda_i = \kappa p_k$ and $\mu_i = \kappa i$. Then we can verify that the constant Poisson(p_k) probabilities

$$p_i(t) = \frac{p_k^i e^{-p_k}}{i!}$$

satisfy

$$\begin{aligned} 0 = \frac{d}{dt} p_i(t) &= \lambda_{i-1} p_{i-1}(t) + \mu_{i+1} p_{i+1}(t) - (\lambda_i + \mu_i) p_i(t) \\ &= \kappa p_k \frac{p_k^{i-1} e^{-p_k}}{(i-1)!} + \kappa(i+1) \frac{p_k^{i+1} e^{-p_k}}{(i+1)!} - (\kappa p_k + \kappa i) \frac{p_k^i e^{-p_k}}{i!} \end{aligned}$$

for all i. Therefore the steady-state and initial state are Poisson(p_k), so marginally $z_k(t) \sim$ Poisson(p_k) for all t.

Description 2 Now that we have established that this continuous-time process has the correct marginals, we show that it is equivalent to the second description. This is mostly by definition, but allows us to then show the equivalence of these descriptions to the third description.

In the second description, we have a Poisson process with rate κp_k and each event dies independently with exponential rate κ.

In this process, during a fixed amount of time t, we expect a Poisson$(t\kappa p_k)$ number of events. The probability that one event happens as $t \downarrow 0$ is $(t\kappa p_k)e^{-t\kappa p_k} = t\kappa p_k + o(t)$ and the probability of more than one event happening is $o(t)$, so this is by definition a birth process with rate κp_k. If there are i active events at the current point in time, the probability that one dies in the next t amount of time as $t \downarrow 0$ is i times the probability that any particular event dies, which is $i(1 - e^{-\kappa t}) = t\kappa i + o(t)$, and the probability that more than one event dies is $o(t)$, so this is a death process with rate κi. As part of proving that the discrete-time birth-death process is equivalent to this, we will show that the number of events active at time zero with this definition is Poisson(p_k), agreeing with the distribution of $z_k(0)$.

Description 3 So the last step is to show that the discrete-time process is equivalent to the continuous-time Poisson process with exponential deaths and as a side effect, we will show that the number of active events at time zero in the second description is Poisson(p_k).

In our definition of the discrete-time process, there are two independent processes that happen from $z_{(i-1)k}$ to z_{ik}. Suppose $z_{(i-1)k}$ is observed at time t and z_{ik} is observed at time $t + t_i$. Then the two independent processes that happen from t to $t + t_i$ are:

- Each active event at time t dies independently with probability $c_i = 1 - e^{-\kappa t_i}$ before time $t + t_i$.
- An independent Poisson$(c_i p_k)$ number of events occur and remain active at time $t + t_i$.

These two processes can be captured by the two steps of $z_{(i-1)k} \to y_{ik} \to z_{ik}$. The transition $z_{(i-1)k} \to y_{ik}$ captures the first step and is the pure death process in which y_{ik} is the number of events alive at $t + t_i$ that existed at time t. The transition $y_{ik} \to z_{ik}$ captures the second step, in which $z_{ik} - y_{ik}$ is the number of new events added between t and $t + t_i$. We show that these two steps can be derived from the second description and is therefore equivalent to discrete-time observations of the underlying continuous-time process.

The first step represents the fact that each event in the continuous-time Poisson process dies with exponential rate κ, since $1 - e^{-\kappa t_i}$ is exactly the probability that an

event alive at t dies by $t+t_i$ by the memory-less property of the exponential distribution. Thus, with probability $1-c_i = e^{-\kappa t_i}$, each of the events in $z_{(i-1)k}$ is still alive in z_{ik}. This quantity is exactly y_{ik} which is therefore distributed Binomial$(z_{(i-1)k}, 1-c_i)$.

The second step represents the number of new events that happen between t and $t+t_i$ that are still alive at $t+t_i$. To determine this second quantity, we note that the base rate of the Poisson process is $\kappa p_k ds$. However, an event at time s with $t < s < t+t_i$ only survives until time $t+t_i$ with probability $e^{-\kappa(t+t_i-s)}$. Therefore, the base measure of the Poisson process for events between t and $t+t_i$ surviving until time $t+t_i$ is $\kappa p_k e^{-\kappa(t+t_i-s)} ds$. Integrating this from t to $t+t_i$, we see that the total number of events born between t and $t+t_i$ still alive at time $t+t_i$ is Poisson with rate

$$\begin{aligned}\int_t^{t+t_i} \kappa p_k e^{-\kappa(t+t_i-s)} ds &= p_k\left(1-e^{-\kappa t_i}\right) \\ &\equiv c_i p_k,\end{aligned}$$

which is exactly the number of events added in the transition from y_{ik} to z_{ik}. We can also see from this that by having z_{1k} be the number of events active at time zero with $t_1 = \infty$, then $z_{1k} \sim \text{Poisson}(p_k)$ as desired.

Therefore this two step discrete-time process is equivalent to the observations at select points in time of the continuous-time process in description two. As discussed earlier, since in the two steps, we are summing the results of an independent thinned Poisson process with resulting rate $p_k(1-c_i)$ and a Poisson process with rate $p_k c_i$, the sum is Poisson with rate p_k.

Therefore, we have shown all three descriptions are equivalent and we can choose which one to work with based on what we wish to show. If we wish to show that z_{ik} is consistent under marginalization, the first description is the most convenient, but if we wish to compute the transition probability, the third one is most convenient.

4.A.4.2 Chain-based GP Derivation

We discuss the derivation of the generative process for the chain-based GP discussed in Section 4.5.2.3.

As with the exchangeable GP, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite gamma-Poisson prior with the stochastic process discussed in Section 4.5.2.1 for each column. Both of these derivations are nearly identical to those of the pIGPFM.

Derivation 1 We now derive how to get the cIGPFM from the gamma-Poisson process construction of $p(Z)$. To correspond exactly to the cIGPFM, we set $c = 1$, but leave c as a variable in the below derivation.

For the first row z_1, we wish to sample from the distribution

$$p(z_1) \quad = \quad \int p(z_1|B)dP(B).$$

In this case, the distribution of B is the prior $\mathrm{GP}(c, B_0)$.

Since we have assumed that the total edge length to root for any entry is one, this is equivalent to the first entry in the IGPFM, so as before, we sample a $\mathrm{NB}(\alpha, 1/2)$ number of points distributed according to the CRP.

Unlike in the IGPFM, in the cIGPFM, we now draw p_k for the newly non-zero entries. This is the posterior of p_k with only a single, non-zero observation at z_{1k} as discussed in Section 4.5.2.2. This posterior is proportional to the Poisson likelihood times the improper gamma prior, which is therefore $\mathrm{Gamma}(z_{1k}, 2)$, as desired.

For the i^{th} row, we wish to sample z_i from the posterior distribution

$$p(z_i|z_1, \ldots, z_{i-1}) \quad = \quad \int p(z_i|B, z_1, \ldots, z_{i-1})dP(B|z_1, \ldots, z_{i-1}).$$

Recalling Section 4.5.2.2, we must reason about the discrete and continuous parts of B independently.

We first reason about the discrete part. Since we have sampled p_k, then we can directly sample z_{ik} from Equation (4.15).

We now reason about the continuous part of B for the i^{th} row. By Equation (4.16), the posterior Lévy measure of the continuous part of B given the first $i-1$ rows is

$$\nu(d\omega, dp) \quad = \quad cp^{-1}e^{-\left(c+\sum_{j=1}^{i-1} c_j\right)p}dpB_0(d\omega).$$

Similar to the IGPFM, using Equation (2.14), we get that

$$B(\Omega)|z_1, \ldots, z_{i-1} \sim \mathrm{Gamma}\left(cB_0(\Omega), c + \sum_{j=1}^{i-1} c_j\right).$$

Now using the conditional distribution Equation (4.15) with $z_{(i-1)k} = 0$ and the fact that if $x|\lambda \sim \mathrm{Poisson}(c_i\lambda)$ and $\lambda \sim \mathrm{Gamma}(a, b)$, then by marginalizing out λ, $x \sim \mathrm{NB}\left(a, \frac{b}{b+c_i}\right)$, we get that the total count of unobserved features g_i in z_i is $\mathrm{NB}\left(cB_0(\Omega), \frac{c+\sum_{j=1}^{i-1} c_j}{c+\sum_{j=1}^{i-1} c_j+c_i}\right)$, which substituting in $c = 1$ and $\alpha = B_0(\Omega)$ gives us $\mathrm{NB}\left(\alpha, \frac{\sum_{j=1}^{i-1} c_j+1}{\sum_{j=1}^{i-1} c_j+c_i+1}\right)$ as in Equation (4.17). We must now figure out how to distribute

these g_i new features. The posterior of B is still a GP in this case and $c_i \leq 1$, so we can interpret c_i as a thinning parameter for a full Poisson process and we know that the allocation of all features for the full process is distributed according to the CRP. Thinning the CRP still gives a CRP just with fewer features, so the g_i are allocated according to the CRP.

Finally, we draw p_k for the new features, which is drawing p_k from the posterior distribution given that the only events in the chain for feature k occurred along an edge of length t_i between $z_{(i-1)k}$ and z_{ik} with no events anywhere else in the chain, which is

$$\text{Gamma}\left(z_{ik}, \sum_{j=1}^{i-1} c_j + c_i + 1\right).$$

as in Equation (4.18).

Derivation 2 The alternate derivation of the cIGPFM is similar to the alternate derivation of the IGPFM. We start by placing a prior on finite $N \times K$ matrices and then letting K go to infinity. As with the IGPFM derivation, we start by sampling λ_k which will play the role of p_k in the infinite limit from

$$\lambda_k \sim \text{Gamma}\left(\frac{\alpha}{K}, 1\right) \qquad k \in \{1, \ldots, K\}$$

but now, the z_{ik} are drawn via the stochastic process from Section 4.5.2.1.

Instead of computing the closed form distribution of Z, we will compute how to sample the rows of Z incrementally while marginalizing out all λ_k for any all-zero columns. For the first row, since we have marginalized out all λ_k, then the distribution of each z_{1k} is $\text{NB}\left(\frac{\alpha}{K}, \frac{1}{2}\right)$. By exactly the same steps as for the pIGPFM (since the marginal distribution of the first row is the same in both cases), this implies that as $K \to \infty$, we can sample $g_1 = \sum_k z_{1k}$, the total number of non-zero features, from a $\text{NB}\left(\alpha, \frac{1}{2}\right)$ distribution and then partition it into K^+ groups according to the CRP. Using the same steps, we can again compute the the posterior distribution of λ_k (which play the role of p_k) for non-zero columns. Since this is a conjugate prior, this posterior distribution is

$$\text{Gamma}\left(z_{ik} + \frac{\alpha}{K}, 2\right) \overset{K\to\infty}{\to} \text{Gamma}(z_{ik}, 2)$$

as desired.

For all subsequent rows, as $K \to \infty$, we can sample z_{ik} for non-zero columns exactly from Equation (4.15).

For the all-zero columns of row i, we can check that for some constant ξ_i,

$$p(z_{ik}|z_{(-i)k} = 0, \alpha) \quad \sim \quad \text{NB}\left(\frac{\alpha}{K}, \xi_i\right).$$

If we do this, then by the exact same argument as in the pIGPFM, then as $K \to \infty$, this will imply that we can sample $g_i = \sum_{k=K^+ + 1}^{K^+ + K_i^{\text{new}}} z_{ik}$ from a $\text{NB}(\alpha, \xi_i)$ distribution and then partition it into K_i^{new} groups according to the CRP.

We must now identify ξ_i and we will be done with our derivation. In this case, using properties of the gamma and Poisson distributions, we can see that

$$\xi_i \quad = \quad \frac{\sum_{j=1}^{i-1} c_j + 1}{\sum_{j=1}^{i-1} c_j + c_i + 1}, \tag{4.22}$$

which therefore means we sample $g_i \sim \text{NB}\left(\alpha, \frac{\sum_{j=1}^{i-1} c_j + 1}{\sum_{j=1}^{i-1} c_j + c_i + 1}\right)$ and partition it according to the CRP, agreeing with Equation (4.17).

Finally, for each new non-zero column, we sample λ_k from the posterior distribution which is

$$\text{Gamma}\left(z_{ik} + \frac{\alpha}{K}, \sum_{j=1}^{i-1} c_j + c_i + 1\right) \quad \overset{K \to \infty}{\to} \quad \text{Gamma}\left(z_{ik}, \sum_{j=1}^{i-1} c_j + c_i + 1\right)$$

agreeing with Equation (4.18).

Chapter 5

Non-exchangeable Bayesian Nonparametric Latent Feature Model Inference Algorithms

We have now presented our two non-exchangeable generalizations for each of the beta and gamma processes. In this chapter we discuss how to perform posterior inference in models using these generalizations. We use the marginalized representations, the pIBP, cIBP, pIGPFM, and cIGPFM, in our inference algorithms. As in the exchangeable cases, exact inference is intractable, but approximate posterior inference is possible via Markov chain Monte Carlo (MCMC) (Robert and Casella, 2004). Sequential Monte Carlo approaches such as those proposed by Wood and Griffiths (2006) also work well, especially for the chain-based approaches. To adapt sequential approaches, we just need to take the generative models from Chapter 4 and apply standard techniques. Samples generated from sequential approached can be easily augmented with MCMC steps to improve performance, but we will not discuss sequential methods further.

An important point in our MCMC approaches is that even though we are dealing with a potentially infinite number of columns in Z, we only need to keep track of the non-zero columns. In this chapter, let Z refer to all the non-zero columns of the matrix. The fact that the number of non-zero entries in any given row has a Poisson or negative binomial distribution means that the number of non-zero columns is almost surely finite and generally small.

Given the matrix Z, we assume that data X are generated through a likelihood function $p(X|Z)$. The likelihood as discussed previously may include additional parameters θ that must be sampled as part of the overall MCMC procedure; we will not discuss such parameters in our presentation in this chapter since they are often application specific.

Unlike the IBP and IGPFM, where p_k can always be integrated out, inference in models using the non-exchangeable priors requires treating p_k as an auxiliary variable, sampling it when needed and integrating it out when possible. By sampling p_k for non-zero columns of Z as opposed to integrating it out, we are able to exploit efficient inference algorithms for trees and chains as discussed in each of the sections on conditional distributions in Chapter 4. For the all-zero columns, we will integrate out all corresponding p_k, allowing us to only store a finite number of parameters in each sample.

In the tree-based priors, updating p_k and all z_{ik} for each column takes $O(N+mN)$ time where m is the total number of times a z_{ik} in column k changes value. Once the chain has mixed well m is typically small, so time complexity is only slightly worse than that of models using the exchangeable priors, which is $O(N)$. In the chain-based priors, the updates take $O(N)$ time. However, in the BP generalization, as we will discuss in Section 5.3.3, there is a precomputation step whose complexity is anywhere from $O(N)$ to $O(N^2)$ depending on the structure of the chain. In the case of equally spaced observations, it is $O(N)$, but for more flexible structures, we pay the price for this flexibility with additional computation that must be run once before we start the posterior inference algorithm.

Given an initialization of the non-zero columns of Z, the corresponding p_k for each of these columns, and α, we construct a Markov chain where at each step, we only need to sample each variable from its conditional distribution given all others. After a sufficient burn-in period, samples will be from the desired posterior distribution.

We now describe how to sample each of the variables in models using each of the priors, first considering the variables for "old" columns—those with non-zero entries—and then turning to the addition of "new" columns. The "old" columns correspond to sampling the part of the i^{th} row associated with the discrete part of the posterior Lévy measure given all other rows and the "new" columns correspond to sampling from the continuous part of the posterior Lévy measure. When sampling from the discrete part, we first generate a sample p_k from the discrete Lévy measure, whereas for the continuous part, we integrate out all of the infinitely many associated p_k.

To generate each sample, we repeat the following until we have enough samples:

1. For $i = 1 : N$:
   (a) For each of the non-zero columns ("old columns"), sample z_{ik}. See Section 5.1 for details.
   (b) Sample the number of new columns for the i^{th} row and, for the GP variants, sample their partition. See Section 5.3 for details. For each of these new columns, sample a corresponding p_k. See Section 5.4 for details.
2. For $k = 1 : K^+$:

(a) Sample p_k. See Section 5.2 for details.

3. Sample α. See Section 5.5 for details.

If there are additional likelihood specific parameters θ, we resample θ between each of these iterations.

Fortunately, most derivations have already been done in the previous chapter. Appendix 5.A contains all derivations not previously discussed.

5.1 Sampling z_{ik} for Old Columns

The probability of each z_{ik} given all other variables is

$$p(z_{ik}|Z_{-(ik)}, p_k, X, \alpha) \propto p(X|Z_{-(ik)}, z_{ik})p(z_{ik}|z_{(-i)k}, p_k), \tag{5.1}$$

where the first term is the probability of X given a full assignment of the parameters and depends on the specific prior being used. The second term can be computed easily depending on which generalization we are using. If a column ever becomes entirely zero after one of these steps, we drop it from Z.

5.1.1 pIBP

The term $p(z_{ik}|z_{(-i)k}, p_k)$ can be computed efficiently using the sum-product algorithm as described in Section 4.4.1.2. By appropriately caching messages from the sum-product algorithm, this evaluation can be reduced to $O(1)$ time. We evaluate Equation (5.1) for $z_{ik} = 0$ and $z_{ik} = 1$ and sample z_{ik} from the corresponding posterior distribution. If the value of z_{ik} changes, we then update the messages for sum-product in $O(N)$ time.

5.1.2 pIGPFM

The term $p(z_{ik}|z_{(-i)k}, p_k)$ can again be computed efficiently using the sum-product algorithm as described in Section 4.4.2.2. By appropriately caching messages from the sum-product algorithm, this evaluation can be reduced to $O(1)$ time. Unlike in the pIBP, z_{ik} can take on an infinite number of values in the pIGPFM. Since we cannot evaluate Equation (5.1) for each of these values and then normalize, to be strictly correct, we must perform a Metropolis-Hastings step to sample z_{ik}. A close approximation often used in practice is to note that the prior likelihood decreases exponentially quickly as z_{ik} grows, so in practice, we can pick some large enough value M such that is is unlikely z_{ik} would be larger than M, evaluate Equation (5.1)

for $z_{ik} \in [0, \ldots, M]$ and sample z_{ik} from the corresponding posterior distribution. If the value of z_{ik} changes, we then update the messages for sum-product in $O(N)$ time.

5.1.3 cIBP

We can evaluate

$$\begin{aligned} p(z_{ik}|z_{(-i)k}, p_k) &= p(z_{ik}|z_{(i-1)k}, z_{(i+1)k}, p_k) \\ &\propto p(z_{(i+1)k}|z_{ik}, p_k)p(z_{ik}|z_{(i-1)k}, p_k) \end{aligned}$$

for $z_{ik} \in \{0, 1\}$, where each of the last two equations can be computed via the transition kernel defined in Equation (4.11).

5.1.4 cIGPFM

Just as in the cIBP, for any value of z_{ik}, we can evaluate

$$p(z_{ik}|z_{(-i)k}, p_k) \propto p(z_{(i+1)k}|z_{ik}, p_k)p(z_{ik}|z_{(i-1)k}, p_k)$$

where each of the last two equations can be computed via the transition kernel defined in Equation (4.15). As in the pIGPFM, to be strictly correct, we must perform an Metropolis-Hastings step to sample z_{ik}, but in practice, we often pick a maximum M and sample z_{ik} from $[0, \ldots, M]$.

5.2 Sampling p_k for Old Columns

We only sample p_k for the non-zero columns of Z, a fact that will be useful in subsequent calculations since it makes the posterior distribution proper. The posterior distribution of each p_k is independent of all other $p_{k'}$ and depends only on the k^{th} column of Z, so we discuss how to sample each of these independently.

Except for the special cases of the exchangeable IBP and IGPFM, there is no simple posterior distribution from which we can sample p_k. However we can evaluate the posterior probability of any particular value of p_k (up to a normalizing constant) and can therefore sample from its posterior distributions using Metropolis-Hastings. Let z_k be shorthand for $\{z_{ik}\}_{i=1}^N$, so that for each k, we will independently sample p_k from

$$p(p_k|z_k) \propto p(z_k|p_k)p(p_k)$$

where $p(p_k)$ is the improper beta or gamma prior given by the beta and gamma Lévy

measures. Despite this improper prior, since there is at least one non-zero entry z_{ik} in each column, the posterior of p_k must be proper. We can see this from the following:

$$\begin{aligned} p(p_k|z_k) &\propto p(z_k|p_k)p(p_k) \\ &= p(z_{(-i)k}|z_{ik}, p_k)p(z_{ik}|p_k)p(p_k). \end{aligned}$$

Since in the generalizations of the BP and GP, each z_{ik} is marginally Bernoulli(p_k) or Poisson(p_k), respectively, in the BP generalizations

$$\underbrace{p(z_{ik}|p_k)}_{\sim\text{Bernoulli}(p_k)} \underbrace{p(p_k)}_{\sim p_k^{-1}} \propto \text{Beta}(p_k; 1, 1)$$

and in the GP generalizations

$$\underbrace{p(z_{ik}|p_k)}_{\sim\text{Poisson}(p_k)} \underbrace{p(p_k)}_{\sim p_k^{-1}e^{-p_k}} \propto \text{Gamma}(p_k; z_{ik}, 2).$$

The term $p(z_{(-i)k}|z_{ik}, p_k)$ can be computed via sum-product for trees or chains and therefore, the posterior is always proper for the non-zero columns and we can efficiently evaluate $p(p_k|z_k)$. We use this to sample p_k from its posterior distribution using a Metropolis-Hastings step.

Specifically, given a proposed value of p' for p_k, if we use $q(p'|p_k)$ as the proposal distribution and the beta or gamma Lévy measure as the prior $p(p_k)$, then the Metropolis-Hastings acceptance ratio for p' is

$$\min\left[1, \frac{q(p_k|p')}{q(p'|p_k)}\frac{p(p'|z_k)}{p(p_k|z_k)}\right] = \min\left[1, \frac{q(p_k|p')}{q(p'|p_k)}\frac{p(z_k|p')p(p')}{p(z_k|p_k)p(p_k)}\right]. \tag{5.2}$$

We are then left to chose an appropriate proposal distribution q. For example, we might use $q(p'|p_k) \sim \mathcal{N}(p_k, \sigma_k^2)$ where $\sigma_k^2 = c \cdot p_k(1 - p_k) + \delta$.

5.2.1 pIBP

Section 4.4.1.2 describes how to compute $p(z_k|p_k)$ efficiently in $O(N)$ time using the sum-product algorithm and the chain rule of probabilities. We plug this into Equation (5.2) to sample p_k.

5.2.2 pIGPFM

Section 4.4.2.2 describes how to compute $p(z_k|p_k)$ efficiently in $O(N)$ time using the sum-product algorithm and the chain rule of probabilities. We plug this into

Equation (5.2) to sample p_k.

5.2.3 cIBP

We can evaluate

$$p(z_k|p_k) = \prod_{i=1}^{N} p(z_{ik}|z_{(i-1)k}, p_k)$$

using the transition probabilities in Equation (4.11), allowing us to sample p_k using Equation (5.2).

5.2.4 cIGPFM

We can evaluate

$$p(z_k|p_k) = \prod_{i=1}^{N} p(z_{ik}|z_{(i-1)k}, p_k)$$

using Equation (4.15), allowing us to sample p_k using Equation (5.2).

5.3 Sampling the New Columns

In addition to sampling z_{ik} for all the non-zero features, we must also sample z_{ik} for the infinitely many all-zero features in row i. This is the main tricky step in inference and requires the most care when deriving new generalizations of the BP and GP. As previously mentioned, we only sample p_k for the non-zero columns, so we must integrate out p_k when computing the posterior for the all-zero columns. We first look at

$$p(z_{ik}|z_{(-i)k} = 0, \alpha) = \int_0^1 p(z_{ik}|z_{(-i)k} = 0, p_k) p(p_k|z_{(-i)k} = 0, \alpha) dp_k.$$

Since we have a nonparametric model in which the posterior of p_k given a column with all zero entries is still improper, the probability that z_{ik} is non-zero for any particular k is zero. This makes sense because otherwise, we would be sampling an infinite number of new non-zero entries. However, we know that marginally, each row will only have a finite number of elements, so this cannot be the case.

Since there are an infinite number of these all-zero features, it turns out that we can sample the number of entries that become non-zero from a non-degenerate

distribution in a batch versus working with each feature independently. In the case of the tree-based priors, this distribution was computed exactly in the generative distribution for the pIBP or pIGPFM, assuming that the i^{th} object is the last one observed in the tree. In the case of the cIBP and cIGPFM, this distribution is more complicated as discussed in Sections 5.3.3 and 5.3.4.

Sampling the features for the i^{th} row in a batch, as mentioned in the introduction to this chapter, is equivalent to sampling the non-zero features generated from the continuous part of the posterior Lévy measure given all other rows. We will first calculate the the posterior Lévy measure given all other rows and then show how to generate the number of new features for the i^{th} row.

In the case of the BP generalizations, we will show how to sample K_i^{new}, the number of new non-zero entries in the i^{th} row. We sample K_i^{new} from its posterior distribution:

$$p(K_i^{\text{new}}|X, Z, \alpha) \quad \propto \quad p(X|Z, K_i^{\text{new}})p(K_i^{\text{new}}|Z, \alpha), \tag{5.3}$$

where the term $p(X|Z, K_i^{\text{new}})$ is the likelihood of X given that Z is augmented by K_i^{new} non-zero entries in the i^{th} row. In Sections 5.3.1 and 5.3.3, we show that $p(K_i^{\text{new}}|Z, \alpha)$ is a Poisson distribution with rate depending on the configuration of the tree or chain. To be completely correct, we must evaluate Equation (5.3) for $K_i^{\text{new}} \in \{0, 1, \ldots\}$ up to ∞, normalize, and then sample K_i^{new} from the resulting distribution. We can use either Metropolis-Hastings or a slice sampler like the one in (Teh et al., 2007) to do this exactly or we can approximate this by sampling K_i^{new} from $\{0, 1, \ldots, K_{\max}\}$ for some $K_{\max}$ such that it is extremely unlikely K_i^{new} falls outside of the range due to the Poisson prior on K_i^{new}.

In the case of the GP generalizations, we must sample the number of new features in row i and their allocation. Designate the new non-zero columns containing the allocation of the new features by z_i^{new}. Let g_i be the number of new features in z_i^{new}. Then, since given z_i^{new}, g_i is deterministic, we wish to sample z_i^{new} from

$$p(z_i^{\text{new}}|X, Z, \alpha) \quad \propto \quad p(X|Z, z_i^{\text{new}})p(z_i^{\text{new}}|g_i)p(g_i|Z, \alpha), \tag{5.4}$$

where $p(X|Z, z_i^{\text{new}})$ is the likelihood of X given the old features Z and the new allocation of features z_i^{new} to the i^{th} row, $p(z_i^{\text{new}}|g_i)$ is the probability of the allocation of features given their count, and $p(g_i|Z, \alpha)$ is the probability that g_i new features are present in row i. We will discuss the latter two terms in Sections 5.3.2 and 5.3.4, showing that they are Ewens distribution (the CRP in which we only care about counts in partitions) and a negative binomial, respectively, where the parameters of the negative binomial have to do with the structure of the tree or chain. To be completely correct, we must evaluate Equation (5.4) for all possible counts g_i and all

allocations of those counts z_i^{new} and then sample from the resulting posterior. Since we cannot do this in practice, we can either sample using a Metropolis-Hastings step or, given that the distribution on g_i makes it unlikely to have g_i large, we can often evaluate g_i for only a moderate range $\{0, 1, \ldots, g_{\max}\}$ for some $g_{\max}$ and find a reasonable approximation.

Therefore, to sample the new columns for our generalizations, we must identify $p(K_i^{\text{new}}|Z, \alpha)$ for the BP generalizations and $p(z_i^{\text{new}}|g_i)$ and $p(g_i|Z, \alpha)$ for the GP generalizations. Plugging these quantities into Equations (5.3) and (5.4) then allows us to sample the new columns.

5.3.1 pIBP

For the pIBP, we can see that when computing the posterior distribution of the i^{th} row, we can view the tree as if the i^{th} object was ordered last. Therefore, all the calculations from Section 4.4.1 and Appendix 4.A.1 are still valid. Let $\sum t$ be the sum of the lengths of all edges in the tree except the edge that connects node i to the rest of the tree. Let t_i be the length of this last edge. Then it was shown that the posterior Lévy measure of B given $z_{(-i)k} = 0$ is $p^{-1}(1-p)^{\sum t} dp B_0(d\omega)$ and therefore, that the distribution $p(K_i^{\text{new}}|Z, \alpha)$ is

$$\text{Poisson}\left(\alpha\left(\psi\left(\sum t + t_i + 1\right) - \psi\left(\sum t + 1\right)\right)\right).$$

5.3.2 pIGPFM

For the pIGPFM, we can similarly see that when computing the posterior distribution of the i^{th} row, we can view the tree as if the i^{th} object was ordered last. Therefore, all the calculations from Section 4.4.2 and Appendix 4.A.2 are still valid. Let $\sum t$ be the sum of the lengths of all edges in the tree except the edge that connects node i to the rest of the tree. Let t_i be the length of this last edge. Then it was shown that the posterior Lévy measure of B given $z_{(-i)k} = 0$ is $p^{-1}e^{-(\sum t+1)p} dp B_0(d\omega)$ and therefore, that the distribution $p(g_i|Z, \alpha)$ is

$$\text{NB}\left(\alpha, \frac{\sum t + 1}{\sum t + t_i + 1}\right),$$

and $p(z_i^{\text{new}}|g_i)$ just follows Ewens distribution with g_i objects.

5.3.3 cIBP

When $i = N$, all the calculations from Section 4.5.1 and Appendix 4.A.3 are still valid. In these sections, we showed that $p(K_N^{\text{new}}|Z, \alpha)$ is

$$K_N^{\text{new}} \sim \text{Poisson}(\alpha\xi_N),$$

where

$$\xi_N = c_N\left(1 - \frac{\sum_{j=1}^{N-1} c_j}{2} + \frac{\sum_{j<k}^{N-1} c_j c_k}{3} - \frac{\sum_{j<k<l}^{N-1} c_j c_k c_l}{4} + \cdots + (-1)^{N-1}\frac{\prod_{j=1}^{N-1} c_j}{N}\right).$$

When $i \neq N$, we will show in Appendix 5.A.1 that $p(K_i^{\text{new}}|Z, \alpha)$ is

$$K_i^{\text{new}} \sim \text{Poisson}(\alpha\xi_i),$$

where

$$\xi_i = c_i c_{i+1}\Big(\frac{1}{2\cdot 1} - \frac{\sum_{j\notin\{i,i+1\}} c_j}{3\cdot 2} + \frac{\sum_{j<k\notin\{i,i+1\}} c_j c_k}{4\cdot 3} - \frac{\sum_{j<k<l\notin\{i,i+1\}} c_j c_k c_l}{5\cdot 4} + \cdots + (-1)^{N-2}\frac{\prod_{j\notin\{i,i+1\}} c_j}{N(N-1)}\Big). \quad (5.5)$$

For a homogeneous Markov chain, this can again be computed in $O(N)$ time or in the general case, we can use the dynamic program from Figure 4.11 in Appendix 4.A.3.2 to compute this in $O(N^2)$ time. Through shared common computations, we can also compute all ξ_i ahead of time and cache them in $O(N^2)$ time. If c_i only take on a few distinct values, this can be greatly improved.

5.3.4 cIGPFM

When $i = N$, all the calculations from Section 4.5.2 and Appendix 4.A.4 are still valid. In these sections, we showed that $p(g_N|Z, \alpha)$ is

$$g_N \sim \text{NB}\left(\alpha, \frac{1 + \sum_{j=1}^{N-1} c_j}{1 + \sum_{j=1}^{N} c_j}\right),$$

and $p(z_N^{\text{new}}|g_N)$ is Ewens distribution for g_N objects.

When $i \neq N$, we will show in Appendix 5.A.2 that $p(g_i|Z,\alpha)$ is

$$g_i \sim \text{NB}\left(\alpha, \frac{1 - c_{i+1}c_i + \sum_{j=1}^{N} c_j}{1 + \sum_{j=1}^{N} c_j}\right), \tag{5.6}$$

and $p(z_i^{\text{new}}|g_i)$ is Ewens distribution with g_i objects.

5.4 Sampling p_k for New Columns

For each of the new columns generated in the previous step, we must sample an initial value of p_k.

5.4.1 pIBP

Using the same notation as before for edge lengths in which $\sum t$ is the sum of the lengths of all edges in the tree except the edge that connects node i to the rest of the tree and t_i is the length of this last edge, if we are sampling p_k for a new column in which only the i^{th} element is non-zero, then as was shown in Section 4.A.1,

$$\begin{aligned} p(p_k|z_k) &= p(p_k|z_{(-i)k} = 0, z_{ik} = 1) \\ &\propto \left(1 - (1-p_k)^{t_i}\right)(1-p_k)^{\sum t} p_k^{-1}. \end{aligned}$$

To obtain a sample from this distribution, we use the Metropolis-Hastings sampler discussed in Section 5.2.

Though it might not be clear that $p(p_k|z_k)$ is a proper distribution, we could have also expressed the posterior as

$$p(p_k|z_k,\alpha) \propto p(z_{(-i)k} = 0|z_{ik} = 1, p_k)p(p_k|z_{ik} = 1, \alpha),$$

which is clearly a proper distribution since the last term is now proper.

5.4.2 pIGPFM

Using the same notation as before for edge lengths, if we are sampling p_k for a new column in which only the i^{th} element is non-zero, then as was shown in Section 4.A.2,

$$p_k \sim \text{Gamma}\left(z_{ik}, \sum t + t_i + 1\right).$$

5.4.3 cIBP

If we are sampling p_k for column k that only has the i^{th} entry non-zero, then for $i = N$, we showed in Section 4.A.3.2 that

$$p(p_k|z_k) \propto c_N \prod_{j<N}(1 - c_j p_k).$$

For $i \neq N$,

$$\begin{aligned} p(p_k|z_k) &\propto p(p_k)\prod_{j=1}^{N} p(z_{jk}|z_{(j-1)k}, p_k) \\ &\propto c_i c_{i+1}(1-p_k)\prod_{j\notin\{i,i+1\}}(1-c_j p_k). \end{aligned}$$

To obtain a sample from these distributions, we use the Metropolis-Hastings sampler discussed in Section 5.2.

5.4.4 cIGPFM

If we are sampling p_k for column k that only has the i^{th} entry non-zero, then for $i = N$, we showed in Section 4.A.4.2 that

$$p_k \sim \text{Gamma}\left(z_{Nk}, \sum_{j=1}^{N} c_j + 1\right).$$

For $i \neq N$, we can similarly show

$$\begin{aligned} p(p_k|z_k) &\propto p(p_k)\prod_{j=1}^{N} p(z_{jk}|z_{(j-1)k}, p_k) \\ &\propto p_k^{z_{ik}-1} e^{-\left(\sum_{j=1}^{N} c_j + 1\right)p_k} \\ &\sim \text{Gamma}\left(z_{ik}, \sum_{j=1}^{N} c_j + 1\right). \end{aligned}$$

5.5 Sampling α

The parameter $\alpha = B_0(\Omega)$ is an important parameter that determines how many features are present in inferred models. It is therefore critical to also be able to

perform posterior inference on α.

For the IBP, this was originally done by Görür et al. (2006) and for the IGPFM, this was done by Titsias (2008). Both placed a Gamma(α_0, β_0) prior on α and, since likelihoods do not depend on α, the posterior distribution of α is

$$p(\alpha|Z) \propto p(Z|\alpha)p(\alpha).$$

For all our priors, we will show that this posterior is always a gamma distribution with parameters depending on the structure of the model as well as the number of non-zero columns in Z.

5.5.1 pIBP

Let $K^+ = \sum_{j=1}^{N} K_j^{\text{new}}$ be the number of non-zero features in Z and $\sum t$ be the total edge length in the entire tree. Since the total number of non-zero columns in Z is the sum of independent Poisson variables K_i^{new}, each drawn from Equation (4.4) in which we incrementally add K_i^{new} based on $1, \ldots, i-1$, then

$$\begin{aligned} p(\alpha|Z) &\propto p(Z|\alpha)p(\alpha) \\ &\sim \text{Poisson}\left(K^+; \alpha\big(\psi\big(1+\sum t\big) - \psi(1)\big)\right) \cdot \text{Gamma}(\alpha_0, \beta_0) \\ &\sim \text{Gamma}\left(\alpha_0 + K^+, \beta_0 + \psi\Big(1+\sum t\Big) - \psi(1)\right). \end{aligned}$$

5.5.2 pIGPFM

As shown in Equations (4.19) and (4.20), α again only influences Z through K_i^{new}. Therefore, if $\sum t$ is the total edge length in the entire tree, using these equations where ξ_i is defined in Equation (4.21) based on the incremental addition of K_i^{new} new distinct features given $1, \ldots, i-1$, then

$$\begin{aligned} p(\alpha|Z) &\propto p(Z|\alpha)p(\alpha) \\ &\propto \left[\prod_{i=1}^{N} \alpha^{K_i^{\text{new}}} \xi_i^{\alpha}\right] \alpha^{\alpha_0 - 1} e^{-\beta_0 \alpha} \\ &\sim \text{Gamma}\left(\alpha_0 + K^+, \beta_0 - \log\left(\prod_{j=1}^{N} \xi_j\right)\right) \\ &\sim \text{Gamma}\left(\alpha_0 + K^+, \beta_0 + \log\Big(1+\sum t\Big)\right). \end{aligned}$$

5.5.3 cIBP

In the case of the cIBP, each K_i^{new} is drawn from Equation (4.13)

$$K_i^{\text{new}} \sim \text{Poisson}(\alpha\xi_i)$$

where ξ_i is

$$\xi_i = c_i\left(1 - \frac{\sum_{j=1}^{i-1} c_j}{2} + \frac{\sum_{j<k}^{i-1} c_j c_k}{3} - \frac{\sum_{j<k<l}^{i-1} c_j c_k c_l}{4} + \cdots + (-1)^{i-1}\frac{\prod_{j=1}^{i-1} c_j}{i}\right).$$

Therefore,

$$p(\alpha|Z) \sim \mathcal{G}\left(\alpha_0 + K^+, \beta_0 + \sum_{i=1}^{N}\xi_i\right).$$

It is important to always calculate each ξ_i only based on $1,\ldots,i-1$ in order for this to be correct. Plugging in ξ_i, we get

$$\sum_{i=1}^{N}\xi_i = \sum_{i=1}^{N} c_i - \sum_{i<j}\frac{c_i c_j}{2} + \sum_{i<j<k}\frac{c_i c_j c_k}{3} - \cdots + (-1)^{N-1}\frac{\prod_{j=1}^{N} c_j}{N}.$$

5.5.4 cIGPFM

We can again use Equations (4.19) and (4.20) to show that α again only influences Z through K_i^{new}. Therefore, using these equations where ξ_i now defined in Equation (4.22) based on the incremental addition of K_i^{new} given $1,\ldots,i-1$, then

$$\begin{aligned} p(\alpha|Z) &\sim \text{Gamma}\left(\alpha_0 + K^+, \beta_0 - \log\left(\prod_{j=1}^{N}\xi_j\right)\right) \\ &= \text{Gamma}\left(\alpha_0 + K^+, \beta_0 + \log\left(1 + \sum_{j=1}^{N} c_j\right)\right). \end{aligned}$$

5.6 Summary

We have now shown how to perform each of the steps needed for the inference algorithms for our non-exchangeable variations. Most of the computations were already done when deriving the marginal representations of each generalization. Only the distributions of the number of new columns in the chain-based generalizations required

any additional computations and those can be found in Section 5.A.

Appendix 5.A Derivations

We discuss the additional derivations needed in the inference algorithm of new columns in the two chain-based generalizations.

5.A.1 Chain-based BP

We discuss the derivation of the sampling distribution of K_i^{new} for $i \neq N$ in the chain-based BP discussed in Section 5.3.3.

As with all other derivations, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite beta-Bernoulli prior with the stochastic process discussed in Section 4.5.1.1 for each column. In this section, let $z_{(-i)}$ stand for $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N$.

Derivation 1 For the i^{th} row, we wish to sample z_i from continuous part of the posterior distribution

$$p(z_i|z_{(-i)}) \quad = \quad \int p(z_i|B, z_{(-i)}) dP(B|z_{(-i)}).$$

Let $c_{i+(i+1)}$ be $1 - e^{-\kappa(t_i+t_{i+1})}$, the parameter of the transition from z_{i-1} to z_{i+1} if z_i is not observed. By Equation (4.12), the posterior Lévy measure of the continuous part of B given $z_{(-i)}$ is

$$\nu(d\omega, dp) \quad = \quad cp^{-1}(1-p)^{c-1} \left[\prod_{j=1:j\notin\{i,i+1\}}^{N} (1 - c_j p) \right] (1 - c_{i+(i+1)}p) dp B_0(d\omega).$$

Using the transition probabilities for the cIBP defined in Equation (4.11), the probability of observing $z_{ik} = 1$ given that $z_{(-i)}$ are all zero and p_k is

$$\begin{aligned} p(z_{ik} = 1|z_{(-i)k} = 0, p_k) \quad &= \quad \frac{p(z_{(i+1)k} = 0|z_{ik} = 1, p_k) p(z_{ik} = 1|z_{(i-1)k} = 0, p_k)}{p(z_{i+1} = 0|z_{i-1} = 0, p_k)} \\ &= \quad \frac{c_i p_k c_{i+1}(1 - p_k)}{1 - c_{i+(i+1)} p_k}. \end{aligned}$$

Therefore, the number of new features for the i^{th} row is Poisson with parameter λ where

$$\begin{aligned}\lambda &= \int\int_0^1 \frac{c_i p c_{i+1}(1-p)}{1-c_{i+(i+1)}p} c p^{-1}(1-p)^{c-1}\left[\prod_{j=1:j\notin\{i,i+1\}}^{N}(1-c_j p)\right](1-c_{i+(i+1)}p)\,dp\,B_0(d\omega)\\ &= c c_i c_{i+1} B_0(\Omega)\int_0^1 (1-p)^c\left[\prod_{j=1:j\notin\{i,i+1\}}^{N}(1-c_j p)\right]dp.\end{aligned}$$

Now we note that by expanding $\left[\prod_{j=1:j\notin\{i,i+1\}}^{N}(1-c_j p)\right]$, we get a weighted sum of proper beta distributions, so we can evaluate this integral.

$$\begin{aligned}&\int_0^1 (1-p)^c\left[\prod_{j=1:j\notin\{i,i+1\}}^{N}(1-c_i p)\right]dp\\ &= \int_0^1 (1-p)^c\left[1-p\sum_{j\notin\{i,i+1\}}c_j+p^2\sum_{j<k\notin\{i,i+1\}}c_j c_k-\cdots+(-1)^{N-2}p^{N-2}\prod_{j\notin\{i,i+1\}}c_j\right]dp\\ &= \frac{\Gamma(1)\Gamma(c+1)}{\Gamma(c+2)}-\frac{\Gamma(2)\Gamma(c+1)}{\Gamma(c+3)}\sum_{j\notin\{i,i+1\}}c_j+\frac{\Gamma(3)\Gamma(c+1)}{\Gamma(c+4)}\sum_{j<k\notin\{i,i+1\}}c_j c_k-\cdots\\ &\qquad+\frac{\Gamma(N-1)\Gamma(c+1)}{\Gamma(c+N)}(-1)^{N-2}\prod_{j\notin\{i,i+1\}}c_j.\end{aligned}$$

Substituting $c=1$, this simplifies to

$$\begin{aligned}&\int_0^1 (1-p)\left[\prod_{j=1:j\notin\{i,i+1\}}^{N}(1-c_i p)\right]dp\\ &\quad= \frac{1}{2\cdot 1}-\frac{1}{3\cdot 2}\sum_{j\notin\{i,i+1\}}c_j+\frac{1}{4\cdot 3}\sum_{j<k}c_j c_k-\cdots+\frac{1}{N(N-1)}(-1)^{N-2}\prod_{j\notin\{i,i+1\}}c_j.\end{aligned}$$

Therefore, substituting $B_0(\Omega)=\alpha$,

$$\lambda = \alpha c_i c_{i+1}\left(\frac{1}{2\cdot 1}-\frac{\sum_{j\notin\{i,i+1\}}c_j}{3\cdot 2}+\frac{\sum_{j<k}c_j c_k}{4\cdot 3}-\cdots+(-1)^{N-2}\frac{\prod_{j\notin\{i,i+1\}}c_j}{N(N-1)}\right),$$

which agrees with Equation (5.5).

Derivation 2 We now show how to derive Equation (5.5) from the infinite limit of a finite beta-Bernoulli prior. As discussed in Derivation 2 of Section 4.A.1, if we can show that for some constant ξ_i,

$$p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) \quad = \quad \frac{\alpha \xi_i}{K} + o\left(\frac{1}{K}\right),$$

then $K_i^{\text{new}} \sim \text{Poisson}(\alpha \xi_i)$.

We therefore wish to find ξ_i such that $p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) = \alpha \xi_i / K + o(1/K)$. For finite K, we can evaluate the probability that $z_{ik} = 1$ given that $z_{(-i)k} = 0$ in the cIBP.

$$\begin{aligned}
& p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) \\
& \quad \propto \quad \int_0^1 p(z_{(i+1)k} = 0 | z_{ik} = 1, p_k) p(z_{ik} = 1 | z_{(i-1)k} = 0, p_k) p(p_k | \alpha) \\
& \qquad\qquad \times \prod_{j \notin \{i, i+1\}} p(z_{jk} = 0 | z_{(j-1)k} = 0, p_k) dp_k \\
& \quad = \quad c_i c_{i+1} \frac{\alpha}{K} \Bigg(\frac{1}{(\alpha/K + 2)(\alpha/K + 1)} - \frac{1}{(\alpha/K + 3)(\alpha/K + 2)} \sum_{j \notin \{i, i+1\}}^{N} c_j + \\
& \qquad\qquad \ldots + (-1)^{N-2} \frac{1}{(\alpha/K + N)(\alpha/K + N - 1)} \prod_{j \notin \{i, i+1\}}^{N} c_j \Bigg).
\end{aligned}$$

Similarly, no matter what i is

$$\begin{aligned}
& p(z_{ik} = 0 | z_{(-i)k} = 0, \alpha) \\
& \quad \propto \quad \int_0^1 p(p_k) \prod_{j=1}^{N} p(z_{jk} = 0 | z_{(j-1)k} = 0, p_k) dp_k \\
& \quad = \quad 1 - \frac{\alpha/K}{\alpha/K + 1} \sum_{j=1}^{N} c_j + \frac{\alpha/K}{\alpha/K + 2} \sum_{j<k}^{N} c_j c_k - \ldots + (-1)^N \frac{\alpha/K}{\alpha/K + N} \prod_{j=1}^{N} c_j.
\end{aligned}$$

We can now do a first order Taylor series expansion on the resulting normalized probability to get that $p(z_{ik} = 1 | z_{(-i)k} = 0, \alpha) = \alpha \xi_i / K + o(1/K)$ where

$$\begin{aligned}
\xi_i \quad = \quad c_i c_{i+1} \Big(& \frac{1}{2} - \frac{\sum_{j \notin \{i, i+1\}} c_j}{3 \cdot 2} + \frac{\sum_{j<l \notin \{i, i+1\}} c_j c_l}{4 \cdot 3} - \frac{\sum_{j<k<l \notin \{i, i+1\}} c_j c_k c_l}{5 \cdot 4} + \\
& \cdots + (-1)^{N-2} \frac{\prod_{j \notin \{i, i+1\}}^{N} c_j}{N(N-1)} \Big)
\end{aligned}$$

as desired.

Equivalence to the IBP We note that similar to what we showed in Section 4.A.3.4 for $i = N$, that for $i \neq N$ we still must reduce to the IBP in the special when all $c_i = 1$. In order to have this happen, we must have

$$\begin{aligned}\frac{1}{N} &= \frac{1}{2 \cdot 1} - \frac{\sum_{j \notin \{i,i+1\}} 1}{3 \cdot 2} + \frac{\sum_{j<k} 1}{4 \cdot 3} - \cdots + (-1)^{N-2} \frac{1}{N(N-1)} \\ &= \sum_{j=0}^{N-2} \frac{(-1)^j \binom{N-2}{j}}{(j+2)(j+1)}.\end{aligned}$$

Define

$$a_j \equiv \frac{(-1)^j \binom{N-2}{j}}{(j+2)(j+1)}.$$

We also define $r_0 = \frac{1}{2}$ and for $j > 0$,

$$\begin{aligned}r_j &\equiv \frac{a_j}{a_{j-1}} \\ &= \frac{\frac{(-1)^j \binom{N-2}{j}}{(j+2)(j+1)}}{\frac{(-1)^{j-1} \binom{N-2}{j-1}}{(j+1)(j)}} \\ &= \frac{j-(N-1)}{j+2}\end{aligned}$$

So therefore our original sum is

$$\begin{aligned}
&\sum_{j=0}^{N-2} \frac{(-1)^j \binom{N-2}{j}}{(j+2)(j+1)} \\
&= \sum_{j=0}^{N-2} a_j = \sum_{j=0}^{N-2} \prod_{k=0}^{j} r_k = r_0 + r_0 r_1 + \cdots + \prod_{k=0}^{N-2} r_k \\
&= r_0(1 + r_1(1 + r_2(\cdots(1 + r_{N-3}(1 + r_{N-2}))\cdots))) \\
&= \frac{1}{2}\left(1 - \frac{N-2}{3}\left(1 - \frac{N-3}{4}\left(\cdots\left(1 - \frac{2}{N-1}\left(1 - \frac{1}{N}\right)\right)\cdots\right)\right)\right) \\
&= \frac{1}{2}\left(1 - \frac{N-2}{3}\left(1 - \frac{N-3}{4}\left(\cdots\left(1 - \frac{2}{N-2}\right)\cdots\right)\right)\right) \\
&= \frac{1}{2}\left(1 - \frac{N-2}{3}\left(\frac{3}{N}\right)\right) \\
&= \frac{1}{N},
\end{aligned}$$

as in the case of the IBP.

5.A.2 Chain-based GP

We discuss the derivation of the sampling distribution of g_i and its allocation for $i \neq N$ in the chain-based GP discussed in Section 5.3.4.

As with the exchangeable GP, this derivation can be done by examining the underlying completely random measure or by taking the limits of a finite gamma-Poisson prior with the stochastic process discussed in Section 4.5.2.1 for each column. In this section, let $z_{(-i)}$ stand for $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N$.

Derivation 1 For the i^{th} row, we wish to sample g_i from continuous part of the posterior distribution

$$p(g_i|Z, \alpha) = \int p(g_i|B, Z) dP(B|Z, \alpha)$$

and then allocate these g_i new features to columns. Let $c_{i+(i+1)}$ be $1 - e^{-\kappa(t_i + t_{i+1})}$, the parameter of the transition from z_{i-1} to z_{i+1} if z_i is not observed. Note that $c_{i+(i+1)} = c_i + c_{i+1} - c_i c_{i+1}$. By Equation (4.16), the posterior Lévy measure of the

continuous part of B given $z_{(-i)}$ is

$$\begin{aligned}\nu(d\omega, dp) &= cp^{-1}e^{-\left(c+c_{i+(i+1)}+\sum_{j=1:j\notin\{i,i+1\}}^{N} c_j\right)p}dpB_0(d\omega) \\ &= cp^{-1}e^{-\left(c-c_ic_{i+1}+\sum_{j=1}^{N} c_j\right)p}dpB_0(d\omega).\end{aligned}$$

Therefore, using Equation (2.14), ignoring the discrete part of B and plugging in $B_0(d\omega) = \alpha$, we have that

$$B(\Omega)|Z, \alpha \sim \text{Gamma}\left(c\alpha, c - c_ic_{i+1} + \sum_{j=1}^{N} c_j\right).$$

Using the transition probabilities for the cIGPFM defined in Equation (4.15), the probability of observing z_{ik} given that $z_{(-i)}$ are all zero and p_k is

$$\begin{aligned}p(z_{ik}|z_{(-i)k} = 0, p_k) &= \frac{p(z_{(i+1)k} = 0|z_{ik}, p_k)p(z_{ik}|z_{(i-1)k} = 0, p_k)}{p(z_{i+1} = 0|z_{i-1} = 0, p_k)} \\ &= \frac{(c_ic_{i+1}p_k)^{z_{ik}}}{z_{ik}!}e^{-c_ic_{i+1}p_k}.\end{aligned}$$

So $p(z_{ik}|z_{(-i)k} = 0, p_k) \sim \text{Poisson}(c_ic_{i+1}p_k)$. Using the summation property of Poisson variables, this means that

$$g_i|B(\Omega) \sim \text{Poisson}\left(c_ic_{i+1}B(\Omega)\right).$$

Therefore, again using the fact that if $x|\lambda \sim \text{Poisson}(c\lambda)$ and $\lambda \sim \text{Gamma}(a, b)$, then by marginalizing out λ, $x \sim \text{NB}\left(a, \frac{b}{b+c}\right)$, we get that

$$g_i|Z, \alpha \sim \text{NB}\left(c\alpha, \frac{c - c_ic_{i+1} + \sum_{j=1}^{N} c_j}{c + \sum_{j=1}^{N} c_j}\right),$$

and plugging in $c = 1$,

$$g_i|Z, \alpha \sim \text{NB}\left(\alpha, \frac{1 - c_ic_{i+1} + \sum_{j=1}^{N} c_j}{1 + \sum_{j=1}^{N} c_j}\right),$$

which agrees with Equation (5.6).

Since g_i is the result of a thinned Poisson$(B(\Omega))$, the allocation of $z_i^{\text{new}}|g_i$ must be the result of a thinned CRP, which is itself a CRP. We weight each partition according to the cardinality of that partition, which gives us Ewens distribution.

Derivation 2 As was done in Derivation 2 in Section 4.A.2, we start with the finite approximation in which for each k, we draw $\lambda_k \sim \text{Gamma}(\alpha/K, 1)$ and then draw $\{z_{ik}\}_{i=1}^N$ from the stochastic process from Section 4.5.2.1. In this prior, we must find ξ_i such that $p(z_{ik}|z_{(-i)k} = 0, \alpha) \sim \text{NB}(\alpha/K, \xi_i)$. If we can do this, then we showed in Section 4.A.2 that in the infinite limit, $g_i \sim \text{NB}(\alpha, \xi_i)$ and that $z_i^{\text{new}}|g_i$ is distributed according to Ewens distribution.

So all we must do is show that $\xi_i = \frac{1-c_i c_{i+1}+\sum_{j=1}^N c_j}{1+\sum_{j=1}^N c_j}$ and we will be done with our derivation. We start by noting

$$\begin{aligned} p(z_{ik}|z_{(-i)k} = 0, \alpha) &\propto \int_0^\infty p(z_{ik}, z_{(-i)k} = 0|\lambda_k)p(\lambda_k|\alpha)d\lambda_k \\ &= \int_0^\infty p(\lambda_k|\alpha)\prod_{j=1}^N p(z_{jk}|z_{(j-1)k}, \lambda_k)d\lambda_k. \end{aligned}$$

The terms $p(z_{jk}|z_{(j-1)k}, \lambda_k)$ break up into three cases:

1. $j = i$ In this case, we know that since $z_{(i-1)k} = 0$, z_{ik} is distributed Poisson$(\lambda_k c_i)$.

$$\begin{aligned} p(z_{ik}|z_{(i-1)k=0}, \lambda_k) &\sim \text{Poisson}(z_{ik}; \lambda_k c_i) \\ &= \frac{e^{-\lambda_k c_i}(\lambda_k c_i)^{z_{ik}}}{z_{ik}!} \end{aligned}$$

2. $j = i + 1$ In this case, since $z_{(i+1)k} = 0$, we know that all z_{ik} elements died between z_{ik} and $z_{(i+1)k}$ and that no new ones were born and survived.

$$\begin{aligned} p(z_{(i+1)k} = 0|z_{ik}, \lambda_k) &= p(z_{(i+1)k} = 0|y_{(i+1)k} = 0, \lambda_k)p(y_{(i+1)k} = 0|z_{ik}, \lambda_k) \\ &\sim \text{Poisson}(0; \lambda_k c_{i+1})\text{Binomial}(0; z_{ik}, 1 - c_{i+1}) \\ &= e^{-\lambda_k c_{i+1}} c_{i+1}^{z_{ik}} \end{aligned}$$

3. $j \neq i$ and $j \neq i + 1$ We know that both $z_{jk} = 0$ and $z_{(j-1)k} = 0$, so

$$\begin{aligned} p(z_{jk} = 0|z_{(j-1)k} = 0, \lambda_k) &\sim \text{Poisson}(0; \lambda_k c_j) \\ &= e^{-\lambda_k c_j} \end{aligned}$$

We can therefore calculate $p(z_{ik}|z_{(-i)k} = 0, \alpha)$ as follows

$$\begin{aligned}
&p(z_{ik}|z_{(-i)k} = 0, \alpha) \\
&\quad\propto \int_0^\infty p(z_{(i+1)k} = 0|z_{ik}, \lambda_k) p(z_{ik}|z_{(i-1)k=0}, \lambda_k) p(\lambda_k|\alpha) \prod_{j \notin \{i,i+1\}} p(z_{jk}|z_{(j-1)k}, \lambda_k) d\lambda_k \\
&\quad= \int_0^\infty e^{-\lambda_k c_{i+1}} c_{i+1}^{z_{ik}} \frac{e^{-\lambda_k c_i} (\lambda_k c_i)^{z_{ik}}}{z_{ik}!} \frac{\lambda_k^{\alpha/K-1} e^{-\lambda_k}}{\Gamma(\alpha/K)} \prod_{j \notin \{i,i+1\}} e^{-\lambda_k c_j} d\lambda_k \\
&\quad= \int_0^\infty \underbrace{\frac{e^{-\lambda_k c_{i+1} c_i} (\lambda_k c_{i+1} c_i)^{z_{ik}}}{z_{ik}!}}_{=\text{Poisson}(z_{ik}; \lambda_k c_{i+1} c_i)} \underbrace{\frac{\lambda_k^{\alpha/K-1}}{\Gamma(\alpha/K)} \exp\left(-\lambda_k \left(1 - c_{i+1} c_i + \sum_{j=1}^N c_j\right)\right)}_{\propto \text{Gamma}\left(\lambda_k; \alpha/K, 1 - c_{i+1} c_i + \sum_{j=1}^N c_j\right)} d\lambda_k \\
&\quad\propto NB\left(\frac{\alpha}{K}, \frac{1 - c_{i+1} c_i + \sum_{j=1}^N c_j}{1 + \sum_{j=1}^N c_j}\right).
\end{aligned}$$

So for $i \neq N$,

$$\xi_i = \frac{1 - c_{i+1} c_i + \sum_{j=1}^N c_j}{1 + \sum_{j=1}^N c_j},$$

as desired.

Chapter 6

Applications

There have been many applications of the BP including modeling protein interactions (Chu et al., 2006), binary matrix factorization (Meeds et al., 2007), learning features for similarity judgement (Navarro and Griffiths, 2007), sparse ICA and factor analysis (Knowles and Ghahramani, 2007; Rai and Daumé, 2008), bipartite graph learning (Wood et al., 2006), hidden Markov models (Van Gael et al., 2009; Fox et al., 2009), and topic models (Williamson et al., 2010b).

The applications of the GP to latent feature models have only included the original toy motivation in (Titsias, 2008) and matrix factorization (Hoffman et al., 2010).

One important commonality of all these applications is that the likelihood is insensitive to the number of all-zero columns. This allows us to use nonparametric priors in which we only need to keep track of the non-zero columns.

In this section, we discuss three applications that we have worked on. In Section 6.1, we introduce the application of link prediction using the IBP. This work was originally published in Miller et al. (2009). In Section 6.2, we demonstrate how to use the pIBP introduced in Section 4.4.1 for choice models, thereby extending the work of Görür et al. (2006). This was originally published in Miller et al. (2008a). Finally, in Section 6.3, we present unpublished work on applying the cIBP and cIGPFM to human genomic data in the form of copy number variations.

6.1 Relational Models

As the availability and importance of relational data—such as the friendships summarized on a social networking website—increases, it becomes increasingly important to have good models for such data. The kinds of latent structure that have been considered for use in predicting links in such networks have been relatively limited. In particular, the machine learning community has focused on latent class models,

adapting Bayesian nonparametric methods to jointly infer how many latent classes there are while learning which entities belong to each class. We pursue a similar approach with a richer kind of latent variable—latent features—using a Bayesian nonparametric approach to simultaneously infer the number of features at the same time we learn which entities have each feature. Our model combines these inferred features with known covariates in order to perform link prediction. We demonstrate that the greater expressiveness of this approach allows us to improve performance on three data sets.

6.1.1 Introduction

Statistical analysis of social networks and other relational data has been an active area of research for over seventy years and is becoming an increasingly important problem as the scope and availability of social network data sets increase (Wasserman and Faust, 1994). In these problems, we observe the interactions between a set of entities and we wish to extract informative representations that are useful for making predictions about the entities and their relationships. One basic challenge is link prediction, where we observe the relationships (or "links") between some pairs of entities in a network (or "graph") and we try to predict unobserved links. For example, in a social network, we might only know some subset of people are friends and some are not, and seek to predict which other people are likely to get along.

Our goal is to improve the expressiveness and performance of generative models based on extracting latent structure representing the properties of individual entities from the observed data, so we will focus on these kinds of models. This rules out approaches like the popular p^* model that uses global quantities of the graph, such as how many edges or triangles are present (Wasserman and Pattison, 1996; Robins et al., 2007). Of the approaches that do link prediction based on attributes of the individual entities, these can largely be classified into class-based and feature-based approaches. There are many models that can be placed under these approaches, so we will focus on the models that are most comparable to our approach.

Most generative models using a class-based representation are based on the stochastic blockmodel, introduced by Wang and Wong (1987) and further developed by Nowicki and Snijders (2001). In the most basic form of the model, we assume there are a finite number of classes that entities can belong to and that these classes entirely determine the structure of the graph, with the probability of a link existing between two entities depending only on the classes of those entities. In general, these classes are unobserved, and inference reduces to assigning entities to classes and inferring the class interactions. One of the important issues that arise in working with this model is determining how many latent classes there are for a given problem. The Infinite Relational Model (IRM) (Kemp et al., 2006) used methods from nonparametric Bayesian

statistics to tackle this problem, allowing the number of classes to be determined at inference time. The Infinite Hidden Relational Model (Xu et al., 2006) further elaborated on this model and the Mixed Membership Stochastic Blockmodel (MMSB) (Airoldi et al., 2009) extended it to allow entities to have mixed memberships.

All these class-based models share a basic limitation in the kinds of relational structure they naturally capture. For example, in a social network, we might find a class which contains "male high school athletes" and another which contains "male high school musicians." We might believe these two classes will behave similarly, but with a class-based model, our options are to either merge the classes or duplicate our knowledge about common aspects of them. In a similar vein, with a limited amount of data, it might be reasonable to combine these into a single class "male high school students," but with more data we would want to split this group into athletes and musicians. For every new attribute like this that we add, the number of classes would potentially double, quickly leading to an overabundance of classes. In addition, if someone is both an athlete and a musician, we would either have to add another class for that or use a mixed membership model, which would say that the more a student is an athlete, the less he is a musician.

An alternative approach that addresses this problem is to use features to describe the entities. There could be a separate feature for "high school student," "male," "athlete," and "musician" and the presence or absence of each of these features is what defines each person and determines their relationships. One class of latent-feature models for social networks has been developed by Hoff et al. (2002) and Hoff (2005, 2008), who proposed real-valued vectors as latent representations of the entities in the network where depending on the model, either the distance, inner product, or weighted combination of the vectors corresponding to two entities affects the likelihood of there being a link between them. However, extending our high school student example, we might hope that instead of having arbitrary real-valued features (which are still useful for visualization), we would infer binary features where each feature could correspond to an attribute like "male" or "athlete." Continuing our earlier example, if we had a limited amount of data, we might not pick up on a feature like "athlete." However, as we observe more interactions, this could emerge as a clear feature. Instead of doubling the numbers of classes in our model, we add an additional feature. Determining the number of features will therefore be of extreme importance.

In this section, we present the *nonparametric latent feature relational model*, a Bayesian nonparametric model in which each entity has binary-valued latent features that influences its relations. In addition, the relations depend on a set of known covariates. This model allows us to simultaneously infer how many latent features there are while at the same time inferring what features each entity has and how those features influence the observations. This model is strictly more expressive than

the stochastic blockmodel. In Section 6.1.2, we describe a simplified version of our model and then the full model. In Section 6.1.3, we discuss how to perform inference. In Section 6.1.4, we illustrate the properties of our model using synthetic data and then show that the greater expressiveness of the latent feature representation results in improved link prediction on three real data sets.

6.1.2 The nonparametric latent feature relational model

Assume we observe the directed relational links between a set of N entities. Let Y be the $N \times N$ binary matrix that contains these links. That is, let $y_{ij} \equiv Y(i,j) = 1$ if we observe a link from entity i to entity j in that relation and $y_{ij} = 0$ if we observe that there is not a link. Unobserved links are left unfilled. Our goal will be to learn a model from the observed links such that we can predict the values of the unfilled entries.

6.1.2.1 Basic model

In our basic model, each entity is described by a set of binary features. We are not given these features a priori and will attempt to infer them. We assume that the probability of having a link from one entity to another is entirely determined by the combined effect of all pairwise feature interactions. If there are K features, then let Z be the $N \times K$ binary matrix where each row corresponds to an entity and each column corresponds to a feature such that $z_{ik} \equiv Z(i,k) = 1$ if the i^{th} entity has feature k and $z_{ik} = 0$ otherwise. and let Z_i denote the feature vector corresponding to entity i. Let W be a $K \times K$ real-valued weight matrix where $w_{kk'} \equiv W(k,k')$ is the weight that affects the probability of there being a link from entity i to entity j if both entity i has feature k and entity j has feature k'.

We assume that links are independent conditioned on Z and W, and that only the features of entities i and j influence the probability of a link between those entities. This defines the likelihood

$$\Pr(Y|Z,W) \quad = \quad \prod_{i,j} \Pr(y_{ij}|Z_i, Z_j, W) \tag{6.1}$$

where the product ranges over all pairs of entities. Given the feature matrix Z and weight matrix W, the probability that there is a link from entity i to entity j is

$$\Pr(y_{ij} = 1|Z,W) = \sigma\left(Z_i W Z_j^\top\right) = \sigma\left(\sum_{k,k'} z_{ik} z_{jk'} w_{kk'}\right) \tag{6.2}$$

where $\sigma(\cdot)$ is a function that transforms values on $(-\infty, \infty)$ to $(0,1)$ such as the

sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ or the probit function $\sigma(x) = \Phi(x)$. An important aspect of this model is that all-zero columns of Z do not affect the likelihood. We will take advantage of this in Section 6.1.2.2.

This model is very flexible. With a single feature per entity, it is equivalent to a stochastic blockmodel. However, since entities can have more than a single feature, the model is more expressive. In the high school student example, each feature can correspond to an attribute like "male," "musician," and "athlete." If we were looking at the relation "friend of" (not necessarily symmetric!), then the weight at the (athlete, musician) entry of W would correspond to the weight that an athlete would be a friend of a musician. A positive weight would correspond to an increased probability, a negative weight a decreased probability, and a zero weight would indicate that there is no correlation between those two features and the observed relation. The more positively correlated features people have, the more likely they are to be friends. Another advantage of this representation is that if our data contained observations of students in two distant locations, we could have a geographic feature for the different locations. While other features such as "athlete" or "musician" might indicate that one person could be a friend of another, the geographic features could have extremely negative weights so that people who live far from each other are less likely to be friends. However, the parameters for the non-geographic features would still be tied for all people, allowing us to make stronger inferences about how they influence the relations. Class-based models would need an abundance of classes to capture these effects and would not have the same kind of parameter sharing.

Given the full set of observations Y, we wish to infer the posterior distribution of the feature matrix Z and the weights W. We do this using Bayes' theorem, $p(Z, W|Y) \propto p(Y|Z, W)p(Z)p(W)$, where we have placed an independent prior on Z and W. Without any prior knowledge about the features or their weights, a natural prior for W involves placing an independent $N(0, \sigma_w^2)$ prior on each w_{ij}. However, placing a prior on Z is more challenging. If we knew how many features there were, we could place an arbitrary parametric prior on Z. However, we wish to have a flexible prior that allows us to simultaneously infer the number of features at the same time we infer all the entries in Z. The Indian Buffet Process is such a prior.

6.1.2.2 The Indian Buffet Process and the basic generative model

As mentioned in the previous section, any features which are all-zero do not affect the likelihood. That means that even if we added an infinite number of all-zero features, the likelihood would remain the same. The Indian Buffet Process (IBP) (Griffiths and Ghahramani, 2006) is a prior on infinite binary matrices such that with probability one, a feature matrix drawn from it for a finite number of entities will only have a

finite number of non-zero features. Moreover, any feature matrix, no matter how many non-zero features it contains, has positive probability under the IBP prior. It is therefore a useful nonparametric prior to place on our latent feature matrix Z.

The generative process to sample matrices from the IBP can be described through a culinary metaphor that gave the IBP its name. In this metaphor, each row of Z corresponds to a diner at an Indian buffet and each column corresponds to a dish at the infinitely long buffet. If a customer takes a particular dish, then the entry that corresponds to the customer's row and the dish's column is a one and the entry is zero otherwise. The culinary metaphor describes how people choose the dishes. In the IBP, the first customer chooses a Poisson(α) number of dishes to sample, where α is a parameter of the IBP. The i^{th} customer tries each previously sampled dish with probability proportional to the number of people that have already tried the dish and then samples a Poisson(α/i) number of new dishes. This process is exchangeable, which means that the order in which the customers enter the restaurant does not affect the configuration of the dishes that people try (up to permutations of the dishes as described by Griffiths and Ghahramani (2006)). This insight leads to a straightforward Gibbs sampler to do posterior inference that we describe in Section 6.1.3.

Using an IBP prior on Z, our basic generative latent feature relational model is:

$$
\begin{aligned}
Z &\sim \text{IBP}(\alpha) \\
w_{kk'} &\sim \mathcal{N}(0, \sigma_w^2) && \text{for all } k, k' \text{ for which features } k \text{ and } k' \text{ are non-zero} \\
y_{ij} &\sim \sigma\left(Z_i W Z_j^\top\right) && \text{for each observation.}
\end{aligned}
$$

6.1.2.3 Full nonparametric latent feature relational model

We have described the basic nonparametric latent feature relational model. We now combine it with ideas from the social network community to get our full model. First, we note that there are many instances of logit models used in statistical network analysis that make use of covariates in link prediction (Wasserman and Pattison, 1996). Here we will focus on a subset of ideas discussed in (Hoff, 2005). Let X_{ij} be a vector that influences the relation y_{ij}, let $X_{p,i}$ be a vector of known attributes of entity i when it is the parent of a link, and let $X_{c,i}$ be a vector of known attributes of entity i when it is a child of a link. For example, in Section 6.1.4.2, when Y represents relationships amongst countries, X_{ij} is a scalar representing the geographic similarity between countries ($X_{ij} = \exp(-d(i,j))$) since this could influence the relationships and $X_{p,i} = X_{c,i}$ is a set of known features associated with each country ($X_{p,i}$ and $X_{c,i}$ would be distinct if we had covariates specific to each country's roles). We then let c be a normally distributed scalar and β, β_p, β_c, a, and b be normally distributed

vectors in our full model in which

$$\Pr(y_{ij} = 1|Z, W, X, \beta, a, b, c) = \sigma\left(Z_i W Z_j^\top + \beta^\top X_{ij} + (\beta_p^\top X_{p,i} + a_i) + (\beta_c^\top X_{c,j} + b_j) + c\right). \tag{6.3}$$

If we do not have information about one or all of X, X_p, and X_c, we drop the corresponding term(s). In this model, c is a global offset that affects the default likelihood of a relation and a_i and b_j are entity and role specific offsets.

So far, we have only considered the case of observing a single relation. It is not uncommon to observe multiple relations for the same set of entities. For example, in addition to the "friend of" relation, we might also observe the "admires" and "collaborates with" relations. We still believe that each entity has a single set of features that determines all its relations, but these features will not affect each relation in the same way. If we are given m relations, label them $Y^1, Y^2, \ldots, Y^m$. We will use the same features for each relation, but we will use an independent weight matrix W^i for each relation Y^i. In addition, covariates might be relation specific or common across all relations. Regardless, they will interact in different ways in each relation. Our full model is now

$$\Pr(Y^1, \ldots, Y^m|Z, \{W^i, X^i, \beta^i, a^i, b^i, c^i\}_{i=1}^m) \;=\; \prod_{i=1}^m \Pr(Y^i|Z, W^i, X^i, \beta^i, a^i, b^i, c^i).$$

6.1.2.4 Variations of the nonparametric latent feature relational model

The model that we have defined is for directed graphs in which the matrix Y^i is not assumed to be symmetric. For undirected graphs, we would like to define a symmetric model. This is easy to do by restricting W^i to be symmetric. If we further believe that the features we learn should not interact, we can assume that W^i is diagonal.

6.1.2.5 Related nonparametric latent feature models

There are two models related to our nonparametric latent feature relational model that both use the IBP as a prior on binary latent feature matrices. The most closely related model is the Binary Matrix Factorization (BMF) model of Meeds et al. (2007). The BMF is a general model with several concrete variants, the most relevant of which was used to predict unobserved entries of binary matrices for image reconstruction and collaborative filtering. If Y is the observed part of a binary matrix, then in this variant, we assume that $Y|U, V, W \sim \sigma(UWV^\top)$ where $\sigma(\cdot)$ is the logistic function, U and V are independent binary matrices drawn from the IBP, and the entries in W are independent draws from a normal distribution. If Y is an $N \times N$ matrix where we assume the rows and columns have the same features (i.e., $U = V$), then this special case of their model is equivalent to our basic (covariate-free) model. While Meeds

et al. (2007) were interested in a more general formalization that is applicable to other tasks, we have specialized and extended this model for the task of link prediction. The other related model is the ADCLUS model (Navarro and Griffiths, 2008). This model assumes we are given a symmetric matrix of non-negative similarities Y and that $Y = ZWZ^\top + \epsilon$ where Z is drawn from the IBP, W is a diagonal matrix with entries independently drawn from a Gamma distribution, and ϵ is independent Gaussian noise. This model does not allow for arbitrary feature interactions nor does it allow for negative feature correlations.

6.1.3 Inference Algorithms

Exact inference in our nonparametric latent feature relational model is intractable (Griffiths and Ghahramani, 2006). However, the IBP prior lends itself nicely to approximate inference algorithms via Markov Chain Monte Carlo (Robert and Casella, 2004). We first describe inference in the single relation, basic model, later extending it to the full model. In our basic model, we must do posterior inference on Z and W. Since with probability one, any sample of Z will have a finite number of non-zero entries, we can store just the non-zero columns of each sample of the infinite binary matrix Z. Since we do not have a conjugate prior on W, we must also sample the corresponding entries of W. Our sampler is as follows:

Given W, resample Z We do this by resampling each row Z_i in succession. When sampling entries in the i^{th} row, we use the fact that the IBP is exchangeable to assume that the i^{th} customer in the IBP was the last one to enter the buffet. Therefore, when resampling z_{ik} for non-zero columns k, if m_k is the number of non-zero entries in column k excluding row i, then

$$\Pr(z_{ik} = 1 | Z_{-ik}, W, Y) \quad \propto \quad m_k \Pr(Y | z_{ik} = 1, Z_{-ik}, W).$$

We must also sample z_{ik} for each of the infinitely many all-zero columns to add features to the representation. Here, we use the fact that in the IBP, the prior distribution on the number of new features for the last customer is Poisson(α/N). As described by Griffiths and Ghahramani (2006), we must then weight this by the likelihood term for having that many new features, computing this for $0, 1, \ldots . k_{\max}$ new features for some maximum number of new features $k_{\max}$ and sampling the number of new features from this normalized distribution. The main difficulty arises because we have not sampled the values of W for the all-zero columns and we do not have a conjugate prior on W, so we cannot compute the likelihood term exactly. We can adopt one of the non-conjugate sampling approaches from the Dirichlet process (Neal, 2000) to this task or use the suggestion by Meeds et al. (2007) to include a Metropolis-Hastings step to propose and either accept or reject some number of new columns and the

corresponding weights. We chose to use a stochastic Monte Carlo approximation of the likelihood. Once the number of new features is sampled, we must sample the new values in W as described below.

Given Z, resample W We sequentially resample each of the weights in W that correspond to non-zero features and drop all weights that correspond to all-zero features. Since we do not have a conjugate prior on W, we cannot directly sample W from its posterior. If $\sigma(\cdot)$ is the probit, we adapt the auxiliary sampling trick from Albert and Chib (1993) to have a Gibbs sampler for the entries of W. If $\sigma(\cdot)$ is the logistic function, no such trick exists and we resort to using a Metropolis-Hastings step for each weight in which we propose a new weight from a normal distribution centered around the old one.

Hyperparameters We can also place conjugate priors on the hyperparameters α and σ_w and perform posterior inference on them. We use the approach from (Görür et al., 2006) for sampling of α.

Multiple relations In the case of multiple relations, we can sample W_i given Z independently for each i as above. However, when we resample Z, we must compute

$$\Pr(z_{ik} = 1 | Z_{-ik}, \{W, Y\}_{i=1}^{m}) \quad \propto \quad m_k \prod_{i=1}^{m} \Pr(Y^i | z_{ik} = 1, Z_{-ik}, W^i).$$

Full model In the full model, we must also update $\{\beta^i, \beta_p^i, \beta_c^i, a^i, b^i, c^i\}_{i=1}^m$. By conditioning on these, the update equations for Z and W^i take the same form, but with Equation (6.3) used for the likelihood. When we condition on Z and W^i, the posterior updates for $(\beta^i, \beta_p^i, \beta_c^i, a^i, b^i, c^i)$ are independent and can be derived from the updates in (Hoff, 2005).

Implementation details Despite the ease of writing down the sampler, samplers for the IBP often mix slowly due to the extremely large state space full of local optima. Even if we limited Z to have K columns, there are 2^{NK} potential feature matrices. In an effort to explore the space better, we can augment the Gibbs sampler for Z by introducing split-merge style moves as described in Meeds et al. (2007) as well as perform annealing or tempering to smooth out the likelihood. However, we found that the most significant improvement came from using a good initialization. A key insight that was mentioned in Section 6.1.2.1 is that the stochastic blockmodel is a special case of our model in which each entity only has a single feature. Stochastic

blockmodels have been shown to perform well for statistical network analysis, so they seem like a reasonable way to initialize the feature matrix. In the results section, we compare the performance of a random initialization to one in which Z is initialized with a matrix learned by the Infinite Relational Model (IRM). To get our initialization point, we ran the Gibbs sampler for the IRM for only 15 iterations and used the resulting class assignments to seed Z.

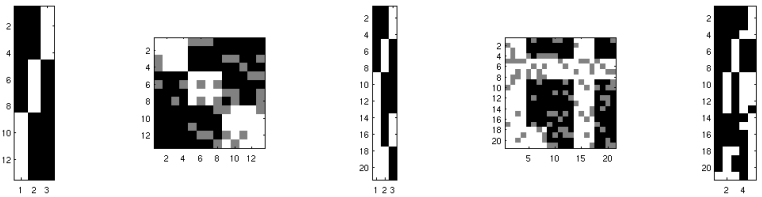
6.1.4 Results

We first qualitatively analyze the strengths and weaknesses of our model on synthetic data, establishing what we can and cannot expect from it. We then compare our model against two class-based generative models, the Infinite Relational Model (IRM) (Kemp et al., 2006) and the Mixed Membership Stochastic Blockmodel (MMSB) (Airoldi et al., 2009), on two data sets from the original IRM paper and a NIPS coauthorship data set, establishing that our model does better than the best of those models on those data sets.

6.1.4.1 Synthetic data

We first focus on the qualitative performance of our model. We applied the basic model to two very simple synthetic data sets generated from known features. These data sets were simple enough that the basic model could attain 100% accuracy on held-out data, but were different enough to address the qualitative characteristics of the latent features inferred. In one data set, the features were the class-based features seen in Figure 6.1(a) and in the other, we used the features in Figure 6.1(c). The observations derived from these features can be seen in Figure 6.1(b) and Figure 6.1(d), respectively.

On both data sets, we initialized Z and W randomly. With the very simple, class-based model, 50% of the sampled feature matrices were identical to the generating feature matrix with another 25% differing by a single bit. However, on the other data set, only 25% of the samples were at most a single bit different than the true matrix. It is not the case that the other 75% of the samples were bad samples, though. A randomly chosen sample of Z is shown in Figure 6.1(e). Though this matrix is different from the true generating features, with the appropriate weight matrix it predicts just as well as the true feature matrix. These tests show that while our latent feature approach is able to learn features that explain the data well, due to subtle interactions between sets of features and weights, the features themselves will not in general correspond to interpretable features. However, we can expect the inferred features to do a good job explaining the data. This also indicates that there are many local optima in the feature space, further motivating the need for good



(a) (b) (c) (d) (e)

Figure 6.1: Features and corresponding observations for synthetic data. In (a), we show features that could be explained by a latent-class model that then produces the observation matrix in (b). White indicates one values, black indicates zero values, and gray indicates held out values. In (c), we show the feature matrix of our other synthetic data set along with the corresponding observations in (d). (e) shows the feature matrix of a randomly chosen sample from our Gibbs sampler.

initialization.

6.1.4.2 Multi-relational data sets

In the original IRM paper, the IRM was applied to several data sets (Kemp et al., 2006). These include a data set containing 54 relations of 14 countries (such as "exports to" and "protests") along with 90 given features of the countries (Rummel, 1999) and a data set containing 26 kinship relationships of 104 people in the Alyawarra tribe in Central Australia (Denham, 1973). See Kemp et al. (2006), Rummel (1999), and Denham (1973) for more details on the data sets.

Our goal in applying the latent feature relational model to these data sets was to demonstrate the effectiveness of our algorithm when compared to two established class-based algorithms, the IRM and the MMSB, and to demonstrate the effectiveness of our full algorithm. For the Alyawarra data set, we had no known covariates. For the countries data set, $X_p = X_c$ was the set of known features of the countries and X was the country distance similarity matrix described in Section 6.1.2.3.

As mentioned in the synthetic data section, the inferred features do not necessarily have any interpretable meaning, so we restrict ourselves to a quantitative comparison. For each data set, we held out 20% of the data during training and we report the AUC, the area under the ROC (Receiver Operating Characteristic) curve, for the held-out data (Huang and Ling, 2005). We report results for inferring a global set of features for all relations as described in Section 6.1.2.3 which we refer to as "global" as well as results when a different set of features is independently learned for each relation and then the AUCs of all relations are averaged together, which we refer to as "single." In addition, we tried initializing our sampler for the latent feature relational

model with either a random feature matrix ("LFRM rand") or class-based features from the IRM ("LFRM w/ IRM"). We ran our sampler for 1000 iterations for each configuration using a logistic squashing function (though results using the probit are similar), throwing out the first 200 samples as burn-in. Each method was given five random restarts.

Table 6.1: AUC on the countries and kinship data sets. Bold identifies the best performance.

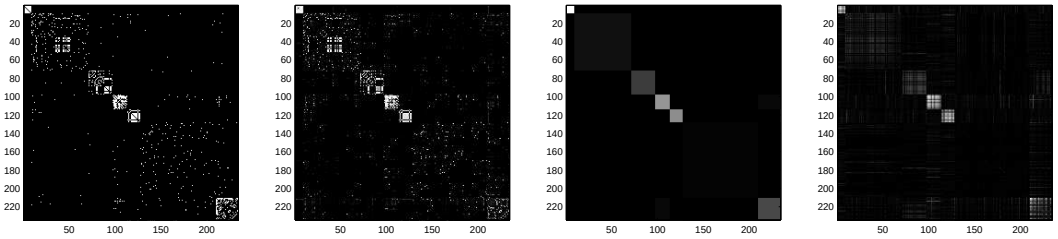
	Countries single	Countries global	Alyawarra single	Alyawarra global
LFRM w/ IRM	0.8521 ± 0.0035	**0.8772** ± 0.0075	0.9346 ± 0.0013	**0.9183** ± 0.0108
LFRM rand	**0.8529** ± 0.0037	0.7067 ± 0.0534	**0.9443** ± 0.0018	0.7127 ± 0.030
IRM	0.8423 ± 0.0034	0.8500 ± 0.0033	0.9310 ± 0.0023	0.8943 ± 0.0300
MMSB	0.8212 ± 0.0032	0.8643 ± 0.0077	0.9005 ± 0.0022	0.9143 ± 0.0097

Results of these tests are in Table 6.1. As can be seen, the LFRM with class-based initialization outperforms both the IRM and MMSB. On the individual relations ("single"), the LFRM with random initialization also does well, beating the IRM initialization on both data sets. However, the random initialization does poorly at inferring the global features due to the coupling of features and the weights for each of the relations. This highlights the importance of proper initialization. To demonstrate that the covariates are helping, but that even without them, our model does well, we ran the global LFRM with class-based initialization without covariates on the countries data set and the AUC dropped to 0.8713 ± 0.0105, which is still the best performance.

On the countries data, the latent feature model inferred on average 5-7 features when seeded with the IRM and 8-9 with a random initialization. On the kinship data, it inferred 9-11 features when seeded with the IRM and 13-19 when seeded randomly.

6.1.4.3 Predicting NIPS coauthorship

As our final example, highlighting the expressiveness of the latent feature relational model, we used the coauthorship data from the NIPS data set compiled in (Globerson et al., 2007). This data set contains a list of all papers and authors from NIPS 1-17. We took the 234 authors who had published with the most other people and looked at their coauthorship information. The symmetric coauthor graph can be seen in Figure 6.2(a). We again learned models for the latent feature relational model, the IRM and the MMSB training on 80% of the data and using the remaining 20% as a test set. For the latent feature model, since the coauthorship relationship is symmetric, we learned a full, symmetric weight matrix W as described in Section



(a) True relations (b) Feature predictions (c) IRM predictions (d) MMSB predictions

Figure 6.2: Predictions for all algorithms on the NIPS coauthorship data set. In (a), a white entry means two people wrote a paper together. In (b-d), the lighter an entry, the more likely that algorithm predicted the corresponding people would interact.

6.1.2.4. We did not use any covariates. A visualization of the predictions for each of these algorithms can be seen in Figure 6.2(b-d). Figure 6.2 really drives home the difference in expressiveness. Stochastic blockmodels are required to group authors into classes, and assumes that all members of classes interact similarly. For visualization, we have ordered the authors by the groups the IRM found. These groups can clearly be seen in Figure 6.2(c). The MMSB, by allowing partial membership is not as restrictive. However, on this data set, the IRM outperformed it. The latent feature relational model is the most expressive of the models and is able to much more faithfully reproduce the coauthorship network.

The latent feature relational model also quantitatively outperformed the IRM and MMSB. We again ran our sampler for 1000 samples initializing with either a random feature matrix or a class-based feature matrix from the IRM and reported the AUC on the held-out data. Using five restarts for each method, the LFRM w/ IRM performed best with an AUC of 0.9509, the LFRM rand was next with 0.9466 and much lower were the IRM at 0.8906 and the MMSB at 0.8705 (all at most ± 0.013). On average, the latent feature relational model inferred 20-22 features when initialized with the IRM and 38-44 features when initialized randomly.

6.2 Tree-Structured Choice Models

Choice models play important roles in both econometrics (McFadden, 2001) and cognitive psychology (Luce, 1959). They describe what happens when people are given two or more options and are asked to choose one of them. In this section, we will restrict our attention to choices between pairs of objects, though the methods presented here can be applied more generally.

Even in the simple case of binary decisions, people's choices are not deterministic. The Elimination By Aspects (EBA) model is a popular attempt to explain this variation (Tversky, 1972). EBA hypothesizes that choices are based on a weighted combination of the features of objects. Keeping our earlier notation, let Z be a feature matrix where $z_{ik} = 1$ if the i^{th} object has the k^{th} feature and $z_{ik} = 0$ otherwise. For each of the features, there is a corresponding weight w_k. The higher the weight, the more influence that feature has. The EBA model defines the probability of choosing object i over object j as

$$p_{ij} = \frac{\sum_k w_k z_{ik}(1 - z_{jk})}{\sum_k w_k z_{ik}(1 - z_{jk}) + \sum_k w_k (1 - z_{ik}) z_{jk}}. \tag{6.4}$$

For comparison with previous results (Görür et al., 2006) we assume extra noise in people's choices, with $\tilde{p}_{ij} = (1 - \epsilon)p_{ij} + 0.5\epsilon$.

If X is the observed choice matrix where x_{ij} contains how many times object i was chosen over object j, then for any given w and Z, the probability of X is

$$P(X|Z, w) = \prod_{i=1}^{N} \prod_{i<j} \binom{x_{ij} + x_{ji}}{x_{ij}} \tilde{p}_{ij}^{x_{ij}} (1 - \tilde{p}_{ij})^{x_{ji}}. \tag{6.5}$$

If the number of features is known, Wickelmaier and Schmid (2004) showed how to estimate the weight vector and feature matrix. In general, though, the number of features is not known. Therefore, Görür et al. (2006) applied the IBP to this model in order to simultaneously infer the number of features, the feature matrix, and the weights of these features, and obtained improved performance over previous models.

In an influential paper, Tversky and Sattath (1979) introduced the preference tree model as an extension of EBA. This model is applicable if the relationships of objects can be captured in a tree structure. In preference trees, each feature has to strictly obey the tree structure. That is, if two objects share a common feature, then all descendants of their most recent common ancestor must have that feature. In some situations, this tree structure may either be known in advance or a good working hypothesis may be available. An example can be found in an experiment reported by Rumelhart and Greeno (1971), in which subjects made 36 pairwise choices of who among a group of nine "well-known personalities" they would like to spend time with. The nine personalities consisted of three politicians, three athletes and three movie stars. It was therefore hypothesized that the tree structure summarizing the prior beliefs about these personalities was similar to the tree shown in Figure 6.3. In this figure, ℓ is the length of the edge from each general category of people to each individual at the leaf.

Just as the IBP can be used to infer features for EBA, the pIBP defines an

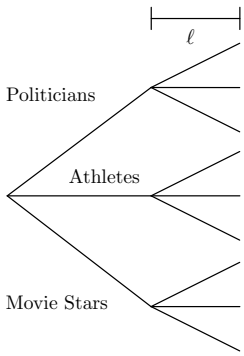


Figure 6.3: Hypothesis about a tree on preferences (Rumelhart and Greeno, 1971).

appropriate prior for the case where features are assumed to follow a tree structure, as in preference trees. The pIBP model for feature generation can be seen as a soft version of preference trees, allowing features to break the tree structure but assigning low probability to these events. Görür et al. (2006) performed a comparison between the IBP as a prior and EBA models with fixed numbers of features as well as a finite preference tree model that was able to use the tree structure. It was shown that EBA with an IBP prior on the feature matrix outperformed all others. As we will show, using the pIBP gives both quantitatively and qualitatively better results in the case where the features are drawn using a tree.

To complete the full specification of the EBA model, we assume that each object has a unique feature as well as an unknown additional number of features that may or may not be shared as shown in Figure 6.4(a). We place a pIBP prior on the nonparametric part of the feature matrix and an independent $\mathcal{G}(1,1)$ prior on each w_k. Inference proceeds as outlined above, with the addition of a Metropolis-Hastings step on w_k. A method similar to that in Görür et al. (2006) was used to deal with the w_k values of new columns.

We generated data from this choice model using the tree from Figure 6.3 with $\ell = 0.1$. The tree induces a "block structure" in the choice matrix, with the correlated features of objects along each branch resulting in similar patterns of choice for those objects. An example feature matrix generated from this model is shown in Figure 6.4(a). The top row shows the feature weights of the corresponding columns; the whiter the feature, the more weight that feature has. The feature matrix, Z, is displayed below where entries that are one are white and zero entries are black. Fifteen such examples were generated. For each of these examples, we computed the true value of choosing object i over j as shown in Figure 6.4(b) where the whiter the $(i,j)^{\text{th}}$ entry, the more likely i is to be chosen over j. Based on these values, we generated data sets with 1, 5, 10, 15, 25, 50, 100, 500, and 1000 choice observations per pair (i,j) following Equation (6.4). An example of an observed matrix X can be

Figure 6.4: Example data demonstrating block structure of the features. (a) True underlying features with corresponding weights in the top row. (b) Underlying probability choice matrix derived from (a) where the lighter the (i, j) entry, the more likely i is to be chosen over j. (c) An example observed choice matrix X drawn from (b) with 5 observations per pair.

seen in Figure 6.4(c) with only 5 observations per pair. The lighter the $(i, j)^{\text{th}}$ entry, the more times i was picked over j.

We used these data to examine the effects of two different factors. First, we wished to show the effect of using the pIBP prior over using the IBP prior as the number of observed choice decisions varied. The use of prior knowledge should always help, but with more observations, the influence of the prior should decrease. Second, we wished to test the effect of varying ℓ in our prior. The three values we tested were $\ell = 0.1$, $\ell = 0.5$, and $\ell = 1.0$. As mentioned in Section 4.4.1.1, the pIBP with $\ell = 1.0$ is the same as the IBP, but does not integrate over p_k analytically.

For each of the nine observation levels on each of the fifteen examples, we performed leave-one-out cross validation with each model. For each model and validation point at each observation level, we ran an MCMC sampler for 3000 iterations from three different random initialization points. The first 1000 samples from each run were discarded as a burn-in period even though all chains appeared to have mixed within 100 iterations. The predictive likelihood was then averaged across every 10^{th} sample for all configurations. These results can be seen in Figure 6.5.

As expected, since the pIBP with $\ell = 0.1$ was using the true tree that generated the data, it was able to beat all other configurations for all numbers of observations except 1000, in which case all algorithms performed similarly. As the number of observations increases, the effect of the prior decreases and the models perform more similarly. The pIBP with $\ell = 0.5$ performs better than the IBP, but not as well as the pIBP with the correct tree. This shows that even without perfect knowledge of the tree structure, by inserting some information into the prior, we are able to outperform algorithms that cannot use the same prior knowledge. We also see that the IBP and pIBP with $\ell = 1.0$ perform nearly identically, so explicitly sampling p_k in the inference algorithm does not influence the results.

In addition to obtaining higher likelihoods, the results from the pIBP were also more concise. In Figure 6.6(a), we show the average feature matrices for the pIBP

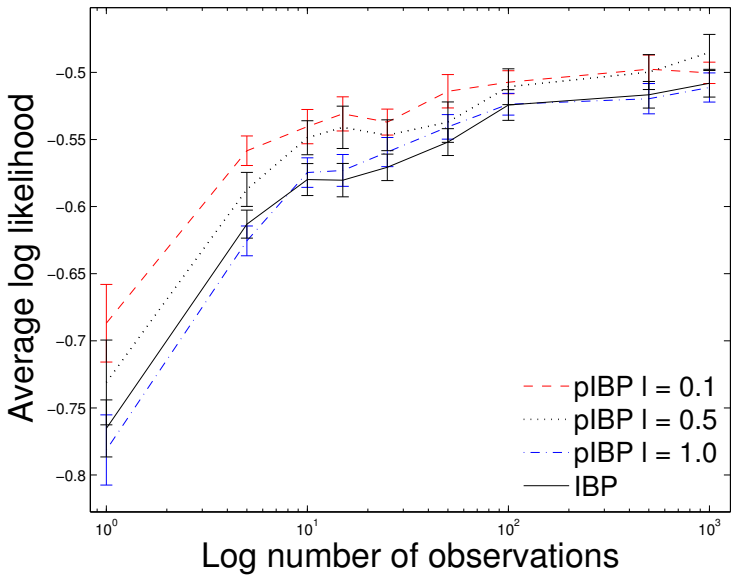


Figure 6.5: Comparison of the average predictive log-likelihood of the IBP and pIBP models with different degrees of prior knowledge along with error bars for choice data. On the x-axis, we vary the (log) number of observations seen for each (i, j) pair.

and IBP when presented with the choice matrix in Figure 6.4(c). In order to obtain these averages, the Z matrices for all samples after the burn-in period were collapsed into equivalent Z matrices in which the weights of all identical columns were summed together. This results in the same probabilities under the EBA model and allows us to average these values across all samples. We also dropped all columns whose weight was below 0.1. As can be seen, the pIBP recovers a feature matrix very similar to the true data while the IBP requires many more features and still does not achieve the same performance. This large number of features necessary to explain the same data was also observed by Görür et al. (2006). Finally, we checked to see how many examples are needed for the choice probability matrix estimated from the samples of Z and W to show the same structure as the true choice probability matrix from Figure 6.4(b). In Figure 6.6(b), we show estimated choice matrices for 1, 10, 100, and 1000 observations per pair. With the pIBP, we observe a block structure immediately, though not all details of the choice matrix are correct. Within very few observations, though, the choice probabilities get close to the true values. In the IBP, we need more than 100 samples before it recovers the block structure.

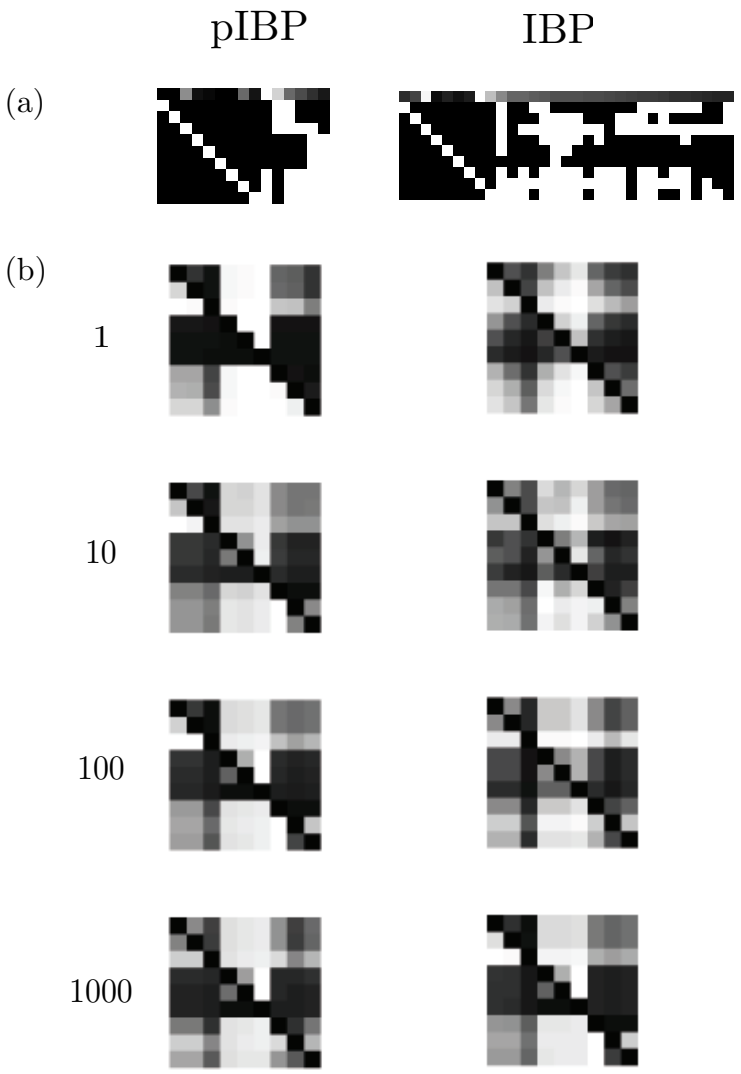


Figure 6.6: Comparison of features inferred and resulting implied block structure from choice data. (a) Mean posterior Z matrices with corresponding weights when presented with data from Figure 6.4. (b) Mean estimated choice matrices P_{ij} for the pIBP and IBP when presented with the different numbers of observations from the data in Figure 6.4.

6.3 Human Genomic Data

In this section, we discuss an application of the cBP and cGP to human genomic data. The particular application we focus on is an attempt at better understanding some of the changes in genomic data that influence cancer in humans.

The Cancer Genome Atlas (TCGA) is a collaborative effort to advance the state of cancer research in the United States lead by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), both at the National Institute of Health (NIH). It is a large-scale operation to coordinate the cancer research of more than 150 researches at dozens of institutions across the country. Genomic data related to both tumors and healthy tissue samples from hundreds of people for twenty different kinds of cancer are in the process of being collected and made available in full to the research community and in an anonymized form to the public at the TCGA website (http://cancergenome.nih.gov/, The Cancer Genome Atlas, 2011). By making all this data available, TCGA aims to significantly accelerate the speed of progress in cancer research.

There are several ways in which genomic variation can influence the formation of tumors. Early studies of DNA focused on the variations observable by microscope. These had to be at least 3 Mb (Megabases, 10^6 nucleotide bases of DNA) long in order to be observed. Later studies of DNA discovered much smaller local variation of the DNA of size smaller than 1 kb as well as SNPs (single-nucleotide polymorphisms), changes in just a single base that have been shown to be very important. More recently, researchers have developed the tools to study variations of intermediate size, between 1 kb and 3 Mb. It has been found that these variations, consisting of inversions, translocations, and copy number variations (CNVs), are as important as SNPs in their effect on the human genome (Feuk et al., 2006; Chin et al., 2011; Freeman et al., 2006).

We focus on the publicly available CNVs of the two forms of cancer in the original 2006 pilot study of TCGA, brain cancer (glioblastoma multiforme, GBM) and ovarian cancer (serous cystadenocarcinoma, OV). CNVs are structural variations consisting of insertions, deletions, and duplications of DNA. The copy number itself is the number of times a particular segment of the DNA can be found. As a simple example, if a particular portion of the genome promotes tumor growth, a higher copy number is generally not good. Similarly, if a section of the DNA inhibits tumor growth, a lower copy number corresponding to a deletion of that section of the genome might result in faster tumor spread. In more complicated examples, particular patterns of interacting copy numbers might give indications of how quickly a tumor will spread as well as what appropriate treatments might be.

The CNV data from TCGA have different levels of preprocessing and smoothing, ranging from level one (least) to level three (most). Each of these levels has

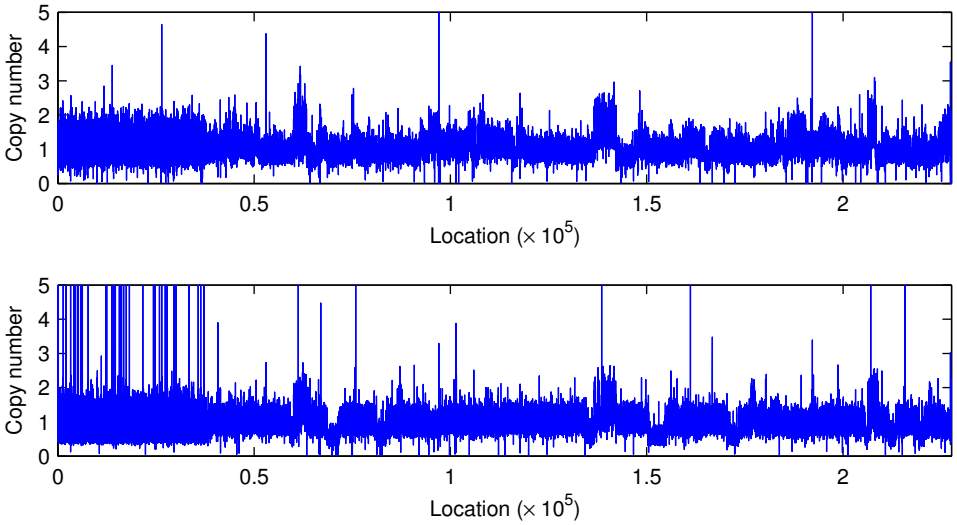


Figure 6.7: Full copy number data for two samples in the GBM data.

some amount of noise that we must take into account when interpreting the data. We use data sets from level two which contain "normalized signals for copy number alterations of aggregated regions." The GBM data was taken from MSKCC HG-CGH-244A and the OV data was taken from HMS HG-CGH-244A. MSKCC and HMS refer to Memorial Sloan-Kettering Cancer Center and Harvard Medical School, the respective locations that collected the data, and HG-CGH-244A refers to Agilent Human Genome CGH Microarray 244A, the platform used for collecting the data. The GBM data contains copy numbers from 476 samples (multiple samples can be from the same person with one or more samples from healthy regions of the body and others from tumors) and the OV data has copy numbers from 204 samples, each with copy numbers taken at 227,612 non-evenly spaced locations. An example of copy number readings from two samples in the GBM data can be found in Figure 6.7. Note the noise as well as the spatial patterns.

Data for all samples at all locations can be found in Figure 6.8. Each row corresponds to a different person and each column corresponds to a particular location. Rows 1–476 are from the GBM data and rows 477–680 are from the OV data. Lighter colors indicate higher count numbers. Note that even though the patterns are different between the GBM data and the OV data, there are some features that show up in both parts of the data.

Our goal will be to develop a method to accurately model this kind of data leveraging the priors for non-exchangeable latent feature models we have developed. Previous work such as (Barnes et al., 2008) has clustered data in order to identify patterns.

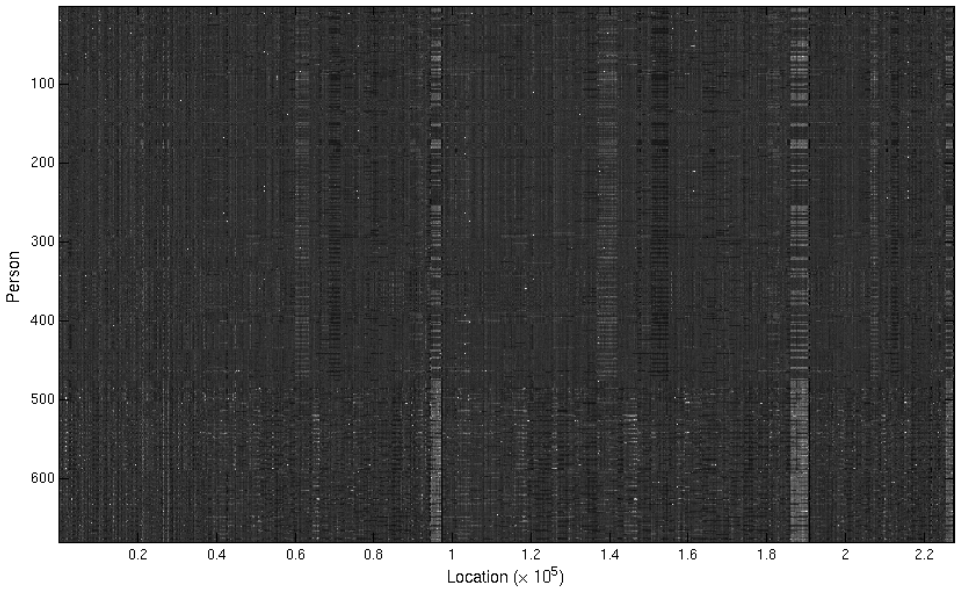


Figure 6.8: Full copy number data for all samples. Samples 1–476 are from the GBM data, samples 477–680 are from the OV data.

Our model will find latent features that explain the patterns in the data. We will show that by using the spatial structure of the data, we are able to explain held-out portions of the data better than if we ignore it.

Given that the underlying data for the copy numbers are count data (the number of times a section of the genome occurs), we develop a non-negative integer factorization of the copy number data that is able to use the spatial dependence of unevenly spaced observations.

Let X be an $N \times D$ matrix where each row corresponds to a sample of D copy number readings as in Figure 6.8. Latent features will be particular patterns of duplications or deletions and any particular sample will consist of a sampling of these features. A priori, we do not know how many of these samples there are, so a nonparametric model is appropriate. Note that the baseline reading of the copy numbers is one, so a deletion can at most reduce the copy number by one to zero, but duplications can be an arbitrary non-negative integer. Therefore, we will model the deletion process and the duplication process separately by subtracting features from a cBP and adding features from a cGP to a baseline of one. Also, we know the spacing of the copy number data and as we can see in Figures 6.7 and 6.8, there is a linear, spatial dependence in the data that the cBP and cGP can use.

Let Z_{GP} be a $D \times K'$ non-negative integer valued matrix containing the duplication

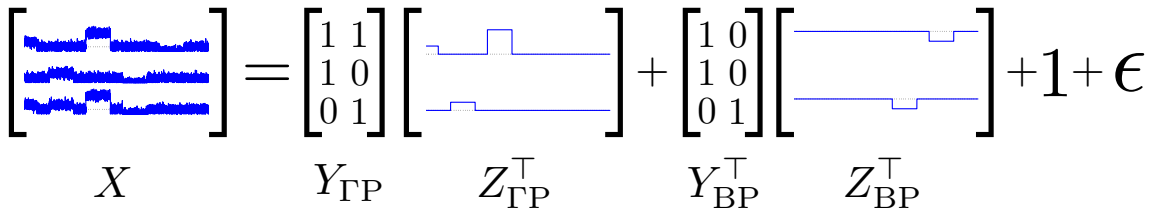


Figure 6.9: Decomposition of X into additive features Z_{GP}, subtractive features $-Z_{\mathrm{BP}}$, the constant $\mathbf{1}$ and the Gaussian noise ϵ. Y_{GP} and Y_{BP} indicate which samples have which features. Since additive features are non-negative integers, we model these using the cGP and since subtractive features must be zero or one, we model these using the cBP.

patterns we wish to infer in which K' is the number of non-zero duplication features and let Z_{BP} be a $D \times K''$ zero-one valued matrix containing the deletion patterns we wish to infer where K'' is the number of non-zero deletion features. We will set the priors

$$\begin{aligned} Z_{\mathrm{GP}} &\sim \mathrm{cIGPFM}(\alpha') \\ Z_{\mathrm{BP}} &\sim \mathrm{cIBP}(\alpha''). \end{aligned}$$

Then each sample will consist of a baseline of one plus some of the features in Z_{GP} minus some of the features in Z_{BP} with additive Gaussian noise. Let Y_{GP} and Y_{BP} be $N \times K'$ and $N \times K''$ zero-one matrices, respectively, indicating which features each data point has. We will sample each entry in Y_{GP} and Y_{BP} independently from a Bernoulli(p) where p is a parameter. If X_i is the i^{th} row of X, $\mathbf{1}$ is the D-vector of all ones, $Y_{\mathrm{GP},i}$ is the i^{th} row of Y_{GP}, $Y_{\mathrm{BP},i}$ is the i^{th} row of Y_{BP}, and σ^2 is the variance of the additive Gaussian noise, this gives us the likelihood

$$p(X|Z_{\mathrm{GP}}, Z_{\mathrm{BP}}, Y_{\mathrm{GP}}, Y_{\mathrm{BP}}, \sigma^2) = \prod_{i=1}^{N} p(X_i|Z_{\mathrm{GP}}, Z_{\mathrm{BP}}, Y_{\mathrm{GP},i}, Y_{\mathrm{BP},i}, \sigma^2)$$

where

$$\prod_{i=1}^{N} p(X_i|Z_{\mathrm{GP}}, Z_{\mathrm{BP}}, Y_{\mathrm{GP},i}, Y_{\mathrm{BP},i}, \sigma^2) \sim \mathcal{N}\left(\mathbf{1} + Y_{\mathrm{GP},i} Z_{\mathrm{GP}}^\top - Y_{\mathrm{BP},i} Z_{\mathrm{BP}}^\top, \sigma^2\right).$$

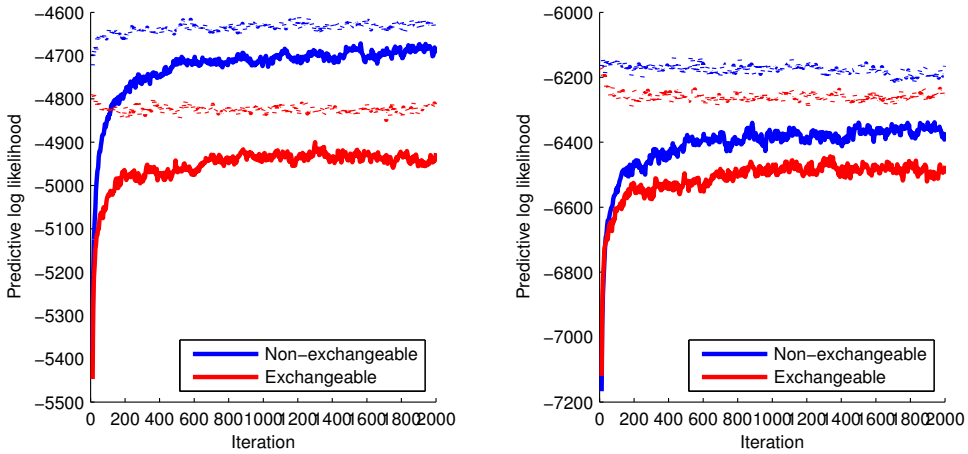
This is shown for a toy example in Figure 6.9.

We ran two experiments on different data sets with this model, one a synthetic data set to verify performance of the model and the other the CNV TCGA data. In both we compared the performance of the non-exchangeable chain-based priors to a fully exchangeable prior in which the two Z matrices are drawn from the exchangeable

IBP and IGPFM. Inference was done via an MCMC sampler.

In the first experiment, we used the model to generate 100 data points at 50 equally spaced locations with $\alpha = 10$, $t_i = 1$ for all i, $\kappa = 0.1$, and $\sigma^2 = 1.5$, meaning moderate spatial dependence with a reasonable amount of noise. We then initialized the samplers for both the exchangeable and non-exchangeable models in one of two ways. In the first initialization, we randomly initialized all features. In the second initialization, we initialized the samplers with the true features that generated the data.

In Figure 6.10, we show the averages of the first $2,000$ iterations of the samplers for both kinds of initializations and both the exchangeable and non-exchangeable models with a five-fold cross-validation. As expected, the non-exchangeable model is able to use the spatial information to infer better features. The noise level was high enough that the spatial information did help as seen by the decline in performance for the exchangeable model when initialized by the true features. When the noise level was much smaller, then both models did equally well since the likelihood was very informative. This just verifies that the model works as expected.



(a) Training predictive log likelihood (b) Test predictive log likelihood

Figure 6.10: Average five-fold training and test predictive log likelihood of the first $2,000$ samples from the exchangeable and non-exchangeable models on synthetic data generated from the non-exchangeable model. The blue lines are from the non-exchangeable models and the red lines are from the exchangeable models. The solid lines are initialized with random features, the dotted lines are initialized with the true features.

We then applied the model to the TCGA copy number data from Figure 6.8 in

which we did not know the true features or any of the true parameters. Unfortunately, inference algorithms for the model does not currently scale to allow D to be on the order of $200,000$, so we broke the data down into pieces of $D = 500$ that would allow us to run $2,000$ iterations of the sampler after $1,000$ iterations of burn-in with five-fold cross validation in three to four hours in Matlab. Rather than run each of the models on the resulting 450 smaller data sets, we randomly chose two segments to test on. Also, instead of doing posterior inference of all the hyperparameters, we did a grid search on a validation set within each fold to find optimal settings of the hyperparameters and did a completely random initialization of the feature matrices. The likelihoods over time are similar to the solid lines in Figure 6.10, so we only discuss the resulting predictive likelihoods.

Table 6.2: Test predictive likelihood on random segments with five-fold cross validation.

	Data 1	Data 2
Non-exchangeable model	-206258.25 $\pm$ 15.30	-210438.47 $\pm$ 12.17
Exchangeable mode	-206643.03 $\pm$ 25.71	-210573.99 $\pm$ 18.55

The results of this experiment can be seen in Table 6.2. The non-exchangeable model is consistently better, though generally by a small amount. We also found that optimal settings for κ varied along the genome, which makes sense from a biological viewpoint.

6.4 Summary

In this chapter, we presented three different applications of Bayesian nonparametric latent feature models. Each was from a different area, showing the versatility of these models. The first application was of the basic exchangeable BP/IBP, demonstrating how this prior can be very useful for the task of latent feature inference for link prediction. This is a promising direction that is still being explored. The second and third applications were to choice data and human genomic data, two applications in which object relationships could be captured by either a tree or chain. These are precisely the scenarios in which our non-exchangeable generalizations are well suited and we demonstrated how the use of this prior information improved performance.

Chapter 7

Conclusion

This dissertation has outlined the three broad areas that people working in the young field of Bayesian nonparametric latent feature models have focused on: extensions and generalizations of the priors, inference algorithms, and applications. We have reviewed the state of the art in all three of these areas and made contributions to each. For priors, we have introduced several non-exchangeable variations of priors for Bayesian nonparametric latent feature models. These variations are suitable in cases in which we know data is either related through a tree or through a chain. Through our work in deriving these priors, we demonstrated how to extend the basic priors and outlined blueprints for how future priors can be developed. For inference algorithms, we have contributed the first variational approximation for the IBP, demonstrating that it can scale better than Gibbs samplers in different data regimes. Finally, for applications, we have introduced three applications including link prediction, tree-structured choice models, and copy numbers from biology.

The field of Bayesian nonparametric latent feature models is still a young field and even though great strides have been made in all three areas of work by us as well as others, there is still much room for improvement. The non-exchangeable generalizations presented here, while more expressive than the original priors, do not capture all non-exchangeable object relationships. As applications demand, more priors and models should be developed and we hope that the ideas, desiderata and various derivations presented here can be used as blueprints for even richer priors. In terms of inference algorithms, until these models scale to much larger data sets, the practical application of these priors will remain limited. The main issue is that even for binary latent feature models, if we truncate the number of features at K, there are still 2^{NK} binary matrices that we must perform inference over. This is an extremely high dimensional discrete space and both MCMC and variational algorithms have difficulty in this space. Current research has started to address this issue, but there is no one algorithm which will work for all problems and therefore, this is still an

active area of research. Finally, there are countless unexplored applications of these models and priors that have yet to be explored. As research progresses in all three of these areas, we will see wider use of these models. Until this happens, though, we must continue to push what we know how to do with these models and priors.

Bibliography

Edoardo M. Airoldi, David M. Blei, Eric P. Xing, and Stephen E. Fienberg. Mixed membership stochastic block models. In D. Koller, Y. Bengio, D. Schuurmans, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS) 21*. Red Hook, NY: Curran Associates, 2009.

James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

David J. Aldous. *Exchangeability and Related Topics*. Springer Lecture Notes in Mathematics. Springer-Verlag, 1983. Lectures from the 13th Summer School on Probability Theory held in Saint-Flour, 1983.

Joseph L. Austerweil and Thomas L. Griffiths. Learning invariant features using the transformed Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Chris Barnes, Vincent Plagnol, Tomas Fitzgerald, Richard Redon, Jonathan Marchini, David Clayton, and Matthew E Hurles. A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40 (10):1245–1252, 2008.

Mathew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, UCL, 2003.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, second edition, 2007.

David Blackwell and James B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

David M. Blei and Michael I. Jordan. Variational methods for the Dirichlet process. In *Proceedings of the International Conference on Machine learning (ICML)*, 2004.

Tamara Broderick, Michael I. Jordan, and Jim Pitman. Beta processes, stick-breaking, and power laws. arXiv:1106.0539v1, 2011.

Lynda Chin, William C. Hahn, Gad Getz, and Matthew Meyerson. Making sense of cancer genomic data. *Genes & Development*, 25:534–555, 2011.

Wei Chu, Zoubin Ghahramani, Roland Krause, and David L. Wild. Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *BIOCOMPUTING 2006: Proceedins of the Pacific Symposium on Biocomputing*, 2006.

Robert B. Cooper. *Introduction to Queueing Theory*. North Holland, second edition, 1981.

Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

Woodrow W. Denham. *The Detection of Patterns in Alyawarra Nonverbal Behavior*. PhD thesis, University of Washington, 1973.

Finale Doshi-Velez and Zoubin Ghahramani. Correlated non-parametric latent feature models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2009a.

Finale Doshi-Velez and Zoubin Ghahramani. Accelerated sampling for the Indian buffet process. In *Proceedings of the International Conference on Machine learning (ICML)*, 2009b.

Finale Doshi-Velez, David Knowles, Shakir Mohamed, and Zoubin Ghahramani. Large scale nonparametric Bayesian inference: Data parallelisation in the Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, 2009a.

Finale Doshi-Velez, Kurt T. Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the Indian buffet process. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009b.

Finale Doshi-Velez, Kurt T. Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the Indian buffet process. Technical report, Department of Engineering, University of Cambridge, 2009c.

François Dufresne, Hans U. Gerber, and Elias S. W. Shiu. Risk theory with the gamma process. *ASTIN Bulletin International Actuarial Association*, 21(2):177–192, 1991.

Rick Durrett. *Probability: Theory and Examples*. Duxbury Press, third edition, 2004.

Warren Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley and Sons, Inc., third edition, 1968.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7:85–97, 2006.

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. Sharing features among dynamic systems with beta processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Jennifer L. Freeman, George H. Perry, Lars Feuk, Richard Redon, Steven A. McCarroll, David M. Altshuler, Hiroyuki Aburatani, Keith W. Jones, Chris Tyler-Smith, Matthew E. Hurles, Nigel P. Carter, Stephen W. Scherer, , and Charles Lee. Copy number variation: New insights in genome diversity. *Genome Research*, 16:949–961, 2006.

Bert Fristedt and Lawrence Gray. *A Modern Approach to Probability Theory*. Birkhäuser Boston, 1996.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, second edition, 2003.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions Pattern Analalysis and Machine Intelligence*, 23(6), 2001.

Zoubin Ghahramani, Thomas L. Griffiths, and Peter Sollich. Bayesian nonparametric latent feature models. *Bayesian Statistics*, 8, 2006.

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. *The Journal of Machine Learning Research*, 8: 2265–2295, 2007.

Dilan Görür, Frank Jäkel, and Carl Edward Rasmussen. A choice model with infinitely many latent features. In *Proceedings of the International Conference on Machine learning (ICML)*, 2006.

Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. Technical Report 2005-001, Gatsby Computational Neuroscience Unit, 2005.

Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Nils Lid Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, 18(3):1259–1294, 1990.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Peter D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295, 2005.

Peter D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 2008.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098, 2002.

Matthew D. Hoffman, David M. Blei, and Perry R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the International Conference on Machine learning (ICML)*, 2010.

Jin Huang and Charles X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005. ISSN 1041-4347.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 1997.

Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the American Association for Artificial Intelligence (AAAI)*, 2006.

Yongdai Kim. Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics*, 27(2):562–588, 1999.

John F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

John F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

David Knowles and Zoubin Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Conference on Independent Component Analysis and Signal Separation*, 2007.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Dahua Lin, Eric Grimson, and John Fisher. Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

R. Duncan Luce. *Individual Choice Behavior*. Wiley, 1959.

Daniel McFadden. Economic choices. *American Economic Review*, 91:351–378, 2001.

Edward Meeds, Zoubin Ghahramani, Radford Neal, and Sam Roweis. Modeling dyadic data with binary latent factors. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2007.

Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2008a.

Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. Variations on non-exchangeable nonparametric priors for latent feature model. In *Uncertainty in Artificial Intelligence (UAI): Nonparametric Bayesian Workshop*, 2008b.

Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. Nonparametric latent feature models for link prediction. In Y. Bengio, D. Schuurmans, J. Lafferty, and C. Williams, editors, *Advances in Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2009.

Kurt T. Miller, Michael I. Jordan, and Thomas L. Griffiths. Non-exchangeable Bayesian nonparametric latent feature models. In *International Society of Bayesian Analaysis (ISBA) World Meeting*, 2010.

Daniel L. Navarro and Thomas L. Griffiths. A nonparametric Bayesian method for inferring features from similarity judgments. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Daniel L. Navarro and Thomas L. Griffiths. Latent features in similarity judgment: A nonparametric Bayesian approach. *Neural Computation*, 20(11):2597–2628, 2008.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Department of Statistics, University of Toronto, 1998.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the International Conference on Machine learning (ICML)*, 2009.

John Paisley, Aimee Zaas, Christopher W. Woods, Geoffrey S. Ginsburg, and Lawrence Carin. A stick-breaking construction of the beta process. In *Proceedings of the International Conference on Machine learning (ICML)*, 2010.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Statistics*, 25(2):855–900, 1997.

Piyush Rai and Hal Daumé. The infinite hierarchical factor regression model. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Vinayak Rao and Yee Whye Teh. Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

Garry Robins, Tom Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2):192–215, May 2007.

Donald L. Rumelhart and James G. Greeno. Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, 8(3):370–381, 1971.

Rudolph J. Rummel. Dimensionality of nations project: Attributes of nations and behavior of nation dyads, 1950–1965. ICPSR data file, 1999.

Ken-iti Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.

Mark J. Schervish. *Theory of Statistics*. Springer, 1995.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

Erik B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2006.

Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky. Describing visual scenes using transformed Dirichlet processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Yee Whye Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

Yee Whye Teh and Dilan Görür. Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Yee Whye Teh and Michael I. Jordan. Hierarchical Bayesian nonparametric models with applications. In Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Yee Whye Teh, Dilan Görür, and Zoubin Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

The Cancer Genome Atlas. http://cancergenome.nih.gov/, 2011.

Romain Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, University of California, Berkeley, Berkeley, CA, 2008.

Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

Michalis K. Titsias. The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Amos Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, 79: 281–299, 1972.

Amos Tversky and Shmuel Sattath. Preference trees. *Psychological Review*, 86(6): 542–573, 1979.

Jurgen Van Gael, Yee Whye Teh, and Zoubin Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2): 1–305, 2008.

Yuchung J. Wang and George Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

Stanley Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks: I. an introduction to Markov random graphs and p*. *Psychometrika*, 61(3):401–425, 1996.

Florian Wickelmaier and Christian Schmid. A Matlab function to estimate choice model parameters from pair comparison data. *Behavior Research Methods, Instruments, and Computers*, 36:29–40, 2004.

Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent Indian buffet processes. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010a.

Sinead Williamson, Chong Wang, Katherine A. Heller, and David M. Blei. The IBP-compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the International Conference on Machine learning (ICML)*, 2010b.

Robert L. Wolpert and Katja Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267, 1998a.

Robert L. Wolpert and Katja Ickstadt. Simulation of Lévy random fields. *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics*, 133:227–242, 1998b.

Frank Wood and Thomas L. Griffiths. Particle filtering for nonparametric Bayesian matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Frank Wood, Thomas L. Griffiths, and Zoubin Ghahramani. A non-parametric Bayesian method for inferring hidden causes. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2006.

Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Infinite hidden relational models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2006.