Clusters and Features from Combinatorial Stochastic Processes

by

Tamara Ann Broderick

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Thomas L. Griffiths
Professor Cari Kaufman
Professor James W. Pitman

Fall 2014

Clusters and Features from Combinatorial Stochastic Processes

Abstract

Clusters and Features from Combinatorial Stochastic Processes

by

Tamara Ann Broderick

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

Clustering involves placing entities into mutually exclusive categories. We wish to relax the requirement of mutual exclusivity, allowing objects to belong simultaneously to multiple classes, a formulation that we refer to as “feature allocation.” The first step is a theoretical one. In the case of clustering the class of probability distributions over exchangeable partitions of a dataset has been characterized (via exchangeable partition probability functions and the Kingman paintbox). These characterizations support an elegant nonparametric Bayesian framework for clustering in which the number of clusters is not assumed to be known a priori. We establish an analogous characterization for feature allocation; we define notions of “exchangeable feature probability functions” and “feature paintboxes” that lead to a Bayesian framework that does not require the number of features to be fixed a priori. We focus on particular models within this framework that are both practical for inference and provide desirable modeling properties. And we explore a further generalization to feature allocations where objects may exhibit any non-negative integer number of features, or traits.

The second step is a computational one. Rather than appealing to Markov chain Monte Carlo for Bayesian inference, we develop a method to transform Bayesian methods for feature allocation (and other latent structure problems) into optimization problems with objective functions analogous to K-means in the clustering setting. These yield approximations to Bayesian inference that are scalable to large inference problems.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgments

First and foremost, I thank my advisor, Michael I. Jordan—for his enthusiasm, encouragement, inspiration, support, and mentorship. I am also particularly grateful to Jim Pitman, whose open door has sparked a number of fascinating discussions and collaborations. I thank my dissertation committee, further including Tom Griffiths and Cari Kaufman, for a number of useful discussions over the years. Indeed, my thanks go out to all of the other colleagues with whom I have had the great pleasure of collaborating on various parts of this work: Nick Boyd, Brian Kulis, Lester Mackey, John Paisley, Andre Wibisono, Ashia Wilson. I am fortunate to have worked with a number of other great researchers on projects during this time that, although not appearing here, have been formative in my doctoral experience: Josh Bloom, David Duvenaud, Jan Luts, Ryan Giordano, Joey Gonzalez, Stefanie Jegelka, Mason Liang, James Long, Ron Minnich, Adam Morgan, Xinghao Pan, Barret Rhoden, Joey Richards, Purnamrita Sarkar, Rebecca Steorts, Matt Wand.

I wish to thank everyone in my research communities, especially the Bayesian nonparametrics community, who has been so supportive. I am particularly indebted to David Banks, David Blei, Zoubin Ghahramani, Peter Müller, Peter Orbanz for their invaluable mentorship. Thanks also to: the SAIL group at UC Berkeley, Ryan Adams, David Aldous, Kevin Canini, Sam Clifford, Finale Doshi, Barbara Engelhardt, Steve Evans, Noah Forman, Nick Foti, Emily Fox, Alexander Gnedin, Lauren Hannah, Katherine Heller, Philipp Hennig, Nils Hjort, Lancelot James, Wayne Tai Lee, Mason Liang, Steve MacEachern, Chris Melgaard, Douglas Rizzolo, Dan Roy, Josh Schraiber, Yun Song, Erik Sudderth, Yee Whye Teh, Hanna Wallach, Sinead Williamson, Bin Yu.

This research has benefited from a number of other sources of support. I thank Matt Hoffman, Tom Griffiths, and Chong Wang for generously sharing their code. I also thank Ryan Lovett—our SCF Systems Manager—for his patient troubleshooting and La Shana Porlaris—our Graduate Student Services Advisor—for going above and beyond in helping me navigate bureaucracy. I am grateful to have been funded during my doctorate by the National Science Foundation Graduate Student Fellowship and the Berkeley Fellowship. Further travel support during this time has come from NSF VIGRE, the NIPS Foundation, WiML, and SFASA. Finally, I thank my family and friends; I couldn't have done it without you.

Chapter 1

Introduction

The continued growth of Bayesian statistics may be attributed to the flexibility of hierarchical modeling and the coherent treatment of uncertainty this paradigm facilitates. Despite the successes of Bayesian statistics, its classical use is situated firmly in a small data world; Markov chain Monte Carlo sampling is often too slow for posterior calculation on modern, massive data sets, and finite-dimensional prior distributions are not able to accommodate the inevitable growth in complexity that comes with significant growth in data size.

Bayesian nonparametrics is the area of Bayesian analysis in which the finite-dimensional prior distributions of classical Bayesian analysis are replaced with stochastic processes. One may view Bayesian nonparametrics as supplying modelers with a richer collection of distributions with which to express prior belief. In practice, however, the field has been dominated by two stochastic processes—the Gaussian process and the Dirichlet process—and thus the flexibility promised by the nonparametric approach has arguably not yet been delivered.

This manuscript provides a broader perspective on the kinds of stochastic processes that populate a toolbox for Bayesian nonparametric analysis. In Part I, we see how combinatorial stochastic processes embody mathematical structure that is useful for both model specification and inference. For instance, a significant body of literature develops prior distributions for clustering, where each data point can belong to one and only one group, called a cluster. Chapter 2 introduces an extension of clustering called a *feature allocation*, where each data point can belong to any non-negative integer number of groups, now called *features*. There are a number of practical examples: each member of a social network might belong to multiple friend groups; each document in a corpus might best be described by multiple themes, or topics; and a customer's purchases might correspond to multiple interests.

With flexible, nonparametric models for these problems in place (Part I), we can focus on computationally scalable inference. Part II demonstrates how to retain the strengths of the Bayesian paradigm and nonparametric analysis while simultaneously enabling fast, and even streaming, inference on large data sets.

We begin in Chapter 2 by noting that a number of disparate representations of clustering models have been used in proofs and inference. A key concept in our analysis is *exchangeability*, which expresses the assumption that any particular order in which data points are seen is irrelevant for inference. Kingman (1978) has shown that the distribution of any random, exchangeable clustering is equivalent to the distribution of a construction of the following type: first draw a random partition (potentially countably infinite) of the unit interval, then draw cluster belonging for each data point with distribution given by this partition, called the *Kingman paintbox*. Pitman (1995) has further shown that—given a random, exchangeable clustering—the probability of any configuration depends only on the cluster sizes, via a function called the *exchangeable partition probability function* (EPPF). Both representations are useful in defining and evaluating new clustering models.

We review how these representations can be viewed as sequential augmentations from simple distributions over partitions (EPPFs) to cluster frequencies (the sizes of the partition intervals in Kingman's paintbox) to *subordinators* and *completely random measures*, which associate a general class of labels with the stick lengths and whose labels we typically use as cluster-specific parameters in likelihood models (Broderick, Jordan, and Pitman, 2013). We show how an analogous augmentation regime can be built up for feature frequency models: from simple distributions over feature allocations to feature frequencies to a (different) class of subordinators and completely random measures. We discuss implications of each representation for tractable inference and provide running examples of the *Dirichlet process* (for clustering) and the *beta process* (for feature modeling).

In Chapter 3, we focus more deeply on the beta process model. Choosing a Bayesian prior often amounts to choosing the most tractable model for inference that satisfies known properties of the model. The beta process provides a prior on an unbounded and unknown number of feature frequencies while also allowing tractable inference. However, *power laws* are often observed in real-world data sets; e.g., we expect, from empirical evidence, that the number of features in a data set will grow as a fixed (but typically unknown) power of the size of the data set. We define a *three-parameter beta process* (3BP) as a generalization of the beta process (Broderick, Jordan, and Pitman, 2012). We show that the 3BP exhibits many of the same traits that allow straightforward inference in the beta process. We further prove, and demonstrate via simulation, that the 3BP almost surely exhibits desired power laws. We show the usefulness of this construction by developing a Markov chain Monte Carlo (MCMC) inference scheme and learning a factor model in a computer vision experiment.

Chapter 4 takes the generalization from clustering models to feature models one step further. We define an extension to the feature modeling framework where each data point may belong any non-negative integer amount to any non-negative integer number of features (not just 0 or 1 as in vanilla feature modeling) (Broderick, Mackey, et al., 2014). For instance, we might want to assign multiple words in a document to a topic (e.g., economics, the arts, sports) or multiple patches in an image to an object class (e.g., grass, sky, car). We propose a beta process model with negative binomial likelihood. We prove the conjugacy of these stochastic processes and provide a power-law extension. We develop MCMC inference and demonstrate the model's applicability in segmenting images and analyzing documents through topic identification.

In the remaining chapters of Part I, we examine the space of potential Bayesian nonparametric models in more depth. In Chapter 5, we show that all exchangeable feature allocations have distribution equivalent to a feature paintbox construction. Moreover, we define both (1) an *exchangeable feature probability function* (EFPF) and (2) *feature frequency models* (Broderick, Jordan, and Pitman, 2013; Broderick, Pitman, and Jordan, 2013). The EFPF is similar to the EPPF though now with an explicit dependence on the data set size. A feature frequency model is characterized as having distribution equivalent to the following construction: draw each feature for each data point as an independent, Bernoulli random variable conditioned on some underlying, random, feature-specific frequency. We show that the distributions with EFPFs are exactly the feature frequency models. While one might initially think of feature models as analogous to cluster models, our results situate feature frequency models and clusterings as analogous subclasses of feature models. With these results, we bring the same completeness to feature allocation characterizations as clustering characterizations.

In Chapter 6, we focus on completely random measures (CRMs) as a particular way to generate feature frequencies, not just for feature allocations but more generally for the case where each data point may exhibit features with a non-negative integer multiplicity (cf. Chapter 4) (Broderick, Wilson, and Jordan, 2014). We demonstrate how to calculate posteriors for general CRM-based priors and likelihoods for Bayesian nonparametric models. Motivated by conjugate priors based on exponential family representations of likelihoods, we introduce a notion of exponential families for CRMs, which we call *exponential CRMs*. This construction allows us to specify automatic Bayesian nonparametric conjugate priors for exponential CRM likelihoods. We demonstrate that our exponential CRMs allow particularly straightforward recipes for size-biased and marginal representations of Bayesian nonparametric models.

The focus of Part I is elucidation of a wide range of models, reflecting known desiderata for various generalizations of clustering. Part II aims more specifically at performing scalable inference in the modern Big Data context while maintaining strengths of the Bayesian paradigm, such as flexible hierarchical modeling.

One particular challenge arising from large datasets is streaming data, where we assume computer memory can hold only fixed-size data subsets and that data, once processed, cannot be revisited. We address this challenge while taking advantage of modern distributed computing architectures in Chapter 7 by developing *SDA-Bayes*, a framework for (S)treaming, (D)istributed, (A)synchronous computation of a Bayesian posterior (Broderick, Boyd, et al., 2013). The framework takes advantage of the naturally streaming nature of iterative Bayesian posterior calculation to make streaming updates to the estimated posterior according to a user-specified approximation batch primitive. We demonstrate the usefulness of our framework on an unsupervised topic learning task with two corpuses: Wikipedia (over 3M documents) and the journal Nature (over 300K documents). We use latent Dirichlet allocation (LDA) as a model for assigning documents to topics. We use variational Bayes (VB), a popular and fast posterior approximation method, as the primitive. We demonstrate that our algorithm, though taking only streaming data, performs as well as a popular non-streaming algorithm for learning LDA with VB.

There are certain trade-offs involved in using VB—and further our streaming, distributed approximation—to approximate a Bayesian posterior. In our MAD-Bayes (MAP-based Asymptotic Derivation from Bayes) framework (Chapter 8), though, we consider a more radical trade-off (Broderick, Kulis, and Jordan, 2013). Recognizing the scalability and ease-of-use of K-means, we take limits of Bayesian posteriors to invent novel K-means-like objective functions and algorithms. In particular, the classical mixture of Gaussians model is related to K-means via *small variance asymptotics*: as the covariances of the Gaussians tend to zero, the negative log-likelihood of the mixture of Gaussians model approaches the K-means objective, and the EM algorithm approaches the K-means algorithm. We instead consider applying small-variance asymptotics directly to the posterior in Bayesian nonparametric models. This framework is independent of any specific Bayesian inference algorithm and generalizes to a range of models. To illustrate, we apply our framework to feature learning, where the beta process provides an appropriate Bayesian nonparametric prior. We obtain novel objective functions and algorithms, all of which are scalable and simple to implement. Empirical results in computer vision demonstrate the benefits of the new framework.

Part I

Models, connections, and inference

Chapter 2

Cluster and feature modeling from combinatorial stochastic processes

One of the focal points of the modern literature on Bayesian nonparametrics has been the problem of *clustering*, or *partitioning*, where each data point is modeled as being associated with one and only one of some collection of groups called clusters or partition blocks. Underlying these Bayesian nonparametric models are a set of interrelated stochastic processes, most notably the Dirichlet process and the Chinese restaurant process. In this chapter we provide a formal development of an analogous problem, called *feature modeling*, for associating data points with arbitrary non-negative integer numbers of groups, now called features or topics. We review the existing combinatorial stochastic process representations for the clustering problem and develop analogous representations for the feature modeling problem. These representations include the beta process and the Indian buffet process as well as new representations that provide insight into the connections between these processes. We thereby bring the same level of completeness to the treatment of Bayesian nonparametric feature modeling that has previously been achieved for Bayesian nonparametric clustering.

2.1 Introduction

Bayesian nonparametrics is the area of Bayesian analysis in which the finite-dimensional prior distributions of classical Bayesian analysis are replaced with stochastic processes. While the rationale for allowing infinite collections of random variables into Bayesian inference is often taken to be that of diminishing the role of prior assumptions, it is also possible to view the move to nonparametrics as supplying the Bayesian paradigm with a richer collection of distributions with which to express prior belief, thus in some sense emphasizing the role of the prior. In practice, however, the field has been dominated by two stochastic processes—the Gaussian process and the Dirichlet process—and thus the flexibility promised by the nonparametric approach has arguably not yet been delivered. In the current chapter we aim to provide a broader perspective on the kinds of stochastic processes that can provide a useful toolbox for Bayesian nonparametric analysis. Specifically, we focus on *combinatorial stochastic processes* as embodying mathematical structure that is useful for both model specification and inference.

The phrase "combinatorial stochastic process" comes from probability theory (Pitman, 2006), where it refers to connections between stochastic processes and the mathematical field of combinatorics. Indeed, the focus in this area of probability theory is on random versions of classical combinatorial objects such as partitions, trees, and graphs—and on the role of combinatorial analysis in establishing properties of these processes. As we wish to argue, this connection is also fruitful in a statistical setting. Roughly speaking, in statistics it is often natural to model observed data as arising from a combination of underlying factors. In the Bayesian setting, such models are often embodied as latent variable models in which the latent variable has a compositional structure. Making explicit use of ideas from combinatorics in latent variable modeling can not only suggest new modeling ideas but can also provide essential help with calculations of marginal and conditional probability distributions.

The Dirichlet process already serves as one interesting exhibit of the connections between Bayesian nonparametrics and combinatorial stochastic processes. On the one hand, the Dirichlet process is classically defined in terms of a partition of a probability space (Ferguson, 1973), and there are many well-known connections between the Dirichlet process and urn models (Blackwell and MacQueen, 1973; Hoppe, 1984). In the current chapter, we will review and expand upon some of these connections, beginning our treatment (non-traditionally) with the notion of an *exchangeable partition probability function* (EPPF) and, from there, discussing related urn models, stick-breaking representations, subordinators, and random measures.

On the other hand, the Dirichlet process is limited in terms of the statistical notion of "combination of underlying factors" that we referred to above. Indeed, the Dirichlet process is generally used in a statistical setting to express the idea that each data point is associated with one and only one underlying factor. In contrast to such *clustering models*, we wish to also consider *featural models*, where each data point is associated with a set of underlying features and it is the interaction among these features that gives rise to an observed data point. Focusing on the case in which these features are binary, we develop some of the combinatorial stochastic process machinery needed to specify featural priors. Specifically, we develop a counterpart to the EPPF, which we refer to as the *exchangeable feature probability function* (EFPF), that characterizes the combinatorial structure of certain featural models. We again develop connections between this combinatorial function and suite of related stochastic processes, including urn models, stick-breaking representations, subordinators, and random measures. As we will discuss, a particular underlying random measure in this case is the *beta process*, originally studied by Hjort (1990) as a model of random hazard functions in survival analysis, but adapted by Thibaux and Jordan (2007) for applications in featural modeling.

For statistical applications it is not enough to develop expressive prior specifications, but it is also essential that inferential computations involving the posterior distribution are tractable. One of the reasons for the popularity of the Dirichlet process is that the associated urn models and stick-breaking representations yield a variety of useful inference algorithms (Neal, 2000). As we will see, analogous algorithms are available for featural models. Thus, as we discuss each of the various representations associated with both the Dirichlet process and the beta process, we will also (briefly) discuss some of the consequences of each for posterior inference.

The remainder of the chapter is organized as follows. We start by reviewing partitions and introducing feature allocations in Section 2.2 in order to define distributions over these models (Section 2.3) via the EPPF in the partition case (Section 2.3) and the EFPF in the feature allocation case (Section 2.3). Illustrating these exchangeable probability functions with examples, we will see that the well-known *Chinese restaurant process* (CRP) (Aldous, 1985) corresponds to a particular EPPF choice (Example 2.3.1) and the *Indian buffet process* (IBP) (Griffiths and Ghahramani, 2006) corresponds to a particular choice of EFPF (Example 2.3.5). From here, we progressively build up richer models by first reviewing stick lengths (Section 2.4), which we will see represent limiting frequencies of certain clusters or features, and then subordinators (Section 2.5), which further associate a random label with each cluster or feature. We illustrate these progressive augmentations for both the CRP (Examples 2.3.1, 2.3.6, 2.4.3, 2.5.7, and 2.5.9) and IBP examples (Examples 2.3.5, 2.3.7, 2.4.4, and 2.5.4). We augment the model once more to obtain a random measure on a general space of cluster or feature parameters in Section 2.6, and discuss how marginalization of this random measure yields the CRP in the case of the Dirichlet process (Example 2.6.1), and the IBP in the case of the beta process (Example 2.6.2). Finally, in Section 2.7, we mention some of the other combinatorial stochastic processes, beyond the Dirichlet process and the beta process, that have begun to be studied in the Bayesian nonparametrics literature, and we provide suggestions for further developments.

2.2 Partitions and feature allocations

While we have some intuitive ideas about what constitutes a cluster or feature model, we want to formalize these ideas before proceeding. We begin with the underlying combinatorial structure on the data indices. We think of $[N] := \{1, \ldots, N\}$ as representing the indices of the first N data points. There are different groupings that we apply in the cluster case (*partitions*) and feature case (*feature allocations*); we describe these below.

First, we wish to describe the space of *partitions* over the indices $[N]$. In particular, a partition π_N of $[N]$ is defined to be a collection of mutually exclusive, exhaustive, non-empty subsets of $[N]$ called *blocks*; that is, $\pi_N = \{A_1, \ldots, A_K\}$ for some number of partition blocks K. An example partition of $[6]$ is $\pi_6 = \{\{1,3,4\},\{2\},\{5,6\}\}$. Similarly, a partition of $\mathbb{N} = \{1, 2, \ldots\}$ is a collection of mutually exclusive, exhaustive, non-empty subsets of $\mathbb{N}$. In this case, the number of blocks may be infinite, and we have $\pi_{\mathbb{N}} = \{A_1, A_2, \ldots\}$. An example partition of $\mathbb{N}$ into two blocks is $\{\{n : n \text{ is even}\}, \{n : n \text{ is odd}\}\}$.

We introduce a generalization of a partition called a *feature allocation* that relaxes both the mutually exclusive and exhaustive restrictions. In particular, a feature allocation f_N of

$[N]$ is defined to be a multiset of non-empty subsets of $[N]$, again called *blocks*, such that each index n can belong to any finite number of blocks. Note that the constraint that no index belong to infinitely many blocks coincides with our intuition for the meaning of these blocks as groups to which the index belongs. Consider an example where the data points are images expressed as pixel arrays, and the latent features represent animals that may or may not appear in each picture. It is impossible to display an infinite number of animals in a picture with finitely many pixels.

We write $f_N = \{A_1, \ldots, A_K\}$ for some number of feature allocation blocks K. An example feature allocation of $[6]$ is $f_6 = \{\{2,3\},\{2,4,6\},\{3\},\{3\},\{3\}\}$. Just as the blocks of a partition are sometimes called *clusters*, so are the blocks of a feature allocation sometimes called *features*. We note that a partition is always a feature allocation, but the converse statement does not hold in general; for instance, f_6 given above is not a partition.

In the remainder of this section, we continue our development in terms of feature allocations since partitions are a special case of the former object. We note that we can extend the idea of random partitions (Aldous, 1985) to consider *random feature allocations.* If $\mathcal{F}_N$ is the space of all feature allocations of $[N]$, then a random feature allocation F_N of $[N]$ is a random element of this space.

We next introduce a few useful assumptions on our random feature allocation. Just as exchangeability of observations is often a central assumption in statistical modeling, so will we make use of *exchangeable feature allocations.* To rigorously define such feature allocations, we introduce the following notation. Let $\sigma : \mathbb{N} \to \mathbb{N}$ be a finite permutation. That is, for some finite value N_σ, we have $\sigma(n) = n$ for all $n > N_\sigma$. Further, for any block $A \subset \mathbb{N}$, denote the permutation applied to the block as follows: $\sigma(A) := \{\sigma(n) : n \in A\}$. For any feature allocation F_N, denote the permutation applied to the feature allocation as follows: $\sigma(F_N) := \{\sigma(A) : A \in F_N\}$. Finally, let F_N be a random feature allocation of $[N]$. Then we say that F_N is exchangeable if $F_N \overset{d}{=} \sigma(F_N)$ for every finite permutation σ.

Our second assumption in what follows will be that we are dealing with a *consistent* feature allocation. We often implicitly imagine the indices arriving one at a time: first 1, then 2, up to N or beyond. We will find it useful, similarly, in defining random feature allocations to suppose that the randomness at stage n somehow agrees with the randomness at stage $n+1$. More formally, we say that a feature allocation f_M of $[M]$ is a *restriction* of a feature allocation f_N of $[N]$ for $M < N$ if

$$f_M = \{A \cap [M] : A \in f_N\}.$$

Let $\mathcal{R}_N(f_M)$ be the set of all feature allocations of $[N]$ whose restriction to $[M]$ is f_M. Then we say that the sequence of random feature allocations (F_n) is *consistent* if for all M and N such that $M < N$, we have that

$$F_N \in \mathcal{R}_N(F_M) \quad a.s. \tag{2.1}$$

With this consistency condition in hand, we can define a random feature allocation F_∞ of $\mathbb{N}$. In particular, such a feature allocation is characterized by the sequence of consistent finite restrictions F_N to $[N]$: $F_N := \{A \cap [N] : A \in F_\infty\}$. Then F_∞ is equivalent to a consistent sequence of finite feature allocations and may be thought of as a random element of the space of such sequences: $F_\infty = (F_n)_n$. We let $\mathcal{F}_\infty$ denote the space of consistent feature allocations, of which each random feature allocation is a random element, and we see that the sigma-algebra associated with this space is generated by the finite-dimensional sigma-algebras of the restricted random feature allocations F_n.

We say that F_∞ is exchangeable if $F_\infty \overset{d}{=} \sigma(F_\infty)$ for every finite permutation σ. That is, when the permutation σ changes no indices above N, we require $F_N \overset{d}{=} \sigma(F_N)$, where F_N is the restriction of F_∞ to $[N]$. A characterization of distributions for F_∞ is provided by Broderick, Pitman, and Jordan (2013), where a similar treatment of the introductory ideas of this section also appears.

In what follows, we consider particular useful ways of representing distributions for exchangeable, consistent random feature allocations with emphasis on partitions as a special case.

2.3 Exchangeable probability functions

Once we know that we can construct (exchangeable and consistent) random partitions and feature allocations, it remains to find useful representations of distributions over these objects.

Exchangeable partition probability function

Consider first an exchangeable, consistent, random partition (Π_n). By the exchangeability assumption, the distribution of the partition should depend only on the (unordered) sizes of the blocks. Therefore, there exists a function p that is symmetric in its arguments such that, for any specific partition assignment $\pi_n = \{A_1, \ldots, A_K\}$, we have

$$\mathbb{P}(\Pi_n = \pi_n) = p(|A_1|, \ldots, |A_K|). \tag{2.2}$$

The function p is called the *exchangeable partition probability function* (EPPF) (Pitman, 1995).

Example 2.3.1 (Chinese restaurant process)**.** The Chinese restaurant process (CRP) (Blackwell and MacQueen, 1973) is an iterative description of a partition via the conditional distributions of the partition blocks to which increasing data indices belong. The Chinese restaurant metaphor forms an equivalence between customers entering a Chinese restaurant and data indices; customers who share a table at the restaurant represent indices belonging to the same partition block.

To generate the label for the first index, the first customer enters the restaurant and sits down at some table, necessarily unoccupied since no one else is in the restaurant. A "dish" is set out at the new table; call the dish "1" since it is the first dish. The customer is assigned

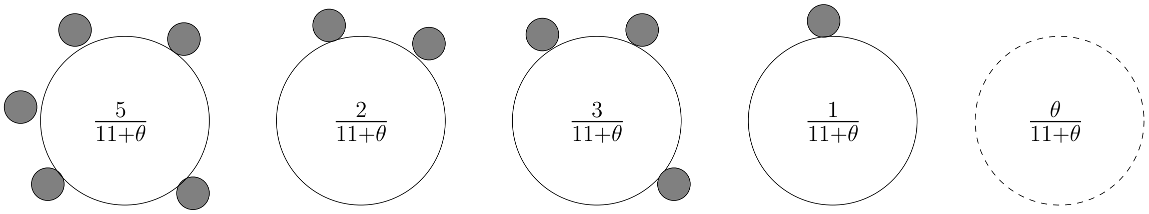


Figure 2.1: The diagram represents a possible CRP seating arrangement after 11 customers have entered a restaurant with parameter θ. Each large white circle is a table, and the smaller gray circles are customers sitting at those tables. If a 12th customer enters, the expressions in the middle of each table give the probability of the new customer sitting there. In particular, the probability of the 12th customer sitting at the first table is $5/(11+\theta)$, and the probability of the 12th customer forming a new table is $\theta/(11+\theta)$.

the label of the dish at her table: $Z_1 = 1$. Recursively, for a restaurant with *concentration parameter* θ, the nth customer sits at an occupied table with probability in proportion to the number of people at the table and at a new table with probability proportional to θ. In the former case, Z_n takes the value of the existing dish at the table, and in the latter case, the next available dish k (equal to the number of existing tables plus one) appears at the new table, and $Z_n = k$. By summing over all possibilities when the nth customer arrives, one obtains the normalizing constant for the distribution across potential occupied tables: $(n-1+\theta)^{-1}$. An example of the distribution over tables for the nth customer is shown in Figure 2.1. To summarize, if we let $K_n := \max\{Z_1, \ldots, Z_n\}$, then the distribution of table assignments for the nth customer is

$$\begin{aligned} &\mathbb{P}(Z_n = k | Z_1, \ldots, Z_{n-1}) \\ &= (n-1+\theta)^{-1} \begin{cases} \#\{m : m < n, Z_m = j\} & \text{for } j \leq K_{n-1} \\ \theta & \text{for } k = K_{n-1}+1 \end{cases} \end{aligned} \tag{2.3}$$

We note that an equivalent generative description follows a Pólya urn style in specifying that each incoming customer sits next to an existing customer with probability proportional to 1 and forms a new table with probability proportional to θ (Hoppe, 1984).

Next, we find the probability of the partition induced by considering the collection of indices sitting at each table as a block in the partition. Suppose that N_k individuals sit at table k so that the set of cardinalities of non-zero table occupancies is $\{N_1, \ldots, N_K\}$ with $N := \sum_{k=1}^{K} N_k$. That is, we are considering the case when N customers have entered the restaurant and sat at K different tables in the specified configuration.

We can see from Eq. (2.3) that when the nth customer enters ($n > 1$), we obtain a factor of $n-1+\theta$ in the denominator. Using the following notation for the rising and falling factorial

$$x_{M\uparrow a} := \prod_{m=0}^{M-1} (x + ma), \quad x_{M\downarrow a} := \prod_{m=0}^{M-1} (x - ma),$$

we find a factor of $(\theta + 1)_{N-1\uparrow 1}$ must occur in the denominator of the probability of the partition of $[N]$. Similarly, each time a customer forms a new table except for the first table, we obtain a factor of θ in the numerator. Combining these factors, we find a factor of θ^{K-1} in the numerator. Finally, each time a customer sits at an existing table with n occupants, we obtain a factor of n in the numerator. Thus, for each table k, we have a factor of $(N_k - 1)!$ once all customers have entered the restaurant.

Having collected all terms in the process, we see that the probability of the resulting configuration is:

$$\mathbb{P}(\Pi_N = \pi_N) = \frac{\theta^{K-1} \prod_{k=1}^{K} (N_k - 1)!}{(\theta + 1)_{N-1\uparrow 1}}. \tag{2.4}$$

We first note that Eq. (2.4) depends only on the block sizes and not on the order of arrival of the customers or dishes at the tables. We conclude that the partition generated according to the CRP scheme is exchangeable. Moreover, as the partition Π_M is the restriction of Π_N to $[M]$ for any $N > M$ by construction, we have that Eq. (2.4) satisfies the consistency condition. It follows that Eq. (2.4) is, in fact, an EPPF. ■

Exchangeable feature probability function

Just as we considered an exchangeable, consistent, random partition above, so we now turn to an exchangeable, consistent, random feature allocation (F_n) now. Let $f_N = \{A_1, \ldots, A_K\}$ be any particular feature allocation. In calculating $\mathbb{P}(F_N = f_N)$, we start by demonstrating in the next example that this probability in some sense undercounts features when they contain exactly the same indices: e.g., $A_j = A_k$ for some $j \neq k$. For instance, consider the following example.

Example 2.3.2 (A two-block, Bernoulli feature allocation)**.** Let $q_A, q_B \in (0, 1)$ represent the frequencies of features A and B. Draw $Z_{A,n} \overset{iid}{\sim} \text{Bern}(q_A)$ and $Z_{B,n} \overset{iid}{\sim} \text{Bern}(q_B)$, independently. Construct the random feature allocation by collecting those indices with successful draws:

$$F_N := \{\{n : n \leq N, Z_{A,n} = 1\}, \{n : n \leq N, Z_{B,n} = 1\}\}.$$

Then the probability of the feature allocation $F_5 = f_5 := \{\{2, 3\}, \{2, 3\}\}$ is

$$q_A^2 (1 - q_A)^3 q_B^2 (1 - q_B)^3,$$

but the probability of the feature allocation $F_5 = f_5' := \{\{2, 3\}, \{2, 5\}\}$ is

$$2 q_A^2 (1 - q_A)^3 q_B^2 (1 - q_B)^3.$$

The difference is that in the latter case the features can be distinguished, and so we must account for the two possible pairings of features to frequencies $\{q_A, q_B\}$.

Now, instead, let $\tilde{F}_N$ be F_N with a uniform random ordering on the features. There is just a single possible ordering of f_5, so the probability of $\tilde{F}_5 = \tilde{f}_5 := (\{2, 3\}, \{2, 3\})$ is again

$$q_A^2 (1 - q_A)^3 q_B^2 (1 - q_B)^3.$$

However, there are two orderings of f'_5, so the probability of $\tilde{F}_5 = \tilde{f}'_5 := (\{2,5\}, \{2,3\})$ is

$$q_A^2 (1 - q_A)^3 q_B^2 (1 - q_B)^3,$$

and the same holds for the other ordering. ■

For reasons suggested by the previous example, we will find it useful to work with the random feature allocation after uniform random ordering, $\tilde{F}_N$. One way to achieve such an ordering and maintain consistency across different N is to associate some independent, continuous random variable with each feature; e.g. assign a uniform random variable on $[0, 1]$ to each feature and order the features according to the order of the assigned random variables. When we view feature allocations constructed as marginals of a *subordinator* in Section 2.5, we will see that this construction is natural.

In general, given a probability of a random feature allocation, $\mathbb{P}(F_N = f_N)$, we can find the probability of a *random ordered feature allocation*, $\mathbb{P}(\tilde{F}_N = \tilde{f}_N)$ as follows. Let H be the number of unique elements of F_N, and let $(\tilde{K}_1, \ldots, \tilde{K}_H)$ be the multiplicities of these unique elements in decreasing size. Then

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = \binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H}^{-1} \mathbb{P}(F_N = f_N), \tag{2.5}$$

where

$$\binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H} := \frac{K!}{\tilde{K}_1! \cdots \tilde{K}_H!}.$$

We will see in Section 2.5 that augmentation of an exchangeable partition with a random ordering is also natural. However, the probability of an ordered random partition is not substantively different from the probability of an unordered version since the factor contributed by ordering a partition is always $1/K!$, where K here is the number of partition blocks.

With this framework in place, we can see that some ordered feature allocations have a probability function p nearly as in Eq. (2.2) that is, moreover, symmetric in its block-size arguments. Consider again the previous example.

Example 2.3.3 (A two-block, Bernoulli feature allocation (continued))**.** Consider any F_N with block sizes N_1 and N_2 constructed as in Example 2.3.2. Then

$$\begin{aligned}\mathbb{P}(\tilde{F}_N = \tilde{f}_N) &= \frac{1}{2} q_A^{N_1} (1 - q_A)^{N - N_1} q_B^{N_2} (1 - q_B)^{N - N_2} \\ &\quad + \frac{1}{2} q_A^{N_2} (1 - q_A)^{N - N_2} q_B^{N_1} (1 - q_B)^{N - N_1} \\ &= p(N, N_1, N_2), \end{aligned} \tag{2.6}$$

where p is some function of the number of indices N and the block sizes (N_1, N_2) that we note is symmetric in all arguments after the first. In particular, we see that the order of N_1 and N_2 was immaterial. ■

We note that in the partition case, $\sum_{k=1}^{K} |A_k| = N$, so N is implicitly an argument to the EPPF. In the feature case, this summation condition no longer holds, so we make the argument N explicit in Eq. (2.6).

However, it is not necessarily the case that such a function, much less a symmetric one, exists for exchangeable feature models—in contrast to the case of exchangeable partitions and the EPPF.

Example 2.3.4 (A general two-block feature allocation)**.** We here describe an exchangeable, consistent random feature allocation whose (ordered) distribution does not depend only on the number of indices N and the sizes of the blocks of the feature allocation.

Let p_1, p_2, p_3, p_4 be fixed frequencies that sum to one. Let Y_n represent the collection of features to which index n belongs. For $n \in \{1, 2\}$, choose Y_n independently and identically according to:

$$Y_n = \begin{cases} \{A\} & \text{with probability } p_1 \\ \{B\} & \text{with probability } p_2 \\ \{A, B\} & \text{with probability } p_3 \\ \emptyset & \text{with probability } p_4 \end{cases} .$$

We form a feature allocation from these labels as follows. For each label (A or B), collect those indices n with the given label appearing in Y_n to form a feature.

Now consider two possible outcome feature allocations: $f_2 = \{\{2\}, \{2\}\}$, and $f_2' = \{\{1\}, \{2\}\}$. The likelihood of any random ordering $\tilde{f}_2$ of f_2 under this model is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) = p_1^0 \, p_2^0 \, p_3^1 \, p_4^1.$$

The likelihood of any ordering $\tilde{f}_2'$ of f' is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}_2') = p_1^1 \, p_2^1 \, p_3^0 \, p_4^0.$$

It follows from these two likelihoods that we can choose values of p_1, p_2, p_3, p_4 such that $\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) \neq \mathbb{P}(\tilde{F}_2 = \tilde{f}_2')$. But $\tilde{f}_2$ and $\tilde{f}_2'$ have the same block counts and N value ($N = 2$). So there can be no such symmetric function p, as in Eq. (2.6), for this model. ■

When a function p exists in the form

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = p(N, |A_1|, \ldots, |A_K|) \tag{2.7}$$

for some random ordered feature allocation $\tilde{f}_N = (A_1, \ldots, A_K)$ such that p is symmetric in all arguments after the first, we call it the *exchangeable feature probability function* (EFPF). Note that the EPPF is not a special case of the EFPF. The EPPF assigns zero probability to any multiset in which an index occurs in more than one element of the multiset; only the sizes of the multiset blocks are relevant in the EFPF case.

We next consider a more complex example of an EFPF.

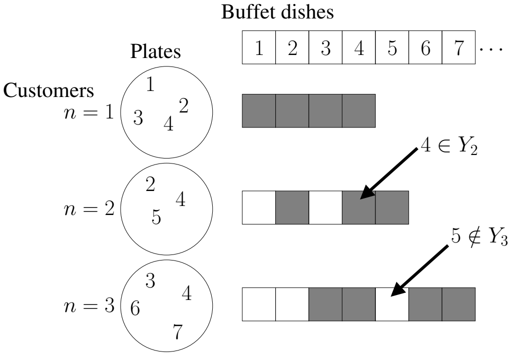


Figure 2.2: Illustration of an Indian buffet process. The buffet (*top*) consists of a vector of dishes, corresponding to features. Each customer—corresponding to a data point—who enters first decides whether or not to eat dishes that the other customers have already sampled and then tries a random number of new dishes, not previously sampled by any customer. A gray box in position (n, k) indicates customer n has sampled dish k, and a white box indicates the customer has not sampled the dish. In the example, the second customer has sampled exactly those dishes indexed by 2, 4, and 5: $Y_2 = \{2, 4, 5\}$.

Example 2.3.5 (Indian buffet process)**.** The Indian buffet process (IBP) (Griffiths and Ghahramani, 2006) is a generative model for a random feature allocation that is specified recursively like the Chinese restaurant process. Also like the CRP, this culinary metaphor forms an equivalence between customers and the indices n that will be partitioned: $n \in \mathbb{N}$. Here, "dishes" again correspond to feature labels just as they corresponded to partition labels for the CRP. But in the IBP case, a customer can sample multiple dishes.

In particular, we start with a single customer, who enters the buffet and chooses $K_1^+ \sim \text{Poisson}(\gamma)$ dishes. Here, $\gamma > 0$ is called the *mass parameter*, and we will also see the *concentration parameter* $\theta > 0$ below. None of the dishes have been sampled by any other customers since no other customers have yet entered the restaurant. We label the dishes $1, \ldots, K_1^+$ if $K_1^+ > 0$. Recursively, the nth customer chooses which dishes to sample in two parts. First, for each dish k that has previously been sampled by any customer in $1, \ldots, n-1$, customer n samples dish k with probability $N_{n-1,k}/(\theta + n - 1)$ for $N_{n,k}$ equal to the number of customers indexed $1, \ldots, n$ who have tried dish k. As each dish represents a feature, and sampling a dish represents that the customer index n belongs to that feature, $N_{n,k}$ is the size of the block of the feature labeled k in the feature allocation of $[n]$. Next, customer n chooses $K_n^+ \sim \text{Poisson}(\theta\gamma/(\theta + n - 1))$ new dishes to try. If $K_n^+ > 0$, then the dishes receive unique labels $K_{n-1} + 1, \ldots, K_n$. Here, K_n represents the number of sampled dishes after n customers: $K_n = K_{n-1} + K_n^+$. An example of the first few steps in the Indian buffet process is shown in Figure 2.2.

With this generative model in hand, we can find the probability of a particular feature allocation. We discover its form by enumeration as for the CRP EPPF in Example 2.3.1. At each round n, we have a Poisson number of new features, K_n^+, represented. The probability factor associated with these choices is a product of Poisson densities.

$$\prod_{n=1}^{N} \frac{1}{K_n^+!} \left(\frac{\theta\gamma}{\theta + n - 1} \right)^{K_n^+} \exp \left(- \frac{\theta\gamma}{\theta + n - 1} \right)$$

Let M_k be the round on which the kth dish, in order of appearance, is first chosen. Then the denominators for future dish choice probabilities are the factors in the product $(\theta + M_k) \cdot (\theta + M_k + 1) \cdots (\theta + N - 1)$. The numerators for the times when the dish is chosen are the factors in the product $1 \cdot 2 \cdots (N_{N,k} - 1)$. The numerators for the times when the dish is not chosen yield $(\theta + M_k - 1) \cdots (\theta + N - 1 - N_{N,k})$. Let $A_{n,k}$ represent the collection of indices in the feature with label k after n customers have entered the restaurant. Then $N_{n,k} = |A_{n,k}|$. Finally, let $\tilde{K}_1, \ldots, \tilde{K}_H$ be the multiplicities of unique feature blocks formed by this model. We note that there are

$$\left[\prod_{n=1}^{N} K_n^+! \right] / \left[\prod_{h=1}^{H} \tilde{K}_h! \right]$$

rearrangements of the features generated by this process that all yield the same feature allocation. Since they all have the same generating probability, we simply multiply by this factor to find the feature allocation probability. Multiplying all factors together and taking $f_n = \{A_{N,1}, \ldots, A_{N,K_N}\}$ yields

$$\begin{aligned}
&\mathbb{P}(F_N = f_N) \\
&= \frac{\prod_{n=1}^{N} K_n^+!}{\prod_{h=1}^{H} \tilde{K}_h!} \cdot \left[\prod_{n=1}^{N} \frac{1}{K_n^+!} \left(\frac{\theta\gamma}{\theta + n - 1} \right)^{K_n^+} \exp \left(- \frac{\theta\gamma}{\theta + n - 1} \right) \right] \\
&\cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(\theta + M_k)}{\Gamma(\theta + N)} \Gamma(N_{N,k}) \frac{\Gamma(\theta + N - N_{N,k})}{\Gamma(\theta + M_k - 1)} \right] \\
&= \left(\prod_{h=1}^{H} \tilde{K}_h! \right)^{-1} \left[\prod_{n=1}^{N} (\theta\gamma)^{K_n^+} \exp \left(- \frac{\theta\gamma}{\theta + n - 1} \right) \right] \cdot \left[\frac{\prod_{k=1}^{K_N} (\theta + M_k - 1)}{\prod_{n=1}^{N} (\theta + n - 1)^{K_n^+}} \right] \\
&\cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(N_{N,k}) \Gamma(\theta + N - N_{N,k})}{\Gamma(\theta + N)} \right] \\
&= \left(\prod_{h=1}^{H} \tilde{K}_h! \right)^{-1} (\theta\gamma)^{K_N} \exp \left(-\theta\gamma \sum_{n=1}^{N} (\theta + n - 1)^{-1} \right) \prod_{k=1}^{K_N} \frac{\Gamma(N_{N,k}) \Gamma(N - N_{N,k} + \theta)}{\Gamma(N + \theta)}.
\end{aligned}$$

It follows from Eq. (2.5) that the probability of a uniform random ordering of the feature allocation is

$$\begin{aligned}
&\mathbb{P}(\tilde{F}_N = \tilde{f}_N) \qquad (2.8) \\
&= \frac{1}{K_N!} (\theta\gamma)^{K_N} \exp \left(-\theta\gamma \sum_{n=1}^{N} (\theta + n - 1)^{-1} \right)
\end{aligned}$$

$$\cdot \prod_{k=1}^{K_N} \frac{\Gamma(N_{N,k})\Gamma(N - N_{N,k} + \theta)}{\Gamma(N + \theta)}.$$

The distribution of $\tilde{F}_N$ has no dependence on the ordering of the indices in $[N]$. Hence, the distribution of F_N depends only on the same quantities—the number of indices and the feature block sizes—and the feature multiplicities. So we see that the IBP construction yields an exchangeable random feature allocation. Consistency follows from the recursive construction and exchangeability. Therefore, Eq. (2.8) is seen to be in EFPF form (cf. Eq. (2.7)). ■

Above, we have seen two examples of how specifying a conditional distribution for the block membership of index n given the block membership of indices in $[n-1]$ yields an exchangeable probability function: e.g. the EPPF in the CRP case (Example 2.3.1) and the EFPF in the IBP case (Example 2.3.5). This conditional distribution is often called a *prediction rule*, and study of the prediction rule in the clustering case may be referred to as *species sampling* (Pitman, 1996; Hansen and Pitman, 1998; Lee et al., 2008). We will see next that the prediction rule can conversely be recovered from the exchangeable probability function specification and therefore the two are equivalent.

Induced allocations and block labeling

In Examples 2.3.1 and 2.3.5 above, we formed partitions and feature allocations in the following way. For partitions, we assigned labels Z_n to each index n. Then we generated a partition of $[N]$ from the sequence $(Z_n)_{n=1}^N$ by saying that indices m and n are in the same partition block ($m \sim n$) if and only if $Z_n = Z_m$. The resulting partition is called the *induced partition* given the labels $(Z_n)_{n=1}^N$. Similarly, given labels $(Z_n)_{n=1}^\infty$, we can form an induced partition of $\mathbb{N}$. It is easy to check that, given a sequence $(Z_n)_{n=1}^\infty$, the induced partitions of the subsequences $(Z_n)_{n=1}^N$, will be consistent.

In the feature case, we first assigned label collections Y_n to each index n. Y_n is interpreted as a set containing the labels of the features to which n belongs. It must have finite cardinality by our definition of a feature allocation. In this case, we generate a feature allocation on $[N]$ from the sequence $(Y_n)_{n=1}^N$ by first letting $\{\phi_k\}_{k=1}^K$ be the set of unique values in $\bigcup_{n=1}^N Y_n$. Then the features are the collections of indices with shared labels: $f_N = \{\{n : \phi_k \in Y_n\} : k = 1, \ldots, K\}$. The resulting feature allocation f_N is called the *induced feature allocation* given the labels $(Y_n)_{n=1}^N$. Similarly, given label collections $(Y_n)_{n=1}^\infty$, where each Y_n has finite cardinality, we can form an induced feature allocation of $\mathbb{N}$. As in the partition case, given a sequence $(Y_n)_{n=1}^\infty$, we can see that the induced feature allocations of the subsequences $(Y_n)_{n=1}^N$ will be consistent.

In reducing to a partition or feature allocation from a set of labels, we shed the information concerning the labels for each partition block or feature. Conversely, we introduce *order-of-appearance* labeling schemes to give partition blocks or features labels when we have, respectively, a partition or feature allocation.

In the partition case, the order-of-appearance labeling scheme assigns the label 1 to the partition block containing index 1. Recursively, suppose we have seen n indices in k different blocks with labels $\{1, \ldots, k\}$. And suppose the $n+1$st index does not belong to an existing block. Then we assign its block the label $k+1$.

In the feature allocation case, we note that index 1 belongs to K_1^+ features. If $K_1^+ = 0$, there are no features to label yet. If $K_1^+ > 0$, we assign these K_1^+ features labels in $\{1, \ldots, K_1^+\}$. Unless otherwise specified, we suppose that the labels are chosen uniformly at random. Let $K_1 = K_1^+$. Recursively, suppose we have seen n indices and K_n different features with labels $\{1, \ldots, K_n\}$. Suppose the $n+1$st index belongs to K_{n+1}^+ features that have not yet been labeled. Let $K_{n+1} = K_n + K_{n+1}^+$. If $K_{n+1}^+ = 0$, there are no new features to label. If $K_{n+1}^+ > 0$, assign these K_{n+1}^+ features labels in $\{K_n + 1, \ldots, K_{n+1}\}$, e.g. uniformly at random.

We can use these labeling schemes to find the prediction rule, which makes use of partition block and feature labels, from the EPPF or EFPF as appropriate. First, consider a partition with EPPF p. Then, given labels $(Z_n)_{n=1}^N$ with $K_N = \max\{Z_1, \ldots, Z_N\}$, we wish to find the distribution of the label Z_{N+1}. Using an order-of-appearance labeling, we know that either $Z_{N+1} \in \{Z_1, \ldots, Z_N\}$ or $Z_{N+1} = K_N + 1$. Let $\pi_N = \{A_{N,1}, \ldots, A_{N,K_N}\}$ be the partition induced by $(Z_n)_{n=1}^N$. Let $N_{N,k} = |A_{N,k}|$. Let $\mathbb{1}(A)$ be the indicator of event A; that is, $\mathbb{1}(A)$ equals 1 if A holds and 0 otherwise. Let $N_{N+1,k} = N_k + \mathbb{1}\{Z_{N+1} = k\}$ for $k = 1, \ldots, K_{N+1}$, and set $N_{N,K_N+1} = 0$ for completeness. $K_{N+1} = K_N + \mathbb{1}\{Z_{N+1} > K_N\}$ is the number of partition blocks in the partition of $[N+1]$. Then the conditional distribution satisfies

$$\mathbb{P}(Z_{N+1} = z | Z_1, \ldots, Z_N) = \frac{\mathbb{P}(Z_1, \ldots, Z_N, Z_{N+1} = z)}{\mathbb{P}(Z_1, \ldots, Z_N)}.$$

But the probability of a certain labeling is just the probability of the underlying partition in this construction, so

$$\mathbb{P}(Z_{N+1} = z | Z_1, \ldots, Z_N) = \frac{p(N_{N+1,1}, \ldots, N_{N+1,K_{N+1}})}{p(N_{N,1}, \ldots, N_{N,K_N})}.$$

Example 2.3.6 (Chinese restaurant process)**.** We continue our Chinese restaurant process example by deriving the Chinese restaurant table assignment scheme from the EPPF in Eq. (2.4). Substituting in the EPPF for the CRP, we find

$$\begin{aligned}
&\mathbb{P}(Z_{N+1} = z | Z_1, \ldots, Z_N) \\
&= \frac{p(N_{N,1}, \ldots, N_{N+1,K_{N+1}})}{p(N_{N,1}, \ldots, N_{N,K_N})} \\
&= \frac{\left(\theta^{K_{N+1}-1} \prod_{k=1}^{K_{N+1}} (N_{N+1,k} - 1)!\right) \left((\theta+1)_{(N+1)-1\uparrow 1}\right)^{-1}}{\left(\theta^{K_N-1} \prod_{k=1}^{K_N} (N_{N,k} - 1)!\right) \left((\theta+1)_{N-1\uparrow 1}\right)^{-1}} \\
&= (N+\theta)^{-1} \left\{ \begin{array}{ll} N_{N,k} & \text{for } z = k \leq K_N \\ \theta & \text{for } z = K_N + 1 \end{array} \right. , \qquad (2.9)
\end{aligned}$$

just as in Eq. (2.3). ■

To find the feature allocation prediction rule, we now imagine a feature allocation with EFPF p. Here we must be slightly more careful about counting due to feature multiplicities. Suppose that after N indices have been seen, we have label collections $(Y_n)_{n=1}^N$, containing a total of K_N features, labeled $\{1, \ldots, K_N\}$. We wish to find the distribution of Y_{N+1}. Suppose $N+1$ belongs to K_{N+1}^+ features that do not contain any index in $[N]$. Using an order-of-appearance labeling, we know that, if $K_{N+1}^+ > 0$, the K_{N+1}^+ new features have labels $K_N + 1, \ldots, K_N + K_{N+1}^+$. Let $f_N = \{A_1, \ldots, A_{K_N}\}$ be the feature allocation induced by $(Y_n)_{n=1}^N$. Let $N_{N,k} = |A_{N,k}|$ be the size of the kth feature. So $N_{N+1,k} = N_{N,k} + \mathbb{1}\{k \in Y_{N+1}\}$, where we let $N_{K_N+j} = 0$ for all of the features that are first exhibited by index $N+1$: $j \in \{1, \ldots, K_{N+1}^+\}$. Further, let the number of features, including new ones, be written $K_{N+1} = K_N + K_{N+1}^+$. Then the conditional distribution satisfies

$$\mathbb{P}(Y_{n+1} = y | Y_1, \ldots, Y_N) = \frac{\mathbb{P}(Y_1, \ldots, Y_N, Y_{N+1} = y)}{\mathbb{P}(Y_1, \ldots, Y_N)}.$$

As we assume that the labels Y are consistent across N, the probability of a certain labeling is just the probability of the underlying ordered feature allocation times a combinatorial term. The combinatorial term accounts first for the uniform ordering of the new features amongst themselves for labeling and then for the uniform ordering of the new features amongst the old features in the overall uniform random ordering.

$$\begin{aligned}\mathbb{P}(Y_{N+1} = y | Y_1, \ldots, Y_N) &= \frac{1}{K_{N+1}^+!} \cdot [(K_N + 1) \cdot (K_N + 2) \cdots K_{N+1}] \\ &\quad \cdot \frac{p(N, N_{N+1,1}, \ldots, N_{N+1,K_{N+1}})}{p(N, N_{N,1}, \ldots, N_{N,K_N})} \\ &= \frac{1}{K_{N+1}^+!} \cdot \frac{K_{N+1}!}{K_N!} \cdot \frac{p(N, N_{N+1,1}, \ldots, N_{N+1,K_{N+1}})}{p(N, N_{N,1}, \ldots, N_{N,K_N})}. \qquad (2.10)\end{aligned}$$

Example 2.3.7 (Indian buffet process)**.** Just as we derived the Chinese restaurant process prediction rule (Eq. (2.9)) from its EPPF (Eq. (2.4)) in Example 2.3.6, so can we derive the Indian buffet process prediction rule from its EFPF (Eq. (2.8)) by using Eq. (2.10). Substituting the IBP EFPF into Eq. (2.10), we find

$$\begin{aligned}&\mathbb{P}(Y_{n+1} = y | Y_1, \ldots, Y_N) \\ &= \frac{1}{K_{N+1}^+!} \cdot \frac{K_{N+1}!}{K_N!} \\ &\cdot \frac{\frac{1}{K_{N+1}!} (\theta\gamma)^{K_{N+1}} \exp\left(-\theta\gamma \sum_{n=1}^{N+1} (\theta + n - 1)^{-1}\right) \prod_{k=1}^{K_{N+1}} \frac{\Gamma(N_{N+1,k}) \Gamma((N+1) - N_{N+1,k} + \theta)}{\Gamma((N+1)+\theta)}}{\frac{1}{K_N!} (\theta\gamma)^{K_N} \exp\left(-\theta\gamma \sum_{n=1}^{N} (\theta + n - 1)^{-1}\right) \prod_{k=1}^{K_N} \frac{\Gamma(N_{N,k}) \Gamma(N - N_{N,k} + \theta)}{\Gamma(N+\theta)}} \\ &= \left[\frac{1}{K_{N+1}^+!} \exp\left(-\frac{\theta\gamma}{\theta + (N+1) - 1}\right) \cdot \left(\frac{\theta\gamma}{\theta + (N+1) - 1}\right)^{K_{N+1}^+} \right]\end{aligned}$$

$$\cdot (\theta + (N+1) - 1)^{K^+_{N+1}} \cdot \left[\prod_{k=K_N+1}^{K_{N+1}} (\theta + (N+1) - 1)^{-1} \right]$$

$$\cdot \prod_{k=1}^{K_N} \frac{N_k^{\mathbb{1}\{k \in z\}} (N - N_{N,k} + \theta)^{\mathbb{1}\{k \notin z\}}}{N + \theta}$$

$$= \text{Poisson}\left(K^+_{N+1} | \frac{\theta\gamma}{\theta + (N+1) - 1} \right) \cdot \prod_{k=1}^{K_N} \text{Bern}\left(\mathbb{1}\{k \in z\} | \frac{N_{N,k}}{N + \theta} \right).$$

The final line is exactly the Poisson distribution for the number of new features times the Bernoulli distributions for the draws of existing features, as described in Example 2.3.5. ■

Inference

The prediction rule formulation of the EPPF or EFPF is particularly useful in providing a means of inferring partitions and feature allocations from a data set. In particular, we assume that we have data points $X_1, \ldots, X_N$ generated in the following manner. In the partition case, we generate an exchangeable, consistent, random partition Π_N according to the distribution specified by some EPPF p. Next, we assign each partition block a random parameter that characterizes that block. To be precise, for the kth partition block to appear according to an order-of-appearance labeling scheme, give this block a new *random* label $\phi_k \sim H$, for some continuous distribution H. For each n, let $Z_n = \phi_k$ where k is the order-of-appearance label of index n. Finally, let

$$X_n \overset{indep}{\sim} \mathcal{L}(Z_n), \tag{2.11}$$

for some distribution $\mathcal{L}$ with parameter Z_n. The choices of both H and $\mathcal{L}$ are specific to the problem domain.

Without attempting to survey the vast literature on clustering, we describe a stylized example to provide intuition for the preceding generative model. In this example, let n index an animal observed in the wild; $Z_n = Z_m$ indicates that animals n and m belong to the same (latent, unobserved) species; $Z_n = Z_m = \phi_k$ is a vector describing the (latent, unobserved) height and weight for that species; and X_n is the observed height and weight of the nth animal.

X_n need not even be directly observed, but Eq. (2.11) together with an EPPF might be part of a larger generative model. In a generalization of the previous stylized example, Z_n indicates the dominant species in the nth geographical region; $Z_n = \phi_k$ indicates some overall species height and weight parameters (for the kth species); X_n indicates the height and weight parameters for species k in the nth region. That is, the height and weight for the species may vary by region. We measure and observe the height and weight $(E_{n,j})_{j=1}^J$ of some J animals in the nth region, believed to be iid draws from a distribution depending on X_n.

Note that the sequence $(Z_n)_{n=1}^N$ is sufficient to describe the partition Π_N since Π_N is the collection of blocks of $[N]$ with the same label values Z_n. The continuity of H is necessary to guarantee the a.s. uniqueness of the block values. So, if we can describe the posterior distribution of $(Z_n)_{n=1}^N$, we can in principle describe the posterior distribution of Π_N.

The posterior distribution of $(Z_n)_{n=1}^N$ conditional on $(X_n)_{n=1}^N$ cannot typically be solved for in closed form, so we turn to a method that approximates this posterior. We will see that prediction rules facilitate the design of a Markov Chain Monte Carlo (MCMC) sampler, in which we approximate the desired posterior distribution by a Markov chain of random samples proven to have the true posterior as its equilibrium distribution.

In the Gibbs sampler formulation of MCMC (S. Geman and D. Geman, 1984), we sample each parameter in turn and conditional on all other parameters in the model. In our case, we will sequentially sample each element of $(Z_n)_{n=1}^N$. The key observation here is that $(Z_n)_{n=1}^N$ is an exchangeable sequence. This observation follows by noting that the partition is exchangeable by assumption, and the sequence (ϕ_k) is exchangeable since it is iid; (Z_n) is an exchangeable sequence since it is a function of (Π_n) and (ϕ_k). Therefore, the distribution of Z_n given the remaining elements $\mathbf{Z}_{-n} := (Z_1, \ldots, Z_{n-1}, Z_{n+1}, \ldots, Z_N)$ is the same as if we thought of Z_n as the final, Nth element in a sequence with $N-1$ preceding values given by $\mathbf{Z}_{-n}$. And the distribution of Z_N given $\mathbf{Z}_{-N}$ is provided by the prediction rule. The full details of the Gibbs sampler for the CRP in Examples 2.3.1 and 2.3.6 were introduced by Escobar (1994); S. N. MacEachern (1994); Escobar and West (1995) and are covered in fuller generality by Neal (2000).

It is worth noting that the sequence of order-of-appearance labels is not exchangeable; for instance, the first label is always 1. However, the prediction rule for Z_N given $(Z_1, \ldots, Z_{N-1})$ breaks into two parts: (1) the probability of Z_N taking a value either in $\{Z_1, \ldots, Z_{N-1}\}$ or a new value and (2) the distribution of Z_N when it takes a new value. When programming such a sampler, it is often useful to simply encode the sets of unique values, which may be done by retaining any set of labels that induce the correct partition (e.g. integer labels) and separately retaining the set of unique parameter values. Indeed, updating the parameter values and partition block assignments separately can lead to improved mixing of the sampler (S. N. MacEachern, 1994).

Similarly, in the feature case, we imagine the following generative model for our data. First, let F_N be a random feature allocation generated according to the EFPF p. For the kth feature block in an order-of-appearance labeling scheme, assign a random label $\phi_k \sim H$ to this block for some continuous distribution H. For each n, let $Y_n = \{\phi_k : k \in J_n\}$, where J_n is here the set of order-of-appearance labels of the features to which n belongs. Finally, as above,

$$X_n \overset{indep}{\sim} \mathcal{L}(Y_n),$$

where the likelihood $\mathcal{L}$ and parameter distribution H are again application-specific and where now $\mathcal{L}$ depends on the variable-size collection of parameters in Y_n.

Griffiths and Ghahramani (2011) provide a review of likelihoods used in practice for feature models. To motivate some of these modeling choices, let us consider some stylized examples that provide helpful intuition. For example, let n index customers at a book-selling website; ϕ_k describes a book topic such as economics, modern art, or science fiction. If ϕ_k describes science fiction books, $\phi_k \in Y_n$ indicates that the nth customer likes to buy science fiction books. But Y_n might have cardinality greater than one (the customer is interested in multiple book topics) or cardinality zero (the customer never buys books). Finally, X_n is a set of book sales for customer n on the book-selling site.

As a second example, let n index pictures in a database; ϕ_k describes a pictorial element such as a train or grass or a cow; $\phi_k \in Y_n$ indicates that picture n contains, e.g., a train; finally, the observed array of pixels X_n that form the picture is generated to contain the pictorial elements in Y_n. As in the clustering case, X_n might not even be directly observed but might serve as a random effect in a deeper hierarchical model.

We observe that although the order-of-appearance label sets are not exchangeable, the sequence (Y_n) is. This fact allows the formulation of a Gibbs sampler via the observation that the distribution of Y_n given the remaining elements $\mathbf{Y}_{-n} := (Y_1, \ldots, Y_{n-1}, Y_{n+1}, \ldots, Y_N)$ is the same as if we thought of Y_n as the final, Nth element in a sequence with $N - 1$ preceding values given by $\mathbf{Y}_{-n}$. The full details of such a sampler for the case of the IBP (Examples 2.3.5 and 2.3.7) are given by Griffiths and Ghahramani (2006).

As in the partition case, in practice when programming the sampler, it is useful to separate the feature allocation encoding from the feature parameter values. Griffiths and Ghahramani (2006) describe how *left order form* matrices give a convenient representation of the feature allocation in this context.

2.4 Stick lengths

Not every symmetric function defined for an arbitrary number of arguments with values in the unit interval is an EPPF (Pitman, 1995), and not every symmetric function with an additional positive integer argument is an EFPF. For instance, the consistency property in Eq. (2.1) implies certain additivity requirements for the function p.

Example 2.4.1 (Not an EPPF)**.** Consider the function p defined with

$$p(1) = 1, \quad p(1,1) = 0.1, \quad p(2) = 0.8, \quad \ldots \tag{2.12}$$

From the information in Eq. (2.12), p may be further defined so as to be symmetric in its arguments for any number of arguments, but since it does not satisfy $p(1) = p(1,1) + p(2)$, it cannot be an EPPF. ■

Example 2.4.2 (Not an EFPF)**.** Consider the function p defined with

$$p(N = 1) = 0.9, \quad p(N = 1, 1) = 0.9, \quad p(N = 1, 1, 1) = 0.9, \quad \ldots \tag{2.13}$$

From the information in Eq. (2.13), p may be further defined so as to be symmetric in its arguments for any number of arguments after the initial N argument, but since $p(N = 1) + p(N = 1, 1) + p(N = 1, 1, 1) > 1$, it cannot be an EFPF. ■

It therefore requires some care to define a suitable distribution over consistent, exchangeable random feature allocations or partitions using the exchangeable probability function framework.

Since we are working with exchangeable sequences of random variables, it is natural to turn to de Finetti's theorem (De Finetti, 1931; Hewitt and Savage, 1955) for clues as to how to proceed. De Finetti's theorem tells us that any exchangeable sequence of random variables can be expressed as an independent and identically distributed sequence when conditioned on an underlying random *mixing measure*. While this theorem may seem difficult to apply directly to, e.g., exchangeable partitions, it may be applied more naturally to an exchangeable sequence of numbers derived from a sequence of partitions. The argument below is due to Aldous (1985).

Suppose that (Π_n) is an exchangeable, consistent sequence of random partitions. Consider the kth partition block to appear according to an order-of-appearance labeling scheme, and give this block a new *random* label, $\phi_k \sim \text{Unif}([0,1])$, such that each random label is drawn independently from the rest. This construction is the same as the one used for parameter generation in Section 2.3, and (Π_n) is exchangeable by the same arguments used there. Let Z_n equal ϕ_k exactly when n belongs to the partition with this label.

If we apply de Finetti's theorem to the sequence (Z_n) and note that (Z_n) has at most countably many different values, we see that there exists some random sequence (ρ_k) such that $\rho_k \in (0,1]$ for all k and, conditioned on the frequencies (ρ_k), (Z_n) has the same distribution as iid draws from (ρ_k). In this description, we have brushed over technicalities associated with partition blocks that contain only one index even as $N \to \infty$ (which may imply $\sum_k \rho_k < 1$).

But if we assume that every partition block eventually contains at least two indices, we can achieve an exchangeable partition of $[N]$ as follows. Let (ρ_k) represent a sequence of values in $(0,1]$ such that $\sum_{k=1}^{\infty} \rho_k \overset{a.s.}{=} 1$. Draw $Z_n \overset{iid}{\sim} \text{Discrete}((\rho_k)_k)$. Let Π_N be the induced partition given $(Z_n)_{n=1}^N$. Exchangeability follows from the iid draws, and consistency follows from the induced partition construction.

When the frequencies (ρ_k) are thought of as subintervals of the unit interval, i.e. a partition of the unit interval, they are collectively called *Kingman's paintbox* (Kingman, 1978). As another naming convention, we may think of the unit interval as a *stick* (Ishwaran and James, 2001). We partition the unit interval by breaking it into various *stick lengths*, which represent the frequencies of each partition block.

A similar construction can be seen to yield exchangeable, consistent random feature allocations. In this case, let (ξ_k) represent a sequence of values in $(0,1]$ such that $\sum_{k=1}^{\infty} \xi_k \overset{a.s.}{<} \infty$. We generate feature collections independently for each index as follows. Start with $Y_n = \emptyset$. For each feature k, add k to the set Y_n, independently from all other features, with probability ξ_k. Let F_N be the induced feature allocation given $(Y_n)_{n=1}^N$. Exchangeability of F_N follows from the iid draws of Y_n, and consistency follows from the induced feature allocation construction. The finite sum constraint ensures each index belongs to a finite number of features a.s.

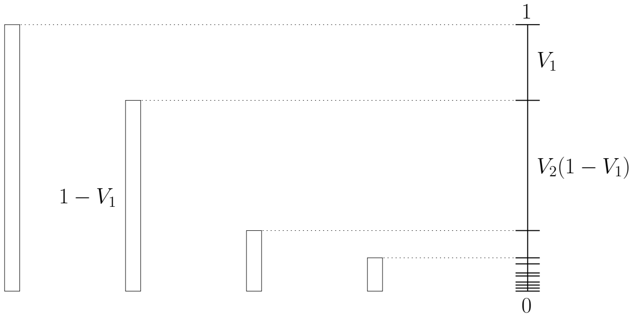


Figure 2.3: An illustration of how stick-breaking divides the unit interval into a sequence of probabilities (Broderick, Jordan, and Pitman, 2012). The stick proportions $(V_1, V_2, \cdots)$ determine what fraction of the remaining stick is appended to the probability sequence at each round.

It remains to specify a distribution on the partition or feature frequencies. The frequencies cannot be iid due to the finite summation constraint in both cases. In the partition case, any infinite set of frequencies cannot even be independent since the summation is fixed to one. One scheme to ensure summation to unity is called *stick-breaking* (McCloskey, 1965; Patil and Taillie, 1977; Sethuraman, 1994; Ishwaran and James, 2001). In stick-breaking, the stick lengths are obtained by recursively breaking off parts of the unit interval to return as the atoms $\rho_1, \rho_2, \ldots$ (cf. Figure 2.3). In particular, we generate stick-breaking proportions $V_1, V_2, \ldots$ as $[0,1]$-valued random variables. Then ρ_1 is the first proportion V_1 times the initial stick length 1; hence $\rho_1 = V_1$. Recursively, after k breaks, the remaining length of the initial unit interval is $\prod_{j=1}^{k}(1-V_j)$. And ρ_{k+1} is the proportion V_{k+1} of the remaining stick; hence $\rho_{k+1} = V_{k+1}\prod_{j=1}^{k}(1-V_j)$.

The stick-breaking construction yields $\rho_1, \rho_2, \ldots$ such that $\rho_k \in [0,1]$ for each k and $\sum_{k=1}^{\infty} \rho_k \le 1$. If the V_k do not decay too rapidly, we will have $\sum_{k=1}^{\infty} \rho_k \overset{a.s.}{=} 1$. In particular, the partition block proportions ρ_k sum to unity a.s. iff there is no remaining stick mass: $\prod_{k=1}^{\infty}(1-V_k) \overset{a.s.}{=} 0$.

We often make the additional, convenient assumption that the V_k are independent. In this case, a necessary and sufficient condition for $\sum_{k=1}^{\infty} \rho_k \overset{a.s.}{=} 1$ is $\sum_{k=1}^{\infty} \mathbb{E}\left[\log(1-V_k)\right] = -\infty$ (Ishwaran and James, 2001). When the V_k are independent and of a canonical distribution, they are easily simulated. Moreover, if we assume that the V_k are such that the ρ_k decay sufficiently rapidly in k, one strategy for simulating a stick-breaking model is to ignore all $k > K$ for some fixed, finite K. This approximation is known as truncation (Ishwaran and James, 2001). It is fortuitously the case that in some models of particular interest, such useful assumptions fall out naturally from the model construction (e.g. Examples 2.4.3 and 2.4.4).

Example 2.4.3 (Chinese restaurant process)**.** In the original exchangeability result due to de Finetti (De Finetti, 1931), the exchangeable random variables were zero/one-valued,

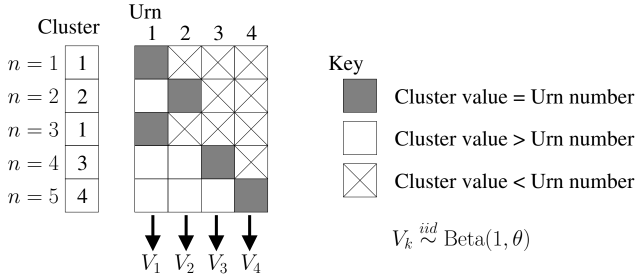

Cluster
Urn
1 2 3 4
$n = 1$ 1
$n = 2$ 2
$n = 3$ 1
$n = 4$ 3
$n = 5$ 4
V_1 V_2 V_3 V_4
Key
Cluster value = Urn number
Cluster value > Urn number
Cluster value < Urn number
$V_k \overset{iid}{\sim} \mathrm{Beta}(1, \theta)$

Figure 2.4: An illustration of the proof based on the Pólya urn that Dirichlet process stick-breaking gives the underlying partition block frequencies for a Chinese restaurant process model. The kth column in the central matrix corresponds to a tallying of when the kth table is chosen (gray), when a table of index larger than k is chosen (white), and when an index smaller than k is chosen ($\times$). If we ignore the $\times$ tallies, the gray and white tallies in each column (after the first) can be modeled as balls drawn from a Pólya urn. The limiting frequency of gray balls in each column is shown below the matrix.

and the mixing measure was a distribution on a single frequency so that the outcomes were conditionally Bernoulli. We will find a similar result in obtaining the stick-breaking proportions associated with the Chinese restaurant process.

We can construct a sequence of binary-valued random variables by dividing the customers in the CRP who are sitting at the first table from the rest; color the former collection of customers gray and the latter collection of customers white. Then, we see that the first customer must be colored gray. And thus we begin with a single gray customer and no white customers. This binary valuation for the first table in the CRP is illustrated by the first column in the matrix in Figure 2.4.

At this point, it is useful to recall the Pólya urn construction (Pólya, 1930; D. A. Freedman, 1965), whereby an urn starts with G_0 gray balls and W_0 white balls. At each round N, we draw a ball from the urn, replace it, and add κ of the same color of ball to the urn. At the end of the round, we have G_N gray balls and W_N white balls. Despite the urn metaphor, the number of balls need not be an integer at any time. By checking Eq. (2.3) which defines the CRP, we can see that the coloring of the gray/white customer matrix assignments starting with the second customer has the same distributions as a sequence of balls from a Pólya urn as a Pólya urn with $G_{1,0} = 1$ initial gray balls, $W_{1,0} = \theta$ initial white balls, and $\kappa_1 = 1$ replacement balls. Let $G_{1,N}$ and $W_{1,N}$ represent the numbers of gray and white balls, respectively, in the urn after N rounds. The important fact about the Pólya urn we use here is that there exists some $V \sim \mathrm{Beta}(G_0/\kappa, W_0/\kappa)$ such that $\kappa^{-1}(G_{N+1} - G_N) \overset{iid}{\sim} \mathrm{Bern}(V)$ for all N. In this particular case of the CRP, then, $G_{1,N+1} - G_{1,N}$ is one if a customer sits at the first table (or zero otherwise), and $G_{1,N+1} - G_{1,N} \overset{iid}{\sim} \mathrm{Bern}(V_1)$ with $V_1 \sim \mathrm{Beta}(1, \theta)$.

We now look at the sequence of customers who sit at the second and subsequent tables. That is, we condition on customers not sitting at the first table or equivalently on the sequence with $G_{1,N+1} - G_{1,N} = 0$. Again, we have that the first customer sits at the second table, by the CRP construction. Now let customers at the second table be colored gray and customers at the third and later tables be colored white. This valuation is illustrated in the second column in Figure 2.4; each × in the figure denotes a data point where the first partition block is chosen and therefore the current Pólya urn is not in play. As before, we begin with one gray customer and no white customers. We can check Eq. (2.3) to see that customer coloring once more proceeds according to a Pólya urn scheme with $G_{2,0} = 1$ initial gray balls, $W_{2,0} = \theta$ initial white balls, and $\kappa_2 = 1$ replacement balls. Thus, contingent on a customer not sitting at the first table, the Nth customer sits at the second table with iid distribution $\mathrm{Bern}(V_2)$ with $V_2 \sim \mathrm{Beta}(1, \theta)$. Since the sequence of individuals sitting at the second table has no other dependence on the sequence of individuals sitting at the first table, we have that V_2 is independent of V_1.

The argument just outlined proceeds recursively to show us that the Nth customer, conditional on not sitting at the first $K-1$ tables for $K \geq 1$, sits at the Kth table with iid distribution $\mathrm{Bern}(V_K)$ and $V_K \sim \mathrm{Beta}(1, \theta)$ with V_K independent of the previous $(V_1, \ldots, V_{K-1})$.

Combining these results, we see that we have the following construction for the customer seating patterns. The V_k are distributed independently and identically according to $\mathrm{Beta}(1, \theta)$. The probability ρ_K of sitting at the Kth table is the probability of not sitting at the first $K-1$ tables, conditional on not sitting at the previous table, times the conditional probability of sitting at the Kth table: $\rho_K = \left[\prod_{k=1}^{K-1}(1 - V_k)\right] \cdot V_K$. Finally, with the vector of table frequencies (ρ_k), each customer sits independently and identically at the corresponding vector of tables according to these frequencies. This process is summarized here:

$$\begin{aligned} V_k &\overset{iid}{\sim} \mathrm{Beta}(1, \theta) \\ \rho_K &:= V_K \prod_{k=1}^{K} (1 - V_k) \\ Z_n &\overset{iid}{\sim} \mathrm{Discrete}((\rho_k)_k). \end{aligned} \tag{2.14}$$

To see that this process is well-defined, first note that $\mathbb{E}\left[\log(1 - V_k)\right]$ exists, is negative, and is the same for all k values. It follows that $\sum_{k=1}^{\infty} \mathbb{E}\left[\log(1 - V_k)\right] = -\infty$, so by the discussion before this example, we must have $\sum_{k=1}^{K} \rho_k \overset{a.s.}{=} 1$. ■

The feature case is easier. Since it does not require the frequencies to sum to one, the random frequencies can be independent so long as they have an a.s. finite sum.

Example 2.4.4 (Indian buffet process)**.** As in the case of the CRP, we can recover the stick lengths for the Indian buffet process using an argument based on an urn model.

Recall that on the first round of the Indian buffet process, $K_1^+ \sim \mathrm{Poisson}(\gamma)$ features are chosen to contain index 1. Consider one of the features, labeled k. By construction, each future data point N belongs to this feature with probability $N_{N-1,k}/(\theta + N - 1)$. Thus, we can model the sequence after the first data point as a Pólya urn of the sort encountered

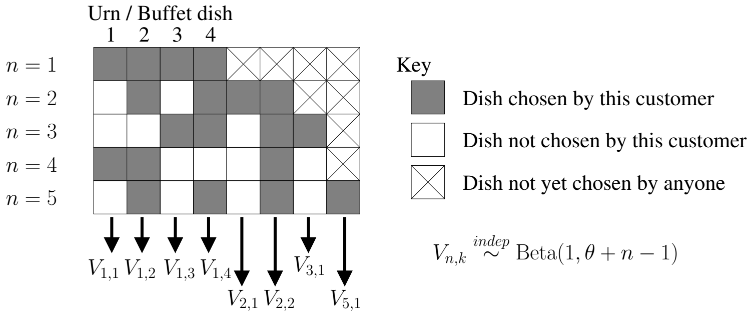

Urn / Buffet dish
1 2 3 4
$n = 1$
$n = 2$
$n = 3$
$n = 4$
$n = 5$
Key
Dish chosen by this customer
Dish not chosen by this customer
Dish not yet chosen by anyone
$V_{1,1}$ $V_{1,2}$ $V_{1,3}$ $V_{1,4}$
$V_{2,1}$ $V_{2,2}$
$V_{3,1}$
$V_{5,1}$
$V_{n,k} \overset{indep}{\sim} \text{Beta}(1, \theta + n - 1)$

Figure 2.5: Illustration of the proof that the frequencies of features in the Indian buffet process are given by beta random variables. For each feature, we can construct a sequence of zero/one variables by tallying whether (gray, one) or not (white, zero) that feature is represented by the given data point. Before the first time a feature is chosen, we mark it with an ×. Each column sequence of gray and white tallies, where we ignore the × marks, forms a Pólya urn with limiting frequencies shown below the matrix.

in Example 2.4.3 with initially $G_{k,0} = 1$ gray balls, $W_{k,0} = \theta$ white balls, and $\kappa_k = 1$ replacement balls. As we have seen, there exists a random variable $V_k \sim \text{Beta}(1, \theta)$ such that representation of this feature by data point N is chosen, iid across all N, as $\text{Bern}(V_k)$. Since the Bernoulli draws conditional on previous draws are independent across all k, the V_k are likewise independent of each other; this fact is also true for k in future rounds. Draws according to such an urn are illustrated in each of the first four columns of the matrix in Figure 2.5.

Now consider any round n. According to the IBP construction, $K_n^+ \sim \text{Poisson}(\gamma\theta/(\theta + n - 1))$ new features are chosen to include index n. Each future data point N (with $N > n$) represents feature k among these features with probability $N_{N-1,k}/(\theta + N - 1)$. In this case, we can model the sequence after the nth data point as a Pólya urn with $G_{k,0} = 1$ initial gray balls, $W_{k,0} = \theta + n - 1$ initial white balls, and $\kappa_k = 1$ replacement balls. So there exists a random variable $V_k \sim \text{Beta}(1, \theta + n - 1)$ such that representation of feature k by data point N is chosen, iid across all N, as $\text{Bern}(V_k)$.

Finally, then, we have the following generative model for the feature allocation by iterating across $n = 1, \ldots, N$ (Thibaux and Jordan, 2007):

$$K_n^+ \overset{indep}{\sim} \text{Poisson}\left(\frac{\gamma\theta}{\theta + n - 1}\right) \tag{2.15}$$

$$K_n = K_{n-1} + K_n^+$$

$$V_k \overset{indep}{\sim} \text{Beta}(1, \theta + n - 1), \quad k = K_{n-1} + 1, \ldots, K_n \tag{2.16}$$

$$I_{n,k} \overset{indep}{\sim} \text{Bern}(V_k), \quad k = 1, \ldots, K_n$$

$I_{n,k}$ is an indicator random variable for whether feature k contains index n. The collection of features to which index n belongs, Y_n, is the collection of features k with $I_{n,k} = 1$. ■

Inference

As we have seen above, the exchangeable probability functions of Section 2.3 are the marginal distributions of the partitions or feature allocations generated according to stick-length models with the stick lengths integrated out. It has been proposed that including the stick lengths in MCMC samplers of these models will improve mixing (Ishwaran and Zarepour, 2000). While it is impossible to sample the countably infinite set of partition block or feature frequencies in these models (cf. Examples 2.4.3 and 2.4.4), a number of ways of getting around this difficulty have been investigated. Ishwaran and Zarepour (2000) examine two separate finite approximations to the full CRP stick length model; one uses a parametric approximation to the full infinite model, and the other creates a truncation by setting the stick break at some fixed size K to be 1: $V_K = 1$. There also exist techniques that avoid any approximations and deal instead directly with the full model: in particular, retrospective sampling (Papaspiliopoulos and Roberts, 2008) and slice sampling (Walker, 2007).

While our discussion thus far has focused on MCMC sampling as a means of approximating the posterior distribution of either the block assignments or both the block assignments and stick lengths, including the stick lengths in a posterior analysis facilitates a different posterior approximation; in particular, *variational methods* can also be used to approximate the posterior. These methods minimize some notion of distance to the posterior over a family of potential approximating distributions (Jordan et al., 1999). The practicality and, indeed, speed of these methods in the case of stick-breaking for the CRP (Example 2.4.3) have been demonstrated by Blei and Jordan (2006).

A number of different models for the stick lengths corresponding to the features of an IBP (Example 2.4.4) have been discovered. The distributions described in Example 2.4.4 are covered by Thibaux and Jordan (2007), who build on work from Hjort (1990); Kim (1999b). A special case of the IBP is examined by Teh, Görür, and Ghahramani (2007), who detail a slice sampling algorithm for sampling from the posterior of the stick lengths and feature assignments. Yet another stick length model for the IBP is explored by Paisley, Zaas, et al. (2010), who show how to apply variational methods to approximate the posterior of their model.

Stick length modeling has the further advantage of allowing inference in cases where it is not straightforward to integrate out the underlying stick lengths to obtain a tractable exchangeable probability function.

2.5 Subordinators

An important point to reiterate about the labels Z_n and label collections Y_n is that when we use the order-of-appearance labeling scheme for partition or feature blocks described above, the random sequences (Z_n) and (Y_n) are not exchangeable. Often, however, we would like to make use of special properties of exchangeability when dealing with these sequences. For instance, if we use Markov Chain Monte Carlo to sample from the posterior distribution of a partition (cf. Section 2.3), we might want to Gibbs sample the cluster assignment of data point n given the assignments of the remaining data points: that is, Z_n given $\{Z_m\}_{m=1}^{N} \backslash \{Z_n\}$. This sampling is particularly easy in some cases (Neal, 2000) if we can treat Z_n as the last random variable in the sequence, but this treatment requires exchangeability.

A way to get around this dilemma was suggested by Aldous (1985) and appeared above in our motivation for using stick lengths. Namely, we assign to the kth partition block a uniform random label $\phi_k \sim \mathrm{Unif}([0,1])$; analogously, we assign to the kth feature a uniform random label $\phi_k \sim \mathrm{Unif}([0,1])$. We can see that in both cases, all of the labels are a.s. distinct. Now, in the partition case, let Z_n be the uniform random label of the partition block to which n belongs. And in the feature case, let Y_n be the (finite) set of uniform random feature labels for the features to which n belongs. We can recover the partition or feature allocation as the induced partition or feature allocation by grouping indices assigned to the same label. Moreover, as discussed above, we now have that each of (Z_n) and (Y_n) is an exchangeable sequence.

If we form partitions or features according to the stick length constructions detailed in Section 2.4, we know that each unique partition or feature label ϕ_k is associated with a frequency ξ_k. We can use this association to form a random measure:

$$\mu = \sum_{k=1}^{\infty} \xi_k \delta_{\phi_k}, \tag{2.17}$$

where δ_{ϕ_k} is a unit point mass located at ϕ_k. In the partition case, $\sum_k \xi_k = 1$, so the random measure is a random probability measure, and we may draw $Z_n \overset{iid}{\sim} \mu$. In the feature case, the weights have a finite sum but do not necessarily sum to one. In the feature case, we draw Y_n by including each ϕ_k for which $\mathrm{Bern}(\xi_k)$ yields a draw of 1.

Another way to codify the random measure in Eq. (2.17) is as a monotone increasing stochastic process on $[0,1]$. Let

$$T_s = \sum_{k=1}^{\infty} \xi_k \mathbb{1}\{\phi_k \le s\}.$$

Then the atoms of μ are in one-to-one correspondence with the jumps of the process T.

This increasing random function construction gives us another means of choosing distributions for the weights ξ_k. We have already seen that these cannot be iid due to the finite summation condition. However, we will see that if we require that the *increments* of a monotone, increasing stochastic process are independent and stationary, then we can use the jumps of that function as the atoms in our random measure for partitions or features.

Definition 2.5.1. A *subordinator* (Bochner, 1955; Bertoin, 1998; Bertoin, 2004) is a stochastic process $(T_s, s \ge 0)$ that has

- Non-negative, non-decreasing paths (a.s.),

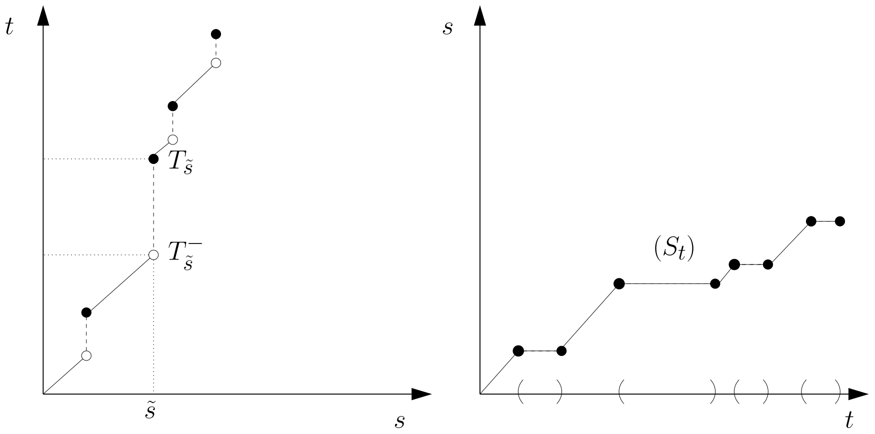


Figure 2.6: *Left*: The sample path (T_s) of a subordinator. $T_{\tilde{s}}^-$ is the limit from the left of (T_s) at $s = \tilde{s}$. *Right*: The right-continuous inverse (S_t) of a subordinator: $S_t := \inf\{s : T_s > t\}$. The open intervals along the t axis correspond to the jumps of the subordinator (T_s).

- Paths that are right-continuous with left limits, and
- Stationary, independent increments.

For our purposes, wherein the subordinator values will ultimately correspond to (perhaps scaled) probabilities, we will assume the subordinator takes values in $[0, \infty)$ though alternative ranges with a sense of ordering are possible.

Subordinators are of interest to us because not only do they exhibit the stationary, independent increments property but they can always be decomposed into two components: a deterministic *drift* component and a *Poisson point process*. Recall that a Poisson point process on space S with rate measure $\nu(dx)$, where $x \in S$, yields a countable subset of points of S. Let $N(A)$ be the number of points of the process in set A for $A \subseteq S$. The process is characterized by the fact that, first, $N(A) \sim \text{Poisson}(\nu(A))$ for any A and, second, for any disjoint $A_1, \ldots, A_K$, we have that $N(A_1), \ldots, N(A_K)$ are independent random variables. See Kingman (1993) for a thorough treatment of these processes. An example subordinator with both drift and jump components is shown on the lefthand side of Figure 2.6.

The subordinator decomposition is detailed in the following result (Bertoin, 1998).

Theorem 2.5.2. *Every subordinator $(T_s, s \geq 0)$ can be written as*

$$T_s = cs + \sum_{k=1}^{\infty} \xi_k \mathbb{1}\{\phi_k \leq s\}, \tag{2.18}$$

for some constant $c \geq 0$ and where $\{(\xi_k, \phi_k)\}_k$ is the countable set of points of a Poisson point process with intensity $\Lambda(d\xi)\ d\phi$, where Λ is a Lévy measure, i.e.

$$\int_0^\infty (1 \wedge \xi) \Lambda(d\xi) < \infty.$$

In particular, then, if a subordinator is finite at time t, the jumps of the subordinator up to t may be used as feature block frequencies if they have support in $[0, 1]$. Or, in general, the normalized jumps may be used as partition block frequencies. We can see from the righthand side of Figure 2.6 that the jumps of a subordinator partition intervals of the form $[0, t)$, as long as the subordinator has no drift component. In either the feature or cluster case, we have substituted the condition of independent and identical distribution for the partition or feature frequencies (i.e., the jumps) with a more natural continuous-time analogue: independent, stationary intervals.

Just as the Laplace transform of a positive random variable characterizes the distribution of that random variable, so does the Laplace transform of the subordinator—which is a positive random variable at any fixed time point—describe this stochastic process (Bertoin, 1998; Bertoin, 2004).

Theorem 2.5.3 (Lévy-Khinchin formula for subordinators)**.** *If $(T_s, s \geq 0)$ is a subordinator, then for $\lambda \geq 0$ we have*

$$\mathbb{E}(e^{-\lambda T_s}) = e^{-\Psi(\lambda)s} \tag{2.19}$$

with

$$\Psi(\lambda) = c\lambda + \int_0^\infty (1 - e^{-\lambda\xi})\Lambda(d\xi), \tag{2.20}$$

where $c \geq 0$ is called the drift constant and Λ is a non-negative, Lévy measure on $(0, \infty)$.

The function $\Psi(\lambda)$ is called the *Laplace exponent* in this context. We note that a subordinator is characterized by its drift constant and Lévy measure.

Using subordinators for feature allocation modeling is particularly easy; since the jumps of the subordinators are formed by a Poisson point process, we can use Poisson process methodology to find the stick lengths and EFPF. To set up this derivation, suppose we generate feature membership from a subordinator by taking Bernoulli draws at each of its jumps with success probability equal to the jump size. Since every jump has strictly positive size, the feature associated with each jump will eventually score a Bernoulli success for some index n with probability one. Therefore, we can enumerate all jumps of the process in order of appearance; that is, we first enumerate all features in which index 1 appears, then all features in which index 2 appears but not index 1, and so on. At the nth iteration, we enumerate all features in which index n appears but not previous indices. Let K_n^+ represent the number of indices so chosen on the nth round. Let $K_0 = 0$ so that recursively $K_n := K_{n-1} + K_n^+$ is the number of subordinator jumps seen by round n, inclusive. Let ξ_k for $k = K_{n-1} + 1, \ldots, K_n$ be the distribution of a particular subordinator jump seen on round n. We now turn to connecting the subordinator perspective to the earlier derivation of stick lengths in Section 2.4.

Example 2.5.4 (Indian buffet process)**.** In our earlier discussion, we found a collection of stick lengths to represent the featural frequencies for the IBP (Eq. (2.16) of Example 2.4.4

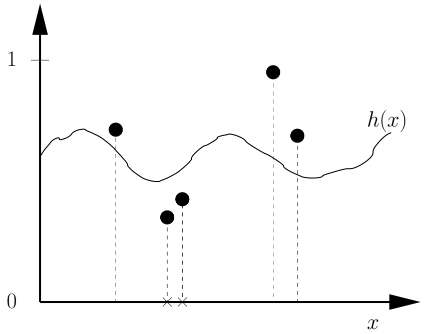


Figure 2.7: An illustration of Poisson thinning. The x-axis values of the filled black circles, emphasized by dotted lines, are generated according to a Poisson process. The $[0, 1]$-valued function $h(x)$ is arbitrary. The vertical axis values of the points are uniform draws in $[0, 1]$. The "thinned" points are the collection of x-axis values corresponding to vertical axis values below $h(x)$ and are denoted with a × symbol.

in Section 2.4). To see the connection to subordinators, we start from the *beta process subordinator* (Kim, 1999b) with zero drift ($c = 0$) and Lévy measure

$$\Lambda(d\xi) = \gamma\theta\xi^{-1}(1 - \xi)^{\theta-1}\, d\xi. \tag{2.21}$$

We will see that the mass parameter $\gamma > 0$ and concentration parameter $\theta > 0$ are the same as those introduced in Example 2.3.5 and continued in Example 2.4.4.

Theorem 2.5.5. *Generate a feature allocation from a beta process subordinator with Lévy measure given by Eq. (2.21). Then the sequence of subordinator jumps (ξ_k), indexed in order of appearance, has the same distribution as the sequence of IBP stick lengths (V_k) described by Eqs. (2.15) and (2.16).*

Proof. Recall the following fact about Poisson thinning (Kingman, 1993), illustrated in Figure 2.7. Suppose that a Poisson point process with rate measure λ generates points with values x. Then suppose that, for each such point x, we keep it with probability $h(x) \in [0, 1]$. The resulting set of points is also a Poisson point process, now with rate measure $\lambda'(A) = \int_A \lambda(dx)h(x)\, dx$.

We prove Theorem 2.5.5 recursively. Define the measure

$$\mu_n(d\xi) := \gamma\theta\xi^{-1}(1 - \xi)^{\theta+n-1}\, d\xi,$$

so that μ_0 is the beta process Lévy measure Λ in Eq. (2.21). We make the recursive assumption that μ_n is distributed as the beta process measure without atoms corresponding to features chosen on the first n iterations.

There are two parts to proving Theorem 2.5.5. First, we show that, on the nth iteration, the number of features chosen and the distribution of the corresponding atom weights agree with Eqs. (2.15) and (2.16), respectively. Second, we check that the recursion assumption holds.

For the first part, note that on the nth round we choose features with probability equal to their atom weight. So we form a thinned Poisson process with rate measure $\xi \cdot \mu_{n-1}(d\xi)$. This rate measure has total mass

$$\int_0^1 \xi \cdot \mu_{n-1}(d\xi) = \gamma \frac{\theta}{\theta + n - 1} =: \gamma_{n-1}.$$

So the number of features chosen is Poisson-distributed with mean $\gamma\theta(\theta + n - 1)^{-1}$, as desired (cf. Eq. (2.15)). And the atom weights have distribution equal to the normalized rate measure

$$\gamma_{n-1}^{-1} \xi \cdot \gamma\theta\xi^{-1}(1-\xi)^{\theta+(n-1)-1} \, d\xi = \mathrm{Beta}(\xi | 1, \theta + n - 1) d\xi,$$

as desired (cf. Eq. (2.16)).

Finally, to check the recursion assumption, we note that those sticks that remain were chosen for having Bernoulli failure draws; i.e., they were chosen with probability equal to one minus their atom weight. So the thinned rate measure for the next round is

$$(1-\xi) \cdot \gamma\theta\xi^{-1}(1-\xi)^{\theta+(n-1)-1} \, d\xi,$$

which is just μ_n. □

The form of the EFPF of the feature allocation generated from the beta process subordinator follows immediately from the stick length distributions we have just derived by the discussion in Example 2.4.4 in Section 2.4. ■

We see from the previous example that feature allocation stick lengths and EFPFs can be obtained in a straightforward manner using the Poisson process representation of the jumps of the subordinator. Partitions, however, are not as easy to analyze, principally due to the fact that the subordinator jumps must first be normalized to obtain a probability measure on $[0, 1]$; a random measure with finite total mass is not sufficient in the partition case. Hence we must compute the stick lengths and EPPF using partition block frequencies from these normalized jumps instead of directly from the subordinator jumps.

In the EPPF case, we make use a of a result that gives us the exchangeable probability function as a function of the Laplace exponent. Though we do not derive this formula here, its derivation can be found in Pitman (2003); the proof relies on, first, calculating the joint distribution of the subordinator jumps and partition generated from the normalized jumps and, second, integrating out the subordinator jumps to find the partition marginal.

Theorem 2.5.6. *Form a probability measure μ by normalizing jumps of the subordinator with Laplace exponent Ψ. Let (Π_n) be a consistent set of exchangeable partitions induced by* *iid draws from μ. For each exchangeable partition $\pi_N = \{A_1, \ldots, A_K\}$ of $[N]$ with $N_k := |A_k|$ for each k,*

$$\begin{aligned}\mathbb{P}(\Pi_N = \pi_N) &= p(N_1, \ldots, N_K) \\ &= \frac{(-1)^{N-K}}{(N-1)!} \int_0^\infty \lambda^{N-1} e^{-\Psi(\lambda)} \prod_{k=1}^{K} \Psi^{(N_k)}(\lambda)\, d\lambda, \end{aligned} \tag{2.22}$$

where $\Psi^{(N_k)}(\lambda)$ is the N_kth derivative of the Laplace exponent Ψ evaluated at λ.

Example 2.5.7 (Chinese restaurant process)**.** We start by introducing the *gamma process*, a subordinator that we will see below generates the Chinese restaurant process EPPF. The gamma process has Laplace exponent $\Psi(\lambda)$ (Eq. (2.19)) characterized by

$$c = 0, \quad \text{and} \quad \Lambda(d\xi) = \theta \xi^{-1} e^{-b\xi}\, d\xi \tag{2.23}$$

for $\theta > 0$ and $b > 0$ (cf. Eq. (2.20) in Theorem 2.5.3). We will see that θ corresponds to the CRP concentration parameter and that b is arbitrary and does not affect the partition model.

We calculate the EPPF using Theorem 2.5.6.

Theorem 2.5.8. *The EPPF for partition block membership chosen according to the normalized jumps (ρ_k) of the gamma subordinator with parameter θ is the CRP EPPF (Eq. (2.4)).*

Proof. By Theorem 2.5.6, if we can find all order derivatives of the Laplace exponent Ψ, we can calculate the EPPF for the partitions generated with frequencies equal to the normalized jumps of this subordinator. The derivatives of Ψ, which are known to always exist (Bertoin, 2000; Rogers and Williams, 2000), are straightforward to calculate if we begin by noting that, from Eq. (2.20) in Theorem 2.5.3, we have in general that

$$\Psi'(\lambda) = c + \int_0^\infty \xi e^{-\lambda\xi} \Lambda(d\xi).$$

Hence, for the gamma process subordinator,

$$\Psi'(\lambda) = \int_0^\infty e^{-\lambda\xi} \theta e^{-b\xi}\, d\xi = \frac{\theta}{\lambda + b}.$$

Then simple integration and differentiation yield

$$\begin{gathered}\Psi(\lambda) = \theta \log(\lambda + b) - \theta \log(b), \\ \text{since } \Psi(0) = 0, \text{ and} \\ \Psi^{(n)}(\lambda) = (-1)^{n-1} \frac{(n-1)!\theta}{(\lambda + b)^n}, \quad n \geq 1.\end{gathered}$$

We can substitute these quantities into the general EPPF formula in Eq. (2.22) of Theorem 2.5.6 to obtain

$$\begin{aligned}
&p(N_1, \ldots, N_K) \\
&= \frac{(-1)^{N-K}}{(N-1)!} \int_0^\infty \lambda^{N-1} (\lambda + b)^{-\theta} b^\theta \prod_{k=1}^K (-1)^{N_k - 1} \frac{(N_k - 1)! \theta}{(\lambda + b)^{N_k}} \, d\lambda \\
&= b^\theta \frac{\theta^K}{(N-1)!} \left[\prod_{k=1}^K (N_k - 1)! \right] b^{N-1-N-\theta+1} \int_0^\infty x^{N-1} (x+1)^{-N-\theta} \, dx \\
&\text{for } x = \lambda / b \\
&= \frac{\theta^K}{(N-1)!} \left[\prod_{k=1}^K (N_k - 1)! \right] \frac{\Gamma(N) \Gamma(\theta)}{\Gamma(N+\theta)} \\
&= \theta^K \left[\prod_{k=1}^K (N_k - 1)! \right] \frac{1}{\theta (\theta + 1)_{N-1 \uparrow 1}}.
\end{aligned}$$

The penultimate line follows from the form of the beta prime distribution. The final line is the CRP EPPF from Eq. (2.4), as desired. We note in particular that the parameter b does not appear in the final EPPF. □

■

Whenever the Laplace exponent of a subordinator is known, Theorem 2.5.6 can similarly be applied to quickly find the EPPF of the partition generated by sampling from the normalized subordinator jumps.

To find the distributions of the stick lengths, i.e., the partition block frequencies, from the subordinator representation for a partition, we must find the distributions of the normalized subordinator jumps.

As in the feature case, we may enumerate the jumps of a subordinator used for partitioning in the order of their appearance. That is, let ρ_1 be the normalized subordinator jump size corresponding to the cluster of the first data point. Recursively, suppose index n joins a cluster to which none of the indices in $[n-1]$ belong, and suppose there are k clusters among $[n-1]$. Then let ρ_{k+1} be the normalized subordinator jump size corresponding to the cluster containing n.

Example 2.5.9 (Chinese restaurant process)**.** We continue with the CRP example.

Theorem 2.5.10. *The normalized subordinator jumps* (ρ_k) *in order of appearance of the gamma subordinator with concentration parameter* θ *(and arbitrary parameter* $b > 0$*) have the same distribution as the CRP stick lengths (Eq. (2.14) of Example 2.4.3 in Section 2.4).*

Proof. First, we introduce some notation. Let $\tau = \sum_k \xi_k$, the sum over all of the jumps of the subordinator. Second, let $\tau_k = \tau - \sum_{j=1}^k \xi_k$, the total sum minus the first k elements (in order of appearance). Note that $\tau = \tau_0$. Finally, let $W_k = \tau_k / \tau_{k-1}$ and $V_k = 1 - W_k$. Then a simple telescoping of factors shows that $\rho_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$:

$$V_k \prod_{j=1}^{k-1}(1 - V_j) = \left(1 - \frac{\tau_k}{\tau_{k-1}}\right) \prod_{j=1}^{k-1} \frac{\tau_j}{\tau_{j-1}} = \frac{\tau_{k-1} - \tau_k}{\tau_0} = \frac{\xi_k}{\tau} = \rho_k.$$

It remains to show that the V_k have the desired distribution. To that end, it is easier to work with the W_k. We will find the following lemma (Pitman, 2006) useful.

Lemma 2.5.11. *Consider a subordinator with Lévy measure Λ, and suppose τ equals the sum of all jumps of the subordinator. Let ρ be the density of Λ with respect to Lebesgue measure. And let f be the density of the distribution of τ with respect to Lebesgue measure. Then*

$$\begin{aligned}
&\mathbb{P}(\tau_0 \in dt_0, \ldots, \tau_k \in dt_k) \\
&= f(t_k)\, dt_k \left(\prod_{j=0}^{k-1} \frac{(t_j - t_{j+1})\rho(t_j - t_{j+1})}{t_j}\, dt_j \right)
\end{aligned}$$

With this lemma in hand, the result follows from a change of variables calculation; we use a bijection between $\{W_1, \ldots, W_k, \tau\}$ and $\{\tau_0, \ldots, \tau_k\}$ defined by $\tau_k = \tau \prod_{j=1}^{k} W_j$. The determinant of the Jacobian for the transformation to the former variables from the latter is

$$J = \prod_{j=1}^{k} \left[\tau \prod_{i=1}^{j-1} W_i \right] = \prod_{j=0}^{k-1} \tau_j(\tau, W_1, \ldots, W_k)$$

In the derivation that follows, we start by expressing results in terms of the τ_j terms with the dependence on $\{\tau, W_1, \ldots, W_k\}$ suppressed to avoid notational clutter: e.g., $J = \prod_{j=0}^{k-1} \tau_j$. At the end, we will evaluate the τ_j terms as functions of $\{\tau, W_1, \ldots, W_k\}$.

For now, then, we have

$$\begin{aligned}
&\mathbb{P}(W_1 \in dw_1, \ldots, W_k \in dw_k, \tau \in dt_0) \\
&= \mathbb{P}(\tau_0 \in dt_0, \ldots, \tau_k \in dt_k) \cdot J \\
&= f(t_k)\, dt_k \left(\prod_{j=0}^{k-1} (t_j - t_{j+1})\rho(t_j - t_{j+1}) \right).
\end{aligned}$$

In the case of the gamma process, we can read $\rho(\xi) = \theta \xi^{-1} e^{-b\xi}$ from Eq. (2.23). The function f is determined by ρ and in this case (Pitman, 2006):

$$f(t) = \mathrm{Gamma}(t | \theta, b) = b^{\theta} \Gamma(\theta)^{-1} t^{\theta - 1} e^{-bt}.$$

So

$$\mathbb{P}(W_1 \in dw_1, \ldots, W_k \in dw_k, \tau \in dt_0)$$

$$\propto t_k^{\theta-1} e^{-bt_0} = t_0^{\theta-1} e^{-bt_0} \prod_{j=1}^{k} w_j^{\theta-1}.$$

Since the distribution factorizes, the $\{W_k\}$ are independent of each other and of τ. Second, we can read off the distributional kernel of each W_k to establish $W_k \overset{iid}{\sim} \text{Beta}(\theta, 1)$, from whence it follows that $V_k \overset{iid}{\sim} \text{Beta}(1, \theta)$. □

■

Inference

In some sense, we skipped ahead in describing inference in Sections 2.3 and 2.4. There, we made use of the fact that random labels for partitions and features imply exhangeability of the data partition block assignments (Z_n) and data feature assignments (Y_n). In the discussion above, we study the object that associates random uniformly distributed labels with each partition or feature. Assuming the labels come from a uniform distribution rather than a general continuous distribution is a special case of the discussion in Section 2.3, and we defer the general case to the next section (Section 2.6).

We have seen above that it is particularly straightforward to obtain an EPPF or EFPF formulation, which yields Gibbs sampling steps as described in Section 2.3, when the stick lengths are generated according to a normalized Poisson process in the partition case or a Poisson process in the feature case. Examples 2.5.4 and 2.5.7 illustrate how to find such exchangeable probability functions. Further, we have already seen the usefulness of the stick representation in inference, and Examples 2.5.4 and 2.5.9 illustrate how stick length distributions may be recovered from the subordinator framework.

2.6 Completely random measures

In our discussion of subordinators, the jump sizes of the subordinator corresponded to the feature frequencies or unnormalized partition frequencies and were the quantities of interest. By contrast, the locations of the jumps mainly served as convenient labels for the frequencies. These locations were chosen uniformly at random from the unit interval. This choice guaranteed the a.s. uniqueness of the labels and the exchangeability of the sequence of index assignments: (Z_n) in the clustering case or (Y_n) in the feature case.

However, a labeling retains exchangeability and a.s. uniqueness as long as the labels are chosen iid from any continuous distribution (not just the uniform distribution). Moreover, in typical applications, we wish to associate some parameter, often referred to as a “random effect,” with each partition block or feature. In the partition case, we usually model the nth data point X_n as being generated according to some likelihood depending on the parameter corresponding to its block assignment. E.g., an individual animal’s height and weight, X_n, varies randomly around the height and weight of its species, Z_n. Likewise, in the feature case, we typically model the observed data point X_n as being generated according to some likelihood depending on the collection of parameters corresponding to its collection of feature block assignments (cf. Eq. (2.11)). E.g., the book-buying pattern of an online consumer, X_n, varies with some noise based on the topics this person likes to read about: Y_n is a collection, possibly empty, of such topics.

In these cases, it can be useful to suppose that the partition block labels (or feature labels) ϕ_k are not necessarily $\mathbb{R}_+$-valued but rather are generated iid according to some continuous distribution H on a general space Φ. Then, whenever k is the order of appearance partition block label of index n, we let $Z_n = \phi_k$. Similarly, whenever k is the order-of-appearance feature label for some feature to which index n belongs, $\phi_k \in Y_n$. Finally, then, we complete the generative model in the partition case by letting $X_n \overset{indep}{\sim} \mathcal{L}(Z_n)$ for some distribution function $\mathcal{L}$ depending on parameter Z_n. And in the feature case, $X_n \overset{indep}{\sim} \mathcal{L}(Y_n)$, where now the distribution function $\mathcal{L}$ depends on the collection of parameters Y_n.

When we take the jump sizes (ξ_k) of a subordinator as the weights of atoms with locations (ϕ_k) drawn iid according to H as described above, we find ourselves with a *completely random measure* μ:

$$\mu = \sum_{k=1}^{\infty} \xi_k \delta_{\phi_k}. \tag{2.24}$$

A completely random measure is a random measure μ such that whenever A and A' are disjoint sets, we have that $\mu(A)$ and $\mu(A')$ are independent random variables.

To see that associating these more general atom locations to the jumps of a subordinator yields a completely random measure, note that Theorem 2.5.2 tells us that the subordinator jump sizes are generated according to a Poisson point process, with some intensity measure $\nu(d\xi)$. The Marking Theorem for Poisson point processes (Kingman, 1993) in turn yields that the tuples $\{(\xi_k, \phi_k)\}_k$ are generated according to a Poisson point process with intensity measure $\nu(d\xi)H(d\phi)$. By Kingman (1967), whenever the tuples $\{(\xi_k, \phi_k)\}_k$ are drawn according to a Poisson point process, the measure in Eq. (2.24) is completely random.

Example 2.6.1 (Dirichlet process)**.** We can form a completely random measure from the gamma process subordinator and a random labeling of the partition blocks. Specifically, suppose that the labels come from a continuous measure H. Then we generate a completely random measure G called a *gamma process* (Ferguson, 1973) in the following way:

$$\nu(d\xi \times d\phi) = \theta \xi^{-1} e^{-b\xi} d\xi \cdot H(d\phi) \tag{2.25}$$

$$\{(\xi_k, \phi_k)\}_k \sim \text{PPP}(\nu) \tag{2.26}$$

$$G = \sum_{k=1}^{\infty} \xi_k \delta_{\phi_k} \tag{2.27}$$

Here, $\text{PPP}(\nu)$ denotes a draw from a Poisson point process with intensity measure ν. The parameters $\theta > 0$ and $b > 0$ are the same as for the gamma process subordinator. A gamma

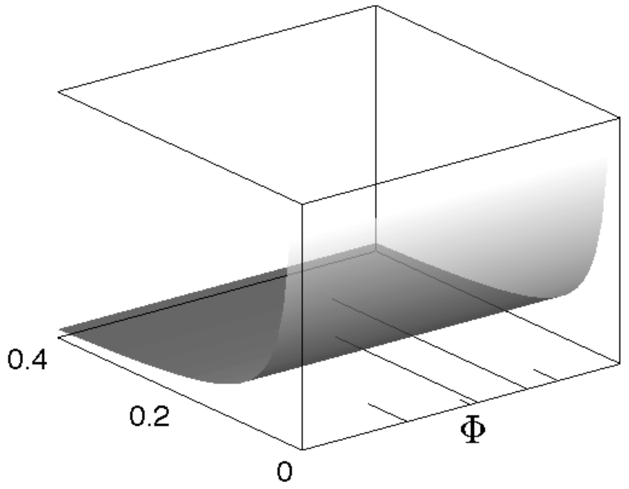


Figure 2.8: The gray manifold depicts the Poisson point process intensity measure ν in Eq. (2.25) for the choice $\Phi = [0, 1]$ and H the uniform distribution on $[0, 1]$. The endpoints of the line segments are points drawn from the Poisson point process as in Eq. (2.26). Taking the positive real-valued coordinate (leftmost axis) as the atom weights, we find the random measure G (a gamma process) on Φ from Eq. (2.27) in the bottom plane.

process draw, along with its generating Poisson point process intensity measure, is illustrated in Figure 2.8.

The *Dirichlet process* (DP) is the random measure formed by normalizing the gamma process (Ferguson, 1973). Since the Dirichlet process atom weights sum to one, it cannot be completely random. We can write the Dirichlet process D generated from the gamma process G above as:

$$\begin{aligned} \tau &= \sum_{k=1}^{\infty} \xi_k \\ \rho_k &= \xi_k / \tau \\ D &= \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k}. \end{aligned}$$

The random variables ρ_k have the same distribution as the Dirichlet process sticks (Eq. (2.14)) or normalized gamma process subordinator jump lengths, as we have seen above (Example 2.5.7). ■

Consider sampling points from a Dirichlet process and forming the induced partition of the data indices. Theorem 2.5.8 shows us that the distribution of the induced partition is the Chinese restaurant process EPPF.

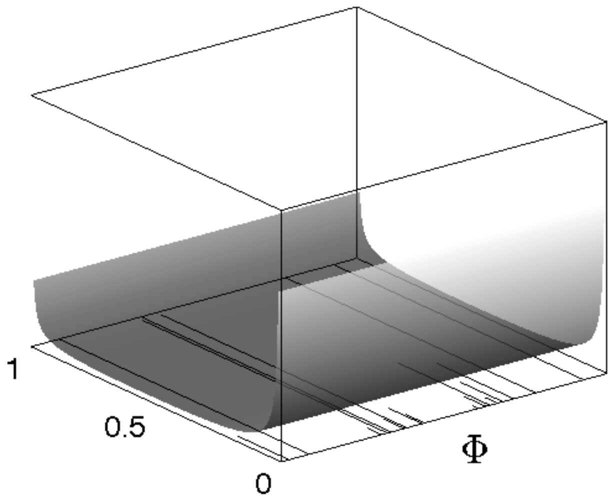


Figure 2.9: The gray manifold depicts the Poisson point process intensity measure ν in Eq. (2.28) for the choice $\Phi = [0, 1]$ and H the uniform distribution on $[0, 1]$. The endpoints of the line segments are points drawn from the Poisson point process as in Eq. (2.29). Taking the $[0, 1]$-valued coordinate (leftmost axis) as the atom weights, we find the measure B (a beta process) on Φ from Eq. (2.30) in the bottom plane.

Example 2.6.2 (Beta process)**.** We can form a completely random measure from the beta process subordinator and a random labeling of the feature blocks. If the labels are generated iid from a continuous measure H, then we say the completely random measure B, generated as follows, is called a *beta process*:

$$\nu(d\xi \times d\phi) = \gamma\theta\xi^{-1}(1-\xi)^{\theta-1} d\xi \cdot H(d\phi) \tag{2.28}$$

$$\{(\xi_k, \phi_k)\}_k \sim \mathrm{PPP}(\nu) \tag{2.29}$$

$$B = \sum_{k=1}^{\infty} \xi_k \delta_{\phi_k}. \tag{2.30}$$

The beta process, along with its generating intensity measure, is depicted in Figure 2.9. The (ξ_k) have the same distribution as the beta process sticks (Eq. (2.16)) or the beta process subordinator jump lengths (Example 2.5.4).

■

Now consider sampling a collection of atom locations according to Bernoulli draws from the atom weights of a beta process and forming the induced feature allocation of the data indices. Theorem 2.5.5 shows us that the distribution of the induced feature allocation is given by the Indian buffet process EFPF.

Inference

In this section, we finally study the full model first outlined in the context of inference of partition and feature structures in Section 2.3. The partition or feature labels described in this section are the same as the block-specific parameters first described in Section 2.3. Since this section focuses on a generalization of the partition or feature labeling scheme beyond the uniform distribution option encoded in subordinators, inference for the atom weights remains unchanged from Sections 2.3, 2.4, and 2.5.

However, we note that, in the course of inferring underlying partition or feature structures, we are often also interested in inferring the parameters of the generative model of the data given the partition block or the feature labels. Conditional on the partition or feature structure, such inference is handled as in a normal hierarchical model with fixed dependencies. Namely, the parameter within a particular block may be inferred from the data points that depend on this block as well as the prior distribution for the parameters. Details for the Dirichlet process example inferred via MCMC sampling are provided by S. N. MacEachern (1994); Escobar and West (1995); Neal (2000); Blei and Jordan (2006) work out details for the Dirichlet process using variational methods. In the beta process case, Griffiths and Ghahramani (2006); Teh, Görür, and Ghahramani (2007); Thibaux and Jordan (2007) describe MCMC sampling, and Paisley, Zaas, et al. (2010) describe a variational approach.

2.7 Discussion

We have pursued a progressive augmentation from (1) simple distributions over partitions and feature allocations in the form of exchangeable probability functions to (2) the representation of stick lengths encoding frequencies of the partition block and feature occurrences to (3) subordinators, which associate random $\mathbb{R}_+$-valued labels with each partition block or feature, and finally to (4) completely random measures, which associate a general class of labels with the stick lengths and whose labels we generally use as parameters in likelihood models built from the partition or feature allocation representation.

Along the way, we have focused primarily on two vignettes. We have shown, via these successive augmentations, that the Chinese restaurant process specifies the marginal distribution of the induced partition formed from iid draws from a Dirichlet process, which is in turn a normalized completely random measure. And we have shown that the Indian buffet process specifies the marginal distribution of the induced feature allocation formed by iid Bernoulli draws across the weights of a beta process.

There are many extensions of these ideas that lie beyond the scope of this chapter. A number of extensions of the CRP and Dirichlet process exist—in either the EPPF form (Pitman, 1996; Blei and Frazier, 2010), the stick length form (Dunson and Park, 2008), or the random measure form (Pitman and Yor, 1997). Likewise, extensions of the IBP and beta process have been explored (Teh, Görür, and Ghahramani, 2007; Paisley, Zaas, et al., 2010; Broderick, Jordan, and Pitman, 2012).

More generally, the framework above demonstrates how alternative partition and feature allocation models may be constructed—either by introducing different EPPFs (Pitman, 1996; Gnedin and Pitman, 2006) or EFPFs, different stick length distributions (Ishwaran and James, 2001), or different random measures (Wolpert and Ickstadt, 2004).

Finally, we note that expanding the set of combinatorial structures with useful Bayesian priors from partitions to the superset of feature allocations suggests that further such structures might be usefully examined. For instance, the *beta negative binomial process* (Broderick, Mackey, et al., 2014; Zhou et al., 2012) provides a prior on a generalization of a feature allocation where we allow the features themselves to be multisets; i.e., each index may have non-negative integer multiplicities of features. Models on trees (Adams, Ghahramani, and Jordan, 2010; McCullagh, Pitman, and Winkel, 2008; Blei, Griffiths, and Jordan, 2010), graphs (W. Li and McCallum, 2006), and permutations (Pitman, 1996) provide avenues for future exploration. And there likely remain further structures to be fitted out with useful Bayesian priors.

Chapter 3

Beta processes, stick-breaking, and power laws

The beta-Bernoulli process provides a Bayesian nonparametric prior for models involving collections of binary-valued features. A draw from the beta process yields an infinite collection of probabilities in the unit interval, and a draw from the Bernoulli process turns these into binary-valued features. Recent work has provided stick-breaking representations for the beta process analogous to the well-known stick-breaking representation for the Dirichlet process. We derive one such stick-breaking representation directly from the characterization of the beta process as a completely random measure. This approach motivates a three-parameter generalization of the beta process, and we study the power laws that can be obtained from this generalized beta process. We present a posterior inference algorithm for the beta-Bernoulli process that exploits the stick-breaking representation, and we present experimental results for a discrete factor-analysis model.

3.1 Introduction

Large data sets are often heterogeneous, arising as amalgams from underlying sub-populations. The analysis of large data sets thus often involves some form of stratification in which groupings are identified that are more homogeneous than the original data. While this can sometimes be done on the basis of explicit covariates, it is also commonly the case that the groupings are captured via discrete latent variables that are to be inferred as part of the analysis. Within a Bayesian framework, there are two widely employed modeling motifs for problems of this kind. The first is the *Dirichlet-multinomial motif*, which is based on the assumption that there are K "clusters" that are assumed to be mutually exclusive and exhaustive, such that allocations of data to clusters can be modeled via a multinomial random variable whose parameter vector is drawn from a Dirichlet distribution. A second motif is the *beta-Bernoulli motif*, where a collection of K binary "features" are used to describe the data, and where each feature is modeled as a Bernoulli random variable whose parameter is obtained from a beta distibution. The latter motif can be converted to the former in principle—we can view particular patterns of ones and zeros as defining a cluster, thus obtaining $M = 2^K$ clusters in total. But in practice models based on the Dirichlet-multinomial motif typically require $O(M)$ additional parameters in the likelihood, whereas those based on the beta-Bernoulli motif typically require only $O(K)$ additional parameters. Thus, if the combinatorial structure encoded by the binary features captures real structure in the data, then the beta-Bernoulli motif can make more efficient usage of its parameters.

The Dirichlet-multinomial motif can be extended to a stochastic process known as the *Dirichlet process*. A draw from a Dirichlet process is a random probability measure that can be represented as follows (McCloskey, 1965; Patil and Taillie, 1977; Ferguson, 1973; Sethuraman, 1994):

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\psi_i}, \tag{3.1}$$

where δ_{ψ_i} represents an atomic measure at location ψ_i, where both the $\{\pi_i\}$ and the $\{\psi_i\}$ are random, and where the $\{\pi_i\}$ are nonnegative and sum to one (with probability one). Conditioning on G and drawing N values independently from G yields a collection of M distinct values, where $M \leq N$ is random and grows (in expectation) at rate $O(\log N)$. Treating these distinct values as indices of clusters, we obtain a model in which the number of clusters is random and subject to posterior inference.

A great deal is known about the Dirichlet process—there are direct connections between properties of G as a random measure (e.g., it can be obtained from a Poisson point process), properties of the sequence of values $\{\pi_i\}$ (they can be obtained from a "stick-breaking process"), and properties of the collection of distinct values obtained by sampling from G (they are characterized by a stochastic process known as the *Chinese restaurant process*). These connections have helped to place the Dirichlet process at the center of Bayesian nonparametrics, driving the development of a wide variety of inference algorithms for models based on Dirichlet process priors and suggesting a range of generalizations (e.g. S. MacEachern, 1999; Ishwaran and James, 2001; Walker, 2007; Kalli, Griffin, and Walker, 2011).

It is also possible to extend the beta-Bernoulli motif to a Bayesian nonparametric framework, and there is a growing literature on this topic. The underlying stochastic process is the *beta process*, which is an instance of a family of random measures known as *completely random measures* (Kingman, 1967). The beta process was first studied in the context of survival analysis by Hjort (1990), where the focus is on modeling hazard functions via the random cumulative distribution function obtained by integrating the beta process. Thibaux and Jordan (2007) focused instead on the beta process realization itself, which can be represented as

$$G = \sum_{i=1}^{\infty} q_i \delta_{\psi_i},$$

where both the q_i and the ψ_i are random and where the q_i are contained in the interval $(0, 1)$. This random measure can be viewed as furnishing an infinite collection of coins, which, when tossed repeatedly, yield a binary featural description of a set of entities in which the number of features with non-zero values is random. Thus, the resulting *beta-Bernoulli process* can be viewed as an infinite-dimensional version of the beta-Bernoulli motif. Indeed, Thibaux and Jordan (2007) showed that by integrating out the random q_i and ψ_i one obtains—by analogy to the derivation of the Chinese restaurant process from the Dirichlet process—a combinatorial stochastic process known as the *Indian buffet process*, previously studied by Griffiths and Ghahramani (2006), who derived it via a limiting process involving random binary matrices obtained by sampling finite collections of beta-Bernoulli variables.

Stick-breaking representations of the Dirichlet process have been particularly important both for algorithmic development and for exploring generalizations of the Dirichlet process. These representations yield explicit recursive formulas for obtaining the weights $\{\pi_i\}$ in Eq. (3.1). In the case of the beta process, explicit non-recursive representations can be obtained for the weights $\{q_i\}$, based on size-biased sampling (Thibaux and Jordan, 2007) and inverse Lévy measure (Wolpert and Ickstadt, 2004; Teh, Görür, and Ghahramani, 2007). Recent work has also yielded recursive constructions that are more closely related to the stick-breaking representation of the Dirichlet process (Teh, Görür, and Ghahramani, 2007; Paisley, Zaas, et al., 2010).

Stick-breaking representations of the Dirichlet process permit ready generalizations to stochastic processes that yield power-law behavior (which the Dirichlet process does not), notably the Pitman-Yor process (Ishwaran and James, 2001; Pitman, 2006). Power-law generalizations of the beta process have also been studied (Teh and Görür, 2009) and stick-breaking-like representations derived. These latter representations are, however, based on the non-recursive sized-biased sampling and inverse-Lévy methods rather than the recursive representations of Teh, Görür, and Ghahramani (2007) and Paisley, Zaas, et al. (2010).

Teh, Görür, and Ghahramani (2007) and Paisley, Zaas, et al. (2010) derived their stick-breaking representations of the beta process as limiting processes, making use of the derivation of the Indian buffet process by Griffiths and Ghahramani (2006) as a limit of finite-dimensional random matrices. In the current chapter we show how to derive stick-breaking for the beta process directly from the underlying random measure. This approach not only has the advantage of conceptual clarity (our derivation is elementary), but it also permits a unified perspective on various generalizations of the beta process that yield power-law behavior.[1] We show in particular that it yields a power-law generalization of the stick-breaking representation of Paisley, Zaas, et al. (2010).

To illustrate our results in the context of a concrete application, we study a discrete factor analysis model previously considered by Griffiths and Ghahramani (2006) and Paisley, Zaas, et al. (2010). The model is of the form

$$X = Z\Phi + E, \tag{3.2}$$

where $X \in \mathbb{R}^{N \times P}$ is the data and $E \in \mathbb{R}^{N \times P}$ is an error matrix. The matrix $\Phi \in \mathbb{R}^{K \times P}$ is a matrix of factors, and $Z \in \mathbb{R}^{N \times K}$ is a binary matrix of factor loadings. The dimension K is

[1] A similar measure-theoretic derivation has been presented recently by Paisley, Blei, and Jordan (2012), who focus on applications to truncations of the beta process.

infinite, and thus the rows of Φ comprise an infinite collection of factors. The matrix Z is obtained via a draw from a beta-Bernoulli process; its nth row is an infinite binary vector of features (i.e., factor loadings) encoding which of the infinite collection of factors are used in modeling the nth data point.

The remainder of the chapter is organized as follows. We introduce the beta process, and its conjugate measure the Bernoulli process, in Section 3.2. In order to consider stick-breaking and power law behavior in the beta-Bernoulli framework, we first review stick-breaking for the Dirichlet process in Section 3.3 and power laws in clustering models in Section 3.4. We consider potential power laws that might exist in featural models in Section 3.4. Our main theoretical results come in the following two sections. First, in Section 3.5, we provide a proof that the stick-breaking representation of Paisley, Zaas, et al. (2010), expanded to include a third parameter, holds for a three-parameter extension of the beta process. Our proof takes a measure-theoretic approach based on a Poisson process. We then make use of the Poisson process framework to establish asymptotic power laws, with exact constants, for the three-parameter beta process in Section 3.6. We also show, in Section 3.6, that there are aspects of the beta-Bernoulli framework that cannot exhibit a power law. We illustrate the asymptotic power laws on a simulated data set in Section 3.7. We present experimental results in Section 8.5, and we present an MCMC algorithm for posterior inference in Appendix 3.A.

3.2 The beta process and the Bernoulli process

The beta process and the Bernoulli process are instances of the general family of random measures known as *completely random measures* (Kingman, 1967). A completely random measure H on a probability space $(\Psi, \mathcal{S})$ is a random measure such that, for any disjoint measurable sets $A_1, \ldots, A_n \in \mathcal{S}$, the random variables $H(A_1), \ldots, H(A_n)$ are independent.

Completely random measures can be obtained from an underlying Poisson point process. Let $\nu(d\psi, du)$ denote a σ-finite measure[2] on the product space $\Psi \times \mathbb{R}$. Draw a realization from a Poisson point process with rate measure $\nu(d\psi, du)$. This yields a set of points $\Pi = \{(\psi_i, U_i)\}_i$, where the index i may range over a countable infinity. Finally, construct a random measure as follows:

$$B = \sum_{i=1}^{\infty} U_i \delta_{\psi_i}, \tag{3.3}$$

where δ_{ψ_i} denotes an atom at ψ_i. This discrete random measure is such that for any measurable set $T \in \mathcal{S}$,

$$B(T) = \sum_{i:\psi_i \in T} U_i.$$

That B is completely random follows from the Poisson point process construction.

[2] The measure ν need not necessarily be σ-finite to generate a completely random measure though we consider only σ-finite measures in this work.

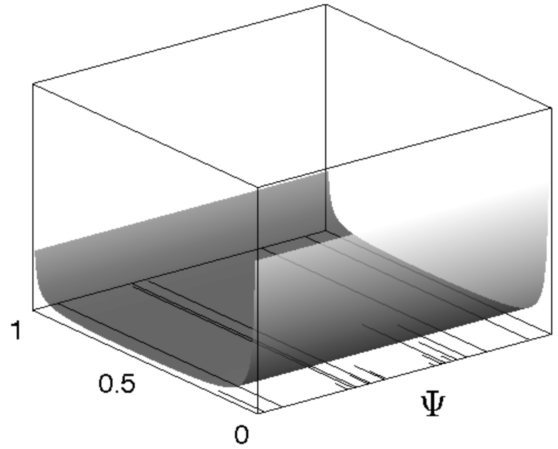


Figure 3.1: The gray surface illustrates the rate density in Eq. (3.4) corresponding to the beta process. The base measure B_0 is taken to be uniform on Ψ. The non-zero endpoints of the line segments plotted below the surface are a particular realization of the Poisson process, and the line segments themselves represent a realization of the beta process.

In addition to the representation obtained from a Poisson process, completely random measures may include a deterministic measure and a set of atoms at fixed locations. The component of the completely random measure generated from a Poisson point process as described above is called the *ordinary component*. As shown by Kingman (1967), completely random measures are essentially characterized by this representation. An example is shown in Figure 3.1.

The *beta process*, denoted $B \sim \mathrm{BP}(\theta, B_0)$, is an example of a completely random measure. As long as the *base measure* B_0 is continuous, which is our assumption here, B has only an ordinary component with rate measure

$$\nu_{\mathrm{BP}}(d\psi, du) = \theta(\psi)u^{-1}(1-u)^{\theta(\psi)-1}\, du\, B_0(d\psi), \quad \psi \in \Psi, u \in [0,1], \tag{3.4}$$

where θ is a positive function on Ψ. The function θ is called the *concentration function* (Hjort, 1990). In the remainder we follow Thibaux and Jordan (2007) in taking θ to be a real-valued constant and refer to it as the *concentration parameter*. We assume B_0 is nonnegative and fixed. The total mass of B_0, $\gamma := B_0(\Psi)$, is called the *mass parameter*. We assume γ is strictly positive and finite. The density in Eq. (3.4), with the choice of B_0 uniform over $[0,1]$, is illustrated in Figure 3.1.

The beta process can be viewed as providing an infinite collection of coin-tossing probabilities. Tossing these coins corresponds to a draw from the *Bernoulli process*, yielding an infinite binary vector that we will treat as a latent feature vector.

More formally, a *Bernoulli process* $Y \sim BeP(B)$ is a completely random measure with potentially both fixed atomic and ordinary components. In defining the Bernoulli process

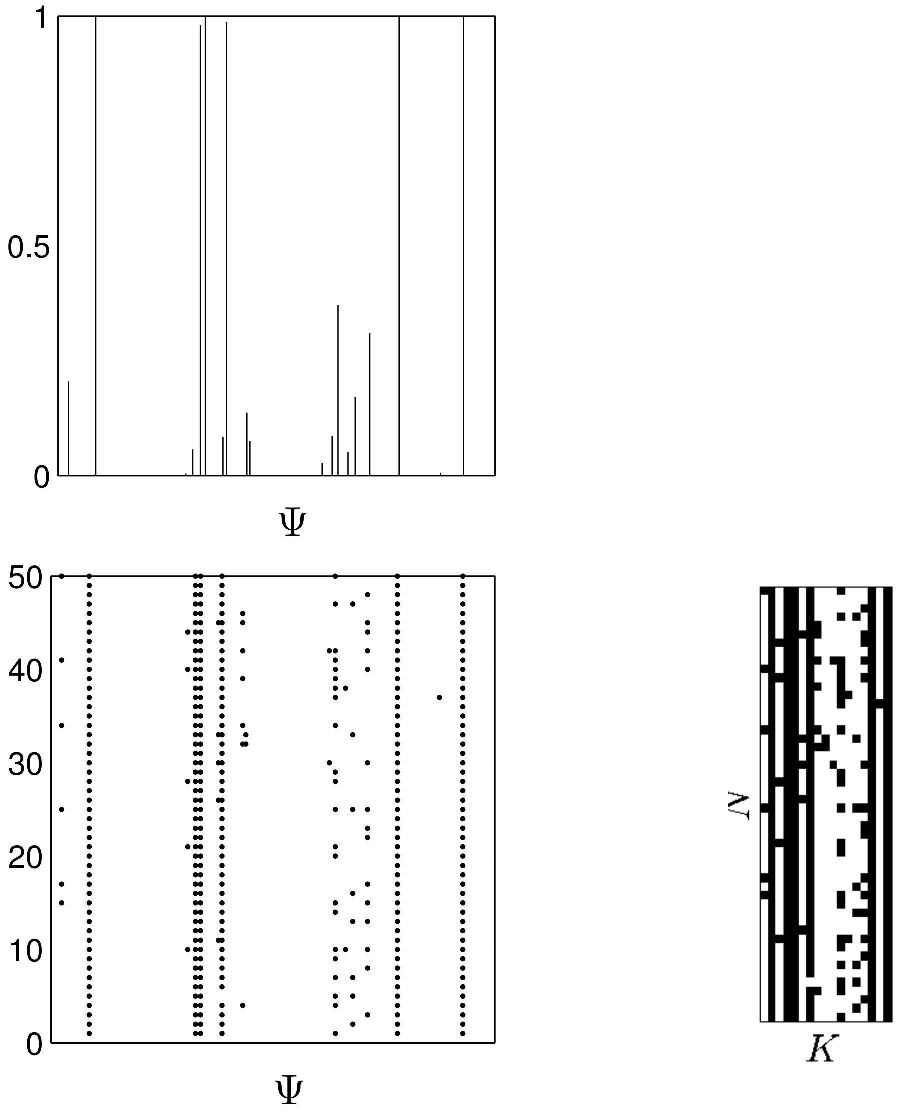


Figure 3.2: *Upper left*: A draw B from the beta process. *Lower left*: 50 draws from the Bernoulli process $BeP(B)$. The vertical axis indexes the draw number among the 50 exchangeable draws. A point indicates a one at the corresponding location on the horizontal axis, $\psi \in \Psi$. *Right*: We can form a matrix from the lower left plot by including only those ψ values with a non-zero number of Bernoulli successes among the 50 draws from the Bernoulli process. Then, the number of columns K is the number of such ψ, and the number of rows N is the number of draws made. A black square indicates a one at the corresponding matrix position; a white square indicates a zero.

we consider only the case in which B is discrete, i.e., of the form in Eq. (3.3), though not necessarily a beta process draw or even random for the moment. Then Y has only a fixed atomic component and has the form

$$Y = \sum_{i=1}^{\infty} b_i \delta_{\psi_i}, \tag{3.5}$$

where $b_i \sim \text{Bern}(u_i)$ for u_i the corresponding atomic mass in the measure B. We can see that $\mathbb{E}(Y|B) = B(\Psi)$ from the mean of the Bernoulli distribution, so the number of non-zero points in any realization of the Bernoulli process is finite when B is a finite measure.

We can link the beta process and N Bernoulli process draws to generate a random feature matrix Z. To that end, first draw $B \sim \text{BP}(\theta, B_0)$ for fixed hyperparameters θ and B_0 and then draw $Y_n \overset{iid}{\sim} \text{BeP}(B)$ for $n \in \{1, \ldots, N\}$. Note that since B is discrete, each Y_n will be discrete as in Eq. (3.5), with point masses only at the atoms $\{\psi_i\}$ of the beta process B. Note also that $\mathbb{E}B(\Psi) = \gamma < \infty$, so B is a finite measure, and it follows that the number of non-zero point masses in any draw Y_n from the Bernoulli process will be finite. Therefore, the total number of non-zero point masses K across N such Bernoulli process draws is finite.

Now reorder the $\{\psi_i\}$ so that the first K are exactly those locations where some Bernoulli process in $\{Y_n\}_{n=1}^N$ has a non-zero point mass. We can form a matrix $Z \in \{0,1\}^{N \times K}$ as a function of the $\{Y_n\}_{n=1}^N$ by letting the (n,k) entry equal one when Y_n has a non-zero point mass at ψ_k and zero otherwise. If we wish to think of Z as having an infinite number of columns, the remaining columns represent the point masses of the $\{Y_n\}_{n=1}^N$ at $\{\psi_k\}_{k>K}$, which we know to be zero by construction. We refer to the overall procedure of drawing Z according to, first, a beta process and then repeated Bernoulli process draws in this way as a *beta-Bernoulli process*, and we write $Z \sim \text{BP} - \text{BeP}(N, \gamma, \theta)$. Note that we have implicitly integrated out the $\{\psi_k\}$, and the distribution of the matrix Z depends on B_0 only through its total mass, γ. As shown by Thibaux and Jordan (2007), this process yields the same distribution on row-exchangeable, infinite-column matrices as the Indian buffet process (Griffiths and Ghahramani, 2006), which describes a stochastic process directly on (equivalence classes of) binary matrices. That is, the Indian buffet process is obtained as an exchangeable distribution on binary matrices when the underlying beta process measure is integrated out. This result is analogous to the derivation of the Chinese restaurant process as the exchangeable distribution on partitions obtained when the underlying Dirichlet process is integrated out. The beta-Bernoulli process is illustrated in Figure 3.2.

3.3 Stick-breaking for the Dirichlet process

The stick-breaking representation of the Dirichlet process (McCloskey, 1965; Patil and Taillie, 1977; Sethuraman, 1994) provides a simple recursive procedure for obtaining the weights $\{\pi_i\}$ in Eq. (3.1). This procedure provides an explicit representation of a draw G from the Dirichlet process, one which can be usefully instantiated and updated in posterior inference

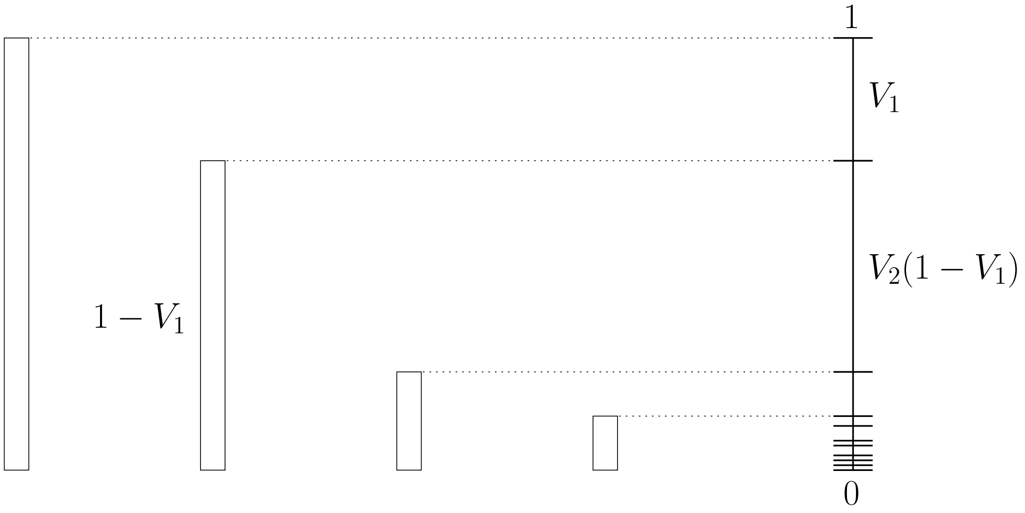


Figure 3.3: A stick-breaking process starts with the unit interval (*far left*). First, a random fraction V_1 of the unit interval is broken off; the remaining stick has length $1 - V_1$ (*middle left*). Next, a random fraction V_2 of the remaining stick is broken off, i.e., a fragment of size $V_2(1-V_1)$; the remaining stick has length $(1-V_1)(1-V_2)$. This process proceeds recursively and generates stick fragments $V_1, V_2(1-V_1), \ldots, V_i \prod_{j<i}(1-V_j), \ldots$. These fragments form a random partition of the unit interval (*far right*).

algorithms (Ishwaran and James, 2001; Blei and Jordan, 2006). We begin this section by reviewing this stick-breaking construction as well as some of the extensions to this construction that yield power-law behavior. We then turn to a consideration of stick-breaking and power laws in the setting of the beta process.

Stick-breaking is the process of recursively breaking off random fractions of the unit interval. In particular, let $V_1, V_2, \ldots$ be some countable sequence of random variables, each with range $[0, 1]$. Each V_i represents the fraction of the remaining stick to break off at step i. Thus, the first stick length generated by the stick-breaking process is V_1. At this point, a fragment of length $1 - V_1$ of the original stick remains. Breaking off V_2 fraction of the remaining stick yields a second stick fragment of $V_2(1 - V_1)$. This process iterates such that the stick length broken off at step i is $V_i \prod_{j<i}(1 - V_j)$. The stick-breaking recursion is illustrated in Figure 3.3.

The Dirichlet process arises from the special case in which the V_i are independent draws from the Beta$(1, \theta)$ distribution (McCloskey, 1965; Patil and Taillie, 1977; Sethuraman, 1994). Thus we have the following representation of a draw $G \sim \mathrm{DP}(\theta, G_0)$:

$$\begin{aligned} G &= \sum_{i=1}^{\infty} \left[V_i \prod_{j=1}^{i-1} (1 - V_j) \right] \delta_{\psi_i} \\ V_i &\overset{iid}{\sim} \mathrm{Beta}(1, \theta) \end{aligned}$$

$$\psi_i \overset{iid}{\sim} G_0, \tag{3.6}$$

where G_0 is referred to as the *base measure* and θ is referred to as the *concentration parameter*.

3.4 Power law behavior

Consider the process of sampling a random measure G from a Dirichlet process and subsequently drawing independently N times from G. The number of unique atoms sampled according to this process will grow as a function of N. The growth associated with the Dirichlet process is relatively slow, however, and when the Dirichlet process is used as a prior in a clustering model one does not obtain the heavy-tailed behavior commonly referred to as a "power law." In this section we first provide a brief exposition of the different kinds of power law that we might wish to obtain in a clustering model and discuss how these laws can be obtained via an extension of the stick-breaking representation. We then discuss analogous laws for featural models.

Power laws in clustering models

First, we establish some notation. Given a number N of draws from a discrete random probability measure G (where G is not necessarily a draw from the Dirichlet process), let $(N_1, N_2, \ldots)$ denote the sequence of counts associated with the unique values obtained among the N draws, where we view these unique values as "clusters." Let

$$K_{N,j} = \sum_{i=1}^{\infty} \mathbb{1}(N_i = j), \tag{3.7}$$

and let

$$K_N = \sum_{i=1}^{\infty} \mathbb{1}(N_i > 0). \tag{3.8}$$

That is, $K_{N,j}$ is the number of clusters that are drawn exactly j times, and K_N is the total number of clusters.

There are two types of power-law behavior that a clustering model might exhibit. First, there is the type of power law behavior reminiscent of Heaps' law (Heaps, 1978; Gnedin, Hansen, and Pitman, 2007) and describing the asymptotic behavior of the number of clusters:

$$K_N \overset{a.s.}{\sim} cN^a, \quad N \to \infty \tag{3.9}$$

for some constants $c > 0, a \in (0, 1)$. Here, $\sim$ means that the limit of the ratio of the left-hand and right-hand side, when they are both real-valued and non-random, is one as the number of data points N grows large. We denote a power law in the form of Eq. (3.9) as *Type I*. Second, there is the type of power law behavior reminiscent of Zipf's law (Zipf, 1949;

Gnedin, Hansen, and Pitman, 2007) and describing the asymptotic behavior of the number of clusters of size j:

$$K_{N,j} \overset{a.s.}{\sim} \frac{a\Gamma(j-a)}{j!\Gamma(1-a)} cN^a, \quad N \to \infty \tag{3.10}$$

again for some constants $c > 0, a \in (0,1)$. We refer to the power law in Eq. (3.10) as *Type II.* Note that Gnedin, Hansen, and Pitman (2007) have shown, and we will see further below, that this particular way of writing the proportionality constant is natural.

Sometimes in the case of Eq. (3.10), we are interested in the behavior in j; therefore we recall $j! = \Gamma(j+1)$ and note the following fact about the Γ-function ratio in Eq. (3.10) (cf. Tricomi and Erdélyi, 1951):

$$\frac{\Gamma(j-a)}{\Gamma(j+1)} \sim j^{-1-a}, \quad j \to \infty. \tag{3.11}$$

Again, we see behavior in the form of a power law at work.

Power-law behavior of Types I and II (and equivalent formulations; see Gnedin, Hansen, and Pitman, 2007) has been observed in a variety of real-world clustering problems including, but not limited to: the number of species per plant genus, the in-degree or out-degree of a graph constructed from hyperlinks on the Internet, the number of people in cities, the number of words in documents, the number of papers published by scientists, and the amount each person earns in income (Mitzenmacher, 2004; Goldwater, Griffiths, and M. Johnson, 2006). Bayesians modeling these situations will prefer a prior that reflects this distributional attribute.

While the Dirichlet process exhibits neither type of power-law behavior, the *Pitman-Yor process* yields both kinds of power law (Pitman and Yor, 1997; Goldwater, Griffiths, and M. Johnson, 2006) though we note that in this case c is a random variable (still with no dependence on N or j). The Pitman-Yor process, denoted $G \sim \mathrm{PY}(\theta, \alpha, G_0)$, is defined via the following stick-breaking representation:

$$\begin{aligned} G &= \sum_{i=1}^{\infty} \left[V_i \prod_{j=1}^{i-1} (1 - V_j) \right] \delta_{\psi_i} \\ V_i &\overset{indep}{\sim} \mathrm{Beta}(1-\alpha, \theta + i\alpha) \\ \psi_i &\overset{iid}{\sim} G_0, \end{aligned} \tag{3.12}$$

where α is known as a *discount parameter.* The case $\alpha = 0$ returns the Dirichlet process (cf. Eq. (3.6)).

Note that in both the Dirichlet process and the Pitman-Yor process, the weights $\{V_i \prod_{j=1}^{i-1}(1 - V_j)\}$ are the weights of the process in size-biased order (Pitman, 2006). In the Pitman-Yor case, the $\{V_i\}$ are no longer identically distributed.

Power laws in featural models

The beta-Bernoulli process provides a specific kind of feature-based representation of entities. In this section we study general featural models and consider the power laws that might arise for such models.

In the clustering framework, we considered N draws from a process that put exactly one mass of size one on some value in Ψ and mass zero elsewhere. In the featural framework we consider N draws from a process that places some non-negative integer number of masses, each of size one, on an almost surely finite set of values in Ψ and mass zero elsewhere. As N_i was the sum of masses at a point labeled $\psi_i \in \Psi$ in the clustering framework, so do we now let N_i be the sum of masses at a point labeled $\psi_i \in \Psi$. We use the same notation as in Section 3.4 to define the number of features K_N (Eq. (3.8)) and the number of features represented by j data points $K_{N,j}$ (Eq. (3.7)). But now we note that the counts N_i no longer sum to N in general.

In the case of featural models, we can still talk about Type I and II power laws, both of which have the same interpretation as in the case of clustering models: asymptotic power law behavior of the number of features and asymptotic power law behavior in the number of features of cardinality j, both as $N \to \infty$.

In the featural case, however, it is also possible to consider a third type of power law. If we let k_n denote the number of features present in the nth draw, we say that k_n shows power law behavior if

$$\mathbb{P}(k_n > M) \sim cM^{-a}$$

for positive constants c and a. We call this last type of power law *Type III*.

3.5 Stick-breaking for the beta process

The weights $\{q_i\}$ for the beta process can be derived by a variety of procedures, including size-biased sampling (Thibaux and Jordan, 2007) and inverse Lévy measure (Wolpert and Ickstadt, 2004; Teh, Görür, and Ghahramani, 2007). The procedures that are closest in spirit to the stick-breaking representation for the Dirichlet process are those due to Paisley, Zaas, et al. (2010) and Teh, Görür, and Ghahramani (2007). Our point of departure is the former, which has the following form:

$$\begin{aligned} B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}} \\ C_i &\overset{iid}{\sim} \text{Poisson}(\gamma) \\ V_{i,j}^{(l)} &\overset{iid}{\sim} \text{Beta}(1, \theta) \\ \psi_{i,j} &\overset{iid}{\sim} \frac{1}{\gamma} B_0. \end{aligned} \tag{3.13}$$

This representation is analogous to the stick-breaking representation of the Dirichlet process in that it represents a draw from the beta process as a sum over independently drawn atoms, with the weights obtained by a recursive procedure. However, it is worth noting that for every (i, j) tuple subscript for $V_{i,j}^{(l)}$, a different stick exists and is broken across the superscript l. Thus, there are no special additive properties across weights in the sum in Eq. (3.13); by contrast, the weights in Eq. (3.12) sum to one almost surely.

The generalization of the one-parameter Dirichlet process to the two-parameter Pitman-Yor process suggests that we might consider generalizing the stick-breaking representation of the beta process in Eq. (3.13) as follows:

$$\begin{aligned} B &= \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}} \\ C_i &\overset{iid}{\sim} \text{Poisson}(\gamma) \\ V_{i,j}^{(l)} &\overset{indep}{\sim} \text{Beta}(1-\alpha, \theta + i\alpha) \\ \psi_{i,j} &\overset{iid}{\sim} \frac{1}{\gamma} B_0. \end{aligned} \tag{3.14}$$

In Section 3.6 we will show that introducing the additional parameter α indeed yields Type I and II power law behavior (but not Type III).

In the remainder of this section we present a proof that these stick-breaking representations arise from the beta process. In contradistinction to the proof of Eq. (3.13) by Paisley, Zaas, et al. (2010), which used a limiting process defined on sequences of finite binary matrices, our approach makes a direct connection to the Poisson process characterization of the beta process. Our proof has several virtues: (1) it relies on no asymptotic arguments and instead comes entirely from the Poisson process representation; (2) it is, as a result, simpler and shorter; and (3) it demonstrates clearly the ease of incorporating a third parameter analogous to the discount parameter of the Pitman-Yor process and thereby provides a strong motivation for the extended stick-breaking representation in Eq. (3.14).

Aiming toward the general stick-breaking representation in Eq. (3.14), we begin by defining a three-parameter generalization of the beta process.[3] We say that $B \sim \text{BP}(\theta, \alpha, B_0)$, where we call α a *discount parameter*, if, for $\psi \in \Psi, u \in [0, 1]$, we have

$$\nu_{\text{BP}}(d\psi, du) = \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} u^{-1-\alpha}(1-u)^{\theta+\alpha-1}\, du\, B_0(d\psi). \tag{3.15}$$

It is straightforward to show that this three-parameter density has similar properties to that of the two-parameter beta process. For instance, choosing $\alpha \in (0, 1)$ and $\theta > -\alpha$ is necessary for the beta process to have finite total mass almost surely; in this case,

$$\int_{\Psi \times \mathbb{R}_+} u\, \nu_{\text{BP}}(d\psi, du) = \gamma < \infty. \tag{3.16}$$

[3] See also Teh and Görür (2009) or Kim and Lee (2001), with $\theta(t) \equiv 1 - \alpha, \beta(t) \equiv \theta + \alpha$, where the left-hand sides are in the notation of Kim and Lee (2001).

We now turn to the main result of this section.

Proposition 3.5.1. *B can be represented according to the process described in Eq. (3.14) if and only if $B \sim \mathrm{BP}(\theta, \alpha, B_0)$.*

Proof. First note that the points in the set

$$P_1 := \left\{(\psi_{1,1}, V_{1,1}^{(1)}), (\psi_{1,2}, V_{1,2}^{(1)}), \ldots, (\psi_{1,C_1}, V_{1,C_1}^{(1)})\right\}$$

are by construction independent and identically distributed conditioned on C_1. Since C_1 is Poisson-distributed, P_1 is a Poisson point process. The same logic gives that in general, for

$$P_i := \left\{\left(\psi_{i,1}, V_{i,1}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i,1}^{(l)})\right), \ldots, \left(\psi_{i,C_i}, V_{i,C_i}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i,C_i}^{(l)})\right)\right\},$$

P_i is a Poisson point process.

Next, define

$$P := \bigcup_{i=1}^{\infty} P_i.$$

As the countable union of Poisson processes with finite rate measures, P is itself a Poisson point process.

Notice that we can write B in Eq. (3.14) as the completely random measure $B = \sum_{(\psi,U)\in P} U\delta_\psi$. Also, for any $B' \sim \mathrm{BP}(\theta, \alpha, B_0)$, we can write $B' = \sum_{(\psi',U')\in\Pi} U'\delta_{\psi'}$, where Π is a Poisson point process with rate measure $\nu_{\mathrm{BP}} = B_0 \times \mu_{\mathrm{BP}}$, and μ_{BP} is a σ-finite measure with density

$$\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} u^{-1-\alpha}(1-u)^{\theta+\alpha-1}\, du. \tag{3.17}$$

Therefore, to show that B has the same distribution as B', it is enough to show that P and Π have the same rate measures.

To that end, let ν denote the rate measure of P. Let $\#S$ indicate the number of elements in set S, and let $\mathbb{1}E$ denote the indicator of the event E; $\mathbb{1}E$ is equal to one when E is true and equal to zero when E is false. Then we have

$$\begin{aligned}
\nu(A \times \tilde{A}) &= \mathbb{E}\#\{(\psi_i, U_i) \in A \times \tilde{A})\} \\
&= \frac{1}{\gamma} B_0(A) \cdot \mathbb{E} \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} \mathbb{1}\{V_{ij}^{(i)} \prod_{l=1}^{i-1}(1 - V_{ij}^{(l)}) \in \tilde{A}\} \\
&= \frac{1}{\gamma} B_0(A) \cdot \sum_{i=1}^{\infty} \mathbb{E} \sum_{j=1}^{C_i} \mathbb{1}\{V_{ij}^{(i)} \prod_{l=1}^{i-1}(1 - V_{ij}^{(l)}) \in \tilde{A}\},
\end{aligned} \tag{3.18}$$

where the last line follows by monotone convergence. Each term in the outer sum can be further decomposed as

$$\begin{aligned}
\mathbb{E}\sum_{j=1}^{C_i} \mathbb{1}\{V_{ij}^{(i)} \prod_{l=1}^{i-1}(1 - V_{ij}^{(l)}) \in \tilde{A}\} &= \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{C_i} \mathbb{1}\{V_{ij}^{(i)} \prod_{l=1}^{i-1}(1 - V_{ij}^{(l)}) \in \tilde{A}\} \middle| C_i\right]\right] \\
&= \mathbb{E}\left[C_i\right] \mathbb{E}\left[\mathbb{1}\{V_{i1}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i1}^{(l)}) \in \tilde{A}\}\right] \\
&\quad \text{since the } V_{ij}^{(l)} \text{ are iid across } j \text{ and independent of } C_i \\
&= \gamma\ \mathbb{E}\mathbb{1}\{V_i \prod_{l=1}^{i-1}(1 - V_l) \in \tilde{A}\} \qquad (3.19) \\
&\quad \text{for } V_i \overset{indep}{\sim} \text{Beta}(1 - \alpha, \theta + i\alpha),
\end{aligned}$$

where the last equality follows since the choice of $\{V_i\}$ gives

$$V_i \prod_{l=1}^{i-1}(1 - V_l) \overset{d}{=} V_{i1}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i1}^{(l)}).$$

Substituting Eq. (3.19) back into Eq. (3.18), canceling γ factors, and applying monotone convergence again yields

$$\nu(A \times \tilde{A}) = B_0(A) \cdot \mathbb{E}\sum_{i=1}^{\infty} \mathbb{1}\{V_i \prod_{l=1}^{i-1}(1 - V_l) \in \tilde{A}\}.$$

We note that both of the measures ν and ν_{BP} factorize:

$$\begin{aligned}
\nu(A \times \tilde{A}) &= B_0(A) \cdot \mathbb{E}\sum_{i=1}^{\infty} \mathbb{1}\{V_i \prod_{l=1}^{i-1}(1 - V_l) \in \tilde{A}\} \\
\nu_{BP}(A \times \tilde{A}) &= B_0(A)\mu_{\text{BP}}(\tilde{A}),
\end{aligned}$$

so it is enough to show that $\mu = \mu_{\text{BP}}$ for the measure μ defined by

$$\mu(\tilde{A}) := \mathbb{E}\sum_{i=1}^{\infty} \mathbb{1}\{V_i \prod_{l=1}^{i-1}(1 - V_l) \in \tilde{A}\}. \qquad (3.20)$$

At this point and later in proving Proposition 3.6.1, we will make use of part of Campbell's theorem, which we copy here from Kingman (1993) for completeness.

Theorem 3.5.2 (Part of Campbell's Theorem). *Let Π be a Poisson process on S with rate measure μ, and let $f : S \to \mathbb{R}$ be measurable. If $\int_S \min(|f(x)|, 1)\ \mu(dx) < \infty$, then*

$$\mathbb{E}\left[\sum_{X \in \Pi} f(X)\right] = \int_S f(x)\ \mu(dx). \qquad (3.21)$$

Now let $\tilde{U}$ be a size-biased pick from $\{V_i \prod_{l=1}^{i-1}(1 - V_l)\}_{i=1}^{\infty}$. By construction, for any bounded, measurable function g, we have

$$\mathbb{E}\left[g(\tilde{U}) | \{V_i\}\right] = \sum_{i=1}^{\infty} V_i \prod_{l=1}^{i-1}(1 - V_l) \cdot g(V_i \prod_{l=1}^{i-1}(1 - V_l)).$$

Taking expectations yields

$$\mathbb{E}g(\tilde{U}) = \mathbb{E}\left[\sum_{i=1}^{\infty} V_i \prod_{l=1}^{i-1}(1 - V_l) g(V_i \prod_{l=1}^{i-1}(1 - V_l))\right] = \int u g(u) \mu(du),$$

where the final equality follows by Campbell's theorem with the choice $f(u) = ug(u)$. Since this result holds for all bounded, measurable g, we have that

$$\mathbb{P}(\tilde{U} \in du) = u\mu(du). \tag{3.22}$$

Finally, we note that, by Eq. (3.20), $\tilde{U}$ is a size-biased sample from probabilities generated by stick-breaking with proportions $\{\mathrm{Beta}(1 - \alpha, \theta + i\alpha)\}$. Such a sample is then distributed $\mathrm{Beta}(1 - \alpha, \theta + \alpha)$ since, as mentioned above, the Pitman-Yor stick-breaking construction gives the size-biased frequencies in order. So, rearranging Eq. (3.22), we can write

$$\begin{aligned}
\mu(du) &= u^{-1}\mathbb{P}(\tilde{U} \in du) \\
&= u^{-1} \frac{\Gamma(1 + \theta)}{\Gamma(1 - \alpha)\Gamma(\theta + \alpha)} u^{(1-\alpha)-1}(1 - u)^{(\theta+\alpha)-1} \\
&\quad \text{using the Beta}(1 - \alpha, \theta + \alpha) \text{ density} \\
&= \mu_{\mathrm{BP}}(du),
\end{aligned}$$

as was to be shown. □

3.6 Power law derivations

By linking the three-parameter stick-breaking representation to the power-law beta process in Eq. (3.15), we can use the results of the following section to conclude that the feature assignments in the three-parameter model follow both Type I and Type II power laws and that they do not follow a Type III power law (Section 3.4). We note that Teh and Görür (2009) found big-O behavior for Types I and II in the three-parameter beta process and Poisson tail behavior in the Type III case. We can strengthen these results to obtain exact asymptotic behavior with constants in the first two cases and also conclude that Type III power laws can never hold in the featural framework whenever the sum of the feature probabilities is almost surely finite, an assumption that would appear to be a necessary component of any physically realistic model.

Type I and II power laws

Our subsequent derivation expands upon the work of Gnedin, Hansen, and Pitman (2007). In that paper, the main thrust of the argument applies to the case in which the feature probabilities are fixed rather than random. In what follows, we obtain power laws of Type I and II in the case in which the feature probabilities are random, in particular when the probabilities are generated from a Poisson process. We will see that this last assumption becomes convenient in the course of the proof. Finally, we apply our results to the specific example of the beta-Bernoulli process.

Recall that we defined K_N, the number of represented clusters in the first N data points, and $K_{N,j}$, the number of clusters represented j times in the first N data points, in Eqs. (3.8) and (3.7), respectively. In Section 3.4, we noted that the same definitions in Eqs. (3.8) and (3.7) hold for featural models if we now let N_i denote the number of data points at time N in which feature i is represented. In terms of the Bernoulli process, N_i would be the number of Bernoulli process draws, out of N, where the ith atom has unit (i.e., nonzero) weight. Thus, K_N is now the number of represented features in the first N data points, and $K_{N,j}$ is the number of features represented j times. It need not be the case that the N_i sum to N here.

Working directly to find power laws in K_N and $K_{N,j}$ as N increases is challenging in part due to N being an integer. A useful technique to surmount this difficulty is called *Poissonization*. In Poissonizing K_N and $K_{N,j}$, we consider new functions $K(t)$ and $K_j(t)$ where the argument t is continuous, in contrast to the integer argument N. We will define $K(t)$ and $K_j(t)$ such that $K(N)$ and $K_j(N)$ have the same asymptotic behavior as K_N and $K_{N,j}$, respectively.

In particular, our derivation of the asymptotic behavior of K_N and $K_{N,j}$ will consist of three parts and will involve working extensively with the mean feature counts

$$\Phi_N := \mathbb{E}[K_N] \quad \text{and} \quad \Phi_{N,j} := \mathbb{E}[K_{N,j}] \quad (j > 1)$$

with $N \in \{1, 2, \ldots\}$ and the Poissonized mean feature counts

$$\Phi(t) := \mathbb{E}[K(t)] \quad \text{and} \quad \Phi_j(t) := \mathbb{E}[K_j(t)] \quad (j > 1)$$

with $t > 0$. First, we will take advantage of Poissonization to find power laws in $\Phi(t)$ and $\Phi_j(t)$ as $t \to \infty$ (Proposition 3.6.1). Then, in order to relate these results back to the original process, we will show that Φ_N and $\Phi(N)$ have the same asymptotic behavior and also that $\Phi_{N,j}$ and $\Phi_j(N)$ have the same asymptotic behavior as $N \to \infty$ (Lemma 3.6.3). Finally, to obtain results for the random process values K_N and $K_{N,j}$, we will conclude by showing that K_N almost surely has the same asymptotic behavior as Φ_N and that $\sum_{k<j} K_{N,k}$ almost surely has the same asymptotic behavior as $\sum_{k<j} \Phi_{N,k}$ as $N \to \infty$ (Proposition 3.6.4).

To obtain power laws for the Poissonized process, we must begin by defining $K(t)$ and $K_j(t)$. To do so, we will construct Poisson processes on the positive half-line, one for each feature. $K(t)$ will be the number of such Poisson processes with points in the interval $[0, t]$;

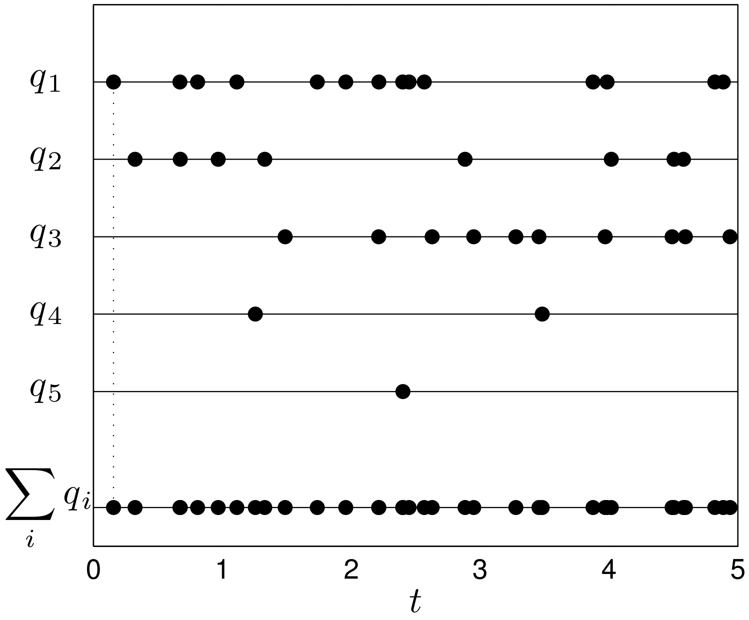


Figure 3.4: The first five sets of points, starting from the top of the figure, illustrate Poisson processes on the positive half-line in the range $t \in [0, 5]$ with respective rates $q_1, \ldots, q_5$. The bottom set of points illustrates the union of all points from the preceding Poisson point processes and is, therefore, itself a Poisson process with rate $\sum_i q_i$. In this example, we have for instance that $K(1) = 2$, $K(4) = 5$, and $K_2(4) = 1$.

similarly, $K_j(t)$ will be the number of Poisson processes with j points in the interval $[0, t]$. This construction is illustrated in Figure 3.4. It remains to specify the rates of these Poisson processes.

Let $(q_1, q_2, \ldots)$ be a countably infinite vector of feature probabilities. We begin by putting minimal restrictions on the q_i. We assume that they are strictly positive, decreasing real numbers. They need not necessarily sum to one, and they may be random. Indeed, we will eventually consider the case where the q_i are the (random) atom weights of a beta process, and then we will have $\sum_i q_i \neq 1$ with probability one.

Let Π_i be a standard Poisson process on the positive real line generated with rate q_i (see, e.g., the top five lines in Figure 3.4). Then $\Pi := \bigcup_i \Pi_i$ is a standard Poisson process on the positive real line with rate $\sum_i q_i$ (see, e.g., the lowermost line in Figure 3.4), where we henceforth assume $\sum_i q_i < \infty$ a.s.

Finally, as mentioned above, we define $K(t)$ to be the number of Poisson processes Π_i with any points in $[0, t]$:

$$K(t) := \sum_i \mathbb{1}\{|\Pi_i \cap [0, t]| > 0\}.$$

And we define $K_j(t)$ to be the number of Poisson processes Π_i with exactly j points in $[0, t]$:

$$K_j(t) := \sum_i \mathbb{1}\{|\Pi_i \cap [0, t]| = j\}.$$

These definitions are very similar to the definitions of K_N and $K_{N,j}$ in Eqs. (3.8) and (3.7), respectively. The principal difference is that the K_N are incremented only at integer N whereas the $K(t)$ can have jumps at any $t \in \mathbb{R}_+$. The same observation holds for the $K_{N,j}$ and $K_j(t)$.

In addition to Poissonizing K_N and $K_{N,j}$ to define $K(t)$ and $K_j(t)$, we will also find it convenient to assume that the $\{q_i\}$ themselves are derived from a Poisson process with rate measure ν. We note that Poissonizing from a discrete index N to a continuous time index t is an approximation and separate from our assumption that the $\{q_i\}$ are generated from a Poisson process though both are fundamentally tied to the ease of working with Poisson processes.

We are now able to write out the mean feature counts in both the Poissonized and original cases. First, the Poissonized definitions of Φ and K allow us to write

$$\Phi(t) := \mathbb{E}[K(t)] = \mathbb{E}[\mathbb{E}[K(t)|q]] = \mathbb{E}[\mathbb{E}[\sum_i \mathbb{1}\{|\Pi_i \cap [0, t]| > 0\}|q]].$$

With a similar approach for $\Phi_j(t)$, we find

$$\Phi(t) = \mathbb{E}[\sum_i (1 - e^{-tq_i})], \quad \Phi_j(t) = \mathbb{E}[\sum_i \frac{(tq_i)^j}{j!} e^{-tq_i}].$$

With the assumption that the $\{q_i\}$ are drawn from a Poisson process with measure ν, we can apply Campbell's theorem (Theorem 3.5.2) to both the original and Poissonized versions of the process to derive the final equality in each of the following lines

$$\Phi(t) = \mathbb{E}[\sum_i (1 - e^{-tq_i})] = \int_0^1 (1 - e^{-tx}) \; \nu(dx) \tag{3.23}$$

$$\Phi_N = \mathbb{E}[\sum_i (1 - (1 - q_i)^N)] = \int_0^1 (1 - (1 - x)^N) \; \nu(dx) \tag{3.24}$$

$$\Phi_j(t) = \mathbb{E}[\sum_i \frac{(tq_i)^j}{j!} e^{-tq_i}] = \frac{t^j}{j!} \int_0^1 x^j e^{-tx} \; \nu(dx) \tag{3.25}$$

$$\Phi_{N,j} = \binom{N}{j} \mathbb{E}[\sum_i q_i^j (1 - q_i)^{N-j}] = \binom{N}{j} \int_0^1 x^j (1 - x)^{N-j} \; \nu(dx). \tag{3.26}$$

Now we establish our first result, which gives a power law in $\Phi(t)$ and $\Phi_j(t)$ when the Poisson process rate measure ν has corresponding power law properties.

Proposition 3.6.1. *Asymptotic behavior of the integral of ν of the following form*

$$\nu_1[0,x] := \int_0^x u\,\nu(du) \sim \frac{\alpha}{1-\alpha} x^{1-\alpha} l(1/x), \quad x \to 0 \tag{3.27}$$

where l is a regularly varying function and $\alpha \in (0,1)$ implies

$$\begin{aligned} \Phi(t) &\sim \Gamma(1-\alpha) t^\alpha l(t), \quad t \to \infty \\ \Phi_j(t) &\sim \frac{\alpha \Gamma(j-\alpha)}{j!} t^\alpha l(t), \quad t \to \infty \quad (j > 1). \end{aligned}$$

Proof. The key to this result is in the repeated use of Abelian or Tauberian theorems. Let A be a map $A : F \to G$ from one function space to another: e.g., an integral or a Laplace transform. For $f \in F$, an Abelian theorem gives us the asymptotic behavior of $A(f)$ from the asymptotic behavior of f, and a Tauberian theorem gives us the asymptotic behavior of f from that of $A(f)$.

First, integrating by parts yields

$$\nu_1[0,x] = -x\bar{\nu}(x) + \int_0^x \bar{\nu}(u)\,du, \quad \bar{\nu}(x) := \int_x^\infty \nu(u)\,du,$$

so the stated asymptotic behavior in ν_1 yields $\bar{\nu}(x) \sim l(1/x)x^{-\alpha} (x \to 0)$ by a Tauberian theorem (Feller, 1966; Gnedin, Hansen, and Pitman, 2007) where the map A is an integral.

Second, another integration by parts yields

$$\Phi(t) = t \int_0^\infty e^{-tx} \bar{\nu}(x)\,dx.$$

The desired asymptotic behavior in Φ follows from the asymptotic behavior in $\bar{\nu}$ and an Abelian theorem (Feller, 1966; Gnedin, Hansen, and Pitman, 2007) where the map A is a Laplace transform. The result for $\Phi_j(t)$ follows from a similar argument when we note that repeated integration by parts of Eq. (3.25) also yields a Laplace transform. □

The importance of assuming that the q_i are distributed according to a Poisson process is that this assumption allowed us to write Φ as an integral and thereby make use of classic Abelian and Tauberian theorems. The importance of Poissonizing the processes K_j and $K_{N,j}$ is that we can write their means as in Eqs. (3.23) and (3.25), which are—up to integration by parts—in the form of Laplace transforms.

Proposition 3.6.1 is the most significant link in the chain of results needed to show asymptotic behavior of the feature counts K_N and $K_{N,j}$ in that it relates power laws in the known feature probability rate measure ν to power laws in the mean behavior of the Poissonized version of these processes. It remains to show this mean behavior translates back to K_N and $K_{N,j}$, first by relating the means of the original and Poissonized processes and then by relating the means to the almost sure behavior of the counts. The next two lemmas address the former concern. Together they establish that the mean feature counts Φ_N and $\Phi_{N,j}$ have the same asymptotic behavior as the corresponding Poissonized mean feature counts $\Phi(N)$ and $\Phi_j(N)$.

Lemma 3.6.2. *Let ν be σ-finite with $\int_0^\infty \nu(du) = \infty$ and $\int_0^\infty u \; \nu(du) < \infty$. Then the number of represented features has unbounded growth almost surely. The expected number of represented features has unbounded growth, and the expected number of features has sublinear growth. That is,*

$$K(t) \uparrow \infty \text{ a.s.}, \quad \Phi(t) \uparrow \infty, \quad \Phi(t) \ll t.$$

Proof. As in Gnedin, Hansen, and Pitman (2007), the first statement follows from the fact that q is countably infinite and each q_i is strictly positive. The second statement follows from monotone convergence. The final statement is a consequence of $\sum_i q_i < \infty$ a.s. □

Lemma 3.6.3. *Suppose the $\{q_i\}$ are generated according to a Poisson process with rate measure as in Lemma 3.6.2. Then, for $N \to \infty$,*

$$|\Phi_N - \Phi(N)| < \frac{2}{N}\Phi_2(N) \to 0$$
$$|\Phi_{N,j} - \Phi_j(N)| < \frac{c_j}{N}\max\{\Phi_j(N), \Phi_{j+2}(N)\} \to 0.$$

for some constants c_j.

Proof. The proof is the same as that of Lemma 1 of Gnedin, Hansen, and Pitman (2007). Establishing the inequalities results from algebraic manipulations. The convergence to zero is a consequence of Lemma 3.6.2. □

Finally, before considering the specific case of the three-parameter beta process, we wish to show that power laws in the means Φ_N and $\Phi_{N,j}$ extend to almost sure power laws in the number of represented features.

Proposition 3.6.4. *Suppose the $\{q_i\}$ are generated from a Poisson process with rate measure as in Lemma 3.6.2. Suppose that $\Phi(t) \sim Ct^\alpha l(t)$ and $\Phi_j(t) \sim C' t^\alpha l'(t)$ for $\alpha \in (0,1)$, $C, C' > 0$, and l and l' slowly varying as $t \to \infty$. Then, for $N \to \infty$,*

$$K_N \overset{a.s.}{\sim} \Phi_N, \quad \sum_{k<j} K_{N,k} \overset{a.s.}{\sim} \sum_{k<j} \Phi_{N,k}.$$

Proof. We wish to show that $K_N / \Phi_N \overset{a.s.}{\to} 1$ as $N \to \infty$. By Borel-Cantelli, it is enough to show that, for any $\epsilon > 0$,

$$\sum_N \mathbb{P}\left(\left|\frac{K_N}{\Phi_N} - 1\right| > \epsilon\right) < \infty.$$

To that end, note

$$\mathbb{P}\left(|K_N - \Phi_N| > \epsilon\Phi_N\right) \leq \mathbb{P}\left(\Phi_N > \epsilon\Phi_N + K_N\right) + \mathbb{P}\left(K_N > \epsilon\Phi_N + \Phi_N\right).$$

The note after Theorem 4 in D. Freedman (1973) gives that

$$\mathbb{P}\left(\Phi_N > \epsilon\Phi_N + K_N\right) \quad \leq \quad \exp\left(-\epsilon^2\Phi_N\right)$$

$$\mathbb{P}\left(K_N > \epsilon\Phi_N + \Phi_N\right) \;\leq\; \exp\left(-\frac{\epsilon^2}{1+\epsilon}\Phi_N\right).$$

So

$$\begin{aligned}\mathbb{P}\left(\left|\frac{K_N}{\Phi_N} - 1\right| > \epsilon\right) &\leq 2\exp\left(-\frac{1}{2}\epsilon^2\Phi_N\right)\\ &\leq c\exp\left(-\frac{1}{2}\epsilon^2 N^\alpha l(N)\right)\end{aligned}$$

for some constant c and sufficiently large N by Lemma 3.6.3 and the assumption on $\Phi(t)$. The last expression is summable in N, and Borel-Cantelli holds.

The proof that $\sum_{k<j} K_{N,k} \overset{a.s.}{\sim} \sum_{k<j} \Phi_{N,j}$ follows the same argument. □

It remains to show that we obtain Type I and II power laws in our special case of the three-parameter beta process, which implies a particular rate measure ν in the Poisson process representation of the $\{q_i\}$. For the three-parameter beta process density in Eq. (3.15), we have

$$\begin{aligned}\nu_1[0,x] &= \int_{\Psi\times(0,x]} u\;\nu_{BP}(d\psi, du)\\ &= \gamma\cdot\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^x u^{-\alpha}(1-u)^{\theta+\alpha-1}\,du\\ &\sim \gamma\cdot\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\int_0^x u^{-\alpha}\,du, \quad x\downarrow 0\\ &= \gamma\cdot\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\cdot\frac{1}{1-\alpha}x^{1-\alpha}.\end{aligned}$$

The final line is exactly the form required by Eq. (3.27) in Proposition 3.6.1, with $l(y)$ equal to the constant function of value

$$C := \frac{\gamma}{\alpha}\cdot\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}. \tag{3.28}$$

Then Proposition 3.6.1 implies that the following power laws hold for the mean of the Poissonized process:

$$\begin{aligned}\Phi(t) &\overset{a.s.}{\sim} \Gamma(1-\alpha)Ct^\alpha, \quad t\to\infty\\ \Phi_j(t) &\overset{a.s.}{\sim} \frac{\alpha\Gamma(j-\alpha)}{j!}Ct^\alpha, \quad t\to\infty \quad (j>1).\end{aligned}$$

Lemma 3.6.3 further yields

$$\Phi_N \overset{a.s.}{\sim} \Gamma(1-\alpha)CN^\alpha, \quad N\to\infty$$

$$\Phi_{N,j} \overset{a.s.}{\sim} \frac{\alpha\Gamma(j-\alpha)}{j!} C N^{\alpha}, \quad N \to \infty \quad (j > 1),$$

and finally Proposition 3.6.4 implies

$$K_N \overset{a.s.}{\sim} \Gamma(1-\alpha) C N^{\alpha}, \quad N \to \infty \tag{3.29}$$

$$K_{N,j} \overset{a.s.}{\sim} \frac{\alpha\Gamma(j-\alpha)}{j!} C N^{\alpha}, \quad N \to \infty \quad (j > 1). \tag{3.30}$$

These are exactly the desired Type I and II power laws (Eqs. (3.9) and (3.10)) for appropriate choices of the constants.

Exponential decay in the number of features

Next we consider a single data point and the number of features that are expressed for that data point in the featural model. We prove results for the general case where the ith feature has probability $q_i \geq 0$ such that $\sum_i q_i < \infty$. Let Z_i be a Bernoulli random variable with success probability q_i and such that all the Z_i are independent. Then $\mathbb{E}[\sum_i Z_i] = \sum_i q_i =: Q$. In this case, a Chernoff bound (Chernoff, 1952; Hagerup and Rub, 1990) tells us that, for any $\delta > 0$, we have

$$\mathbb{P}[\sum_i Z_i \geq (1+\delta)Q] \leq e^{\delta Q}(1+\delta)^{-(1+\delta)Q}.$$

When M is large enough such that $M > Q$, we can choose δ such that $(1+\delta)Q = M$. Then this inequality becomes

$$\mathbb{P}[\sum_i Z_i \geq M] \leq e^{M-Q} Q^M M^{-M} \quad \text{for } M > Q. \tag{3.31}$$

We see from Eq. (3.31) that the number of features $\sum_i Z_i$ that are expressed for a data point exhibits super-exponential tail decay and therefore cannot have a power law probability distribution when the sum of feature probabilities $\sum_i q_i$ is finite. For comparison, let $Z \sim \text{Poisson}(Q)$. Then (Franceschetti et al., 2007)

$$\mathbb{P}[Z \geq M] \leq e^{M-Q} Q^M M^{-M} \quad \text{for } M > Q,$$

the same tail bound as in Eq. (3.31).

To apply the tail-behavior result of Eq. (3.31) to the beta process (with two or three parameters), we note that the total feature probability mass is a.s. finite by Eq. (3.16). Since the same set of feature probabilities is used in all subsequent Bernoulli process draws for the beta-Bernoulli process, the result holds.

3.7 Simulation

To illustrate the three types of power laws discussed above, we simulated beta process atom weights under three different choices of the discount parameter α, namely $\alpha = 0$ (the classic, two-parameter beta process), $\alpha = 0.3$, and $\alpha = 0.6$. In all three simulations, the remaining beta process parameters were kept constant at total mass parameter value $\gamma = 3$ and concentration parameter value $\theta = 1$.

The simulations were carried out using our extension of the Paisley, Zaas, et al. (2010) stick-breaking construction in Eq. (3.14). We generated 2,000 rounds of feature probabilities; that is, we generated 2,000 random variables C_i and $\sum_{i=1}^{2{,}000} C_i$ feature probabilities. With these probabilities, we generated $N = 1{,}000$ data points, i.e., 1,000 vectors of $(\sum_{i=1}^{2{,}000} C_i)$ independent Bernoulli random variables with these probabilities. With these simulated data, we were able to perform an empirical evaluation of our theoretical results.

Figure 3.5 illustrates power laws in the number of represented features K_N on the left (Type I power law) and the number of features represented by exactly one data point $K_{N,1}$ on the right (Type II power law). Both of these quantities are plotted as functions of the increasing number of data points N. The blue points show the simulated values for the classic, two-parameter beta process case with $\alpha = 0$. The center set of black points in each case corresponds to $\alpha = 0.3$, and the upper set of black points in each case corresponds to $\alpha = 0.6$.

We also plot curves obtained from our theoretical results in order to compare them to the simulation. Recall that in our theoretical development, we noted that there are two steps to establishing the asymptotic behavior of K_N and $K_{N,j}$ as N increases. First, we compare the random quantities K_N and $K_{N,j}$ to their respective means, Φ_N and $\Phi_{N,j}$. These means, as computed via numerical quadrature from Eq. (3.24) and directly from Eq. (3.26), are shown by red curves in the plots. Second, we compare the means to their own asymptotic behavior. This asymptotic behavior, which we ultimately proved was shared with the respective K_N or $K_{N,j}$ in Eqs. (3.29) and (3.30), is shown by green curves in the plots.

We can see in both plots that the $\alpha = 0$ behavior is distinctly different from the straight-line behavior of the $\alpha > 0$ examples. In both cases, we can see that any growth in α is slower than can be described by straight-line growth. In particular, when $\alpha = 0$, the expected number of features is

$$\Phi_N = \mathbb{E}[K_N] = \mathbb{E}\left[\sum_{n=1}^{N} \text{Poisson}\left(\gamma \frac{\theta}{n+\theta}\right)\right] = \sum_{n=1}^{N} \gamma \frac{\theta}{n+\theta} \sim \gamma\theta \log(N). \tag{3.32}$$

Similarly, when $\alpha = 0$, the expected number of features represented by exactly one data point, $K_{N,1}$, is (by Eq. (3.26))

$$\begin{aligned}\Phi_{N,1} = \mathbb{E}[K_{N,1}] &= \binom{N}{1} \int_0^1 x^1 (1-x)^{N-1} \cdot \gamma\theta x^{-1} (1-x)^{\theta-1}\, dx \\ &= N\gamma\theta \cdot \frac{\Gamma(1)\Gamma(N-1+\theta)}{\Gamma(N+\theta)} = \gamma\theta \frac{N}{N-1+\theta} \sim \gamma\theta,\end{aligned}$$

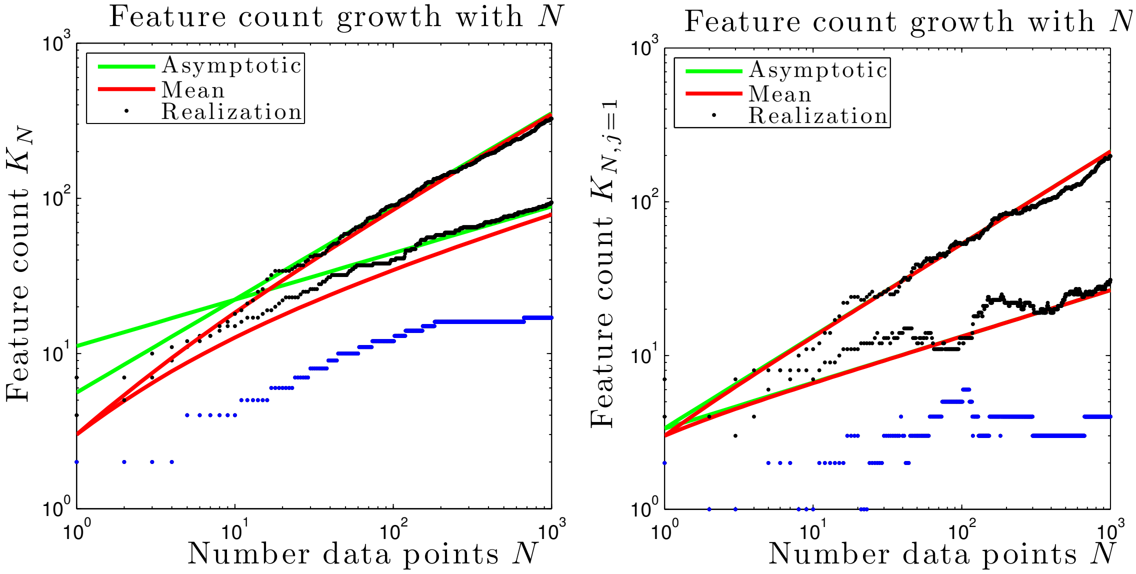


Figure 3.5: Growth in the number of represented features K_N (*left*) and the number of features represented by exactly one data point $K_{N,1}$ (*right*) as the total number of data points N grows. The points in the scatterplot are derived by simulation; blue for $\alpha = 0$, center black for $\alpha = 0.3$, and upper black for $\alpha = 0.6$. The red lines in the *left* plot show the theoretical mean Φ_N (Eq. (3.24)); in the *right* plot, they show the theoretical mean $\Phi_{N,1}$ (Eq. (3.26)). The green lines show the theoretical asymptotic behavior, Eq. (3.29) on the *left* (Type I power law) and Eq. (3.30) on the *right* (Type II power law).

where the second line follows from using the normalization constant of the (proper) beta distribution. Interestingly, while $K_{N,1}$ grows as a power law when $\alpha > 0$, its expectation is constant when $\alpha = 0$. While many new features are instantiated as N increases in the $\alpha = 0$ case, it seems that they are quickly represented by more data points than just the first one.

Type I and II power laws are somewhat easy to visualize since we have one point in our plots for each data point simulated. The behaviors of $K_{N,j}$ as a function of j for fixed N and Type III power laws (or lack thereof) are somewhat more difficult to visualize. In the case of $K_{N,j}$ as a function of j, we might expect that a large number of data points N is necessary to see many groups of size j for j much greater than one. In the Type III case, we have seen that in fact power laws do not hold for any value of α in the beta process. Rather, the number of data points exhibiting more than M features decreases more quickly in M than a power law would predict; therefore, we cannot plot many values of M before this number effectively goes to zero.

Nonetheless, Figure 3.6 compares our simulated data to the approximation of Eq. (3.10) with Eq. (3.11) (*left*) and Type III power laws (*right*). On the left, blue points as usual denote simulated data under $\alpha = 0$; middle black points show $\alpha = 0.3$, and upper black points show $\alpha = 0.6$. Here, we use connecting lines between plotted points to clarify α values. The green

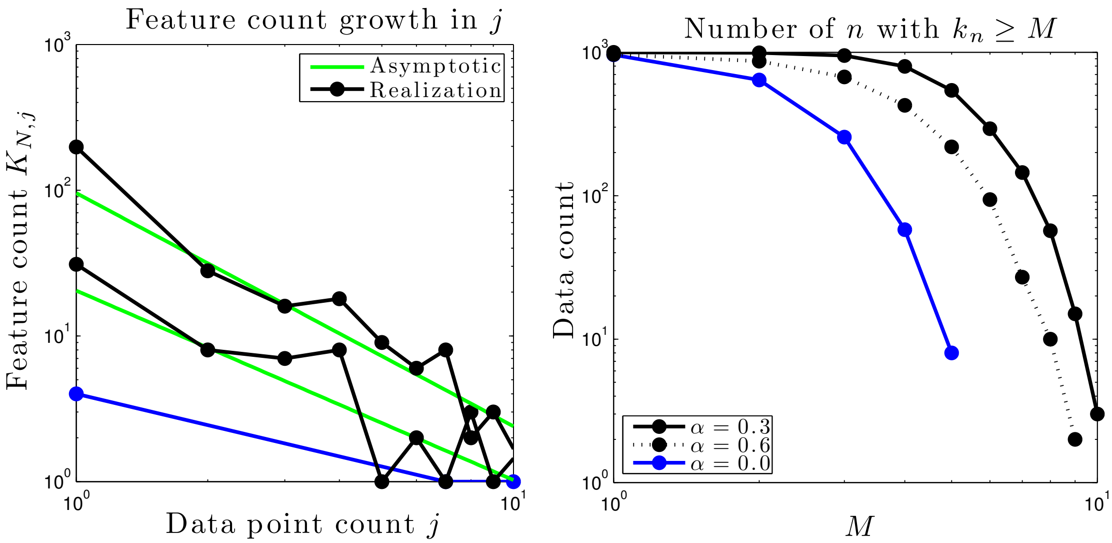


Figure 3.6: *Left*: Change in the number of features with exactly j representatives among N data points for fixed N as a function of j. The blue points, with connecting lines, are for $\alpha = 0$; middle black are for $\alpha = 0.3$, upper black for $\alpha = 0.6$. The green lines show the theoretical asymptotic behavior in j (Eqs. (3.10) and (3.11)) for the two $\alpha > 0$ cases. *Right*: Change in the number of data points, indexed by n, with number of feature assignments k_n greater than some positive, real-valued M as M increases. Neither the $\alpha = 0$ case (blue) nor the $\alpha > 0$ cases (black) exhibit Type III power laws.

lines for the $\alpha > 0$ case illustrate the approximation of Eq. (3.11). Around $j = 10$, we see that the number of feaures exhibited by j data points, $K_{N,j}$, degenerates to mainly zero and one values. However, for smaller values of j we can still distinguish the power law trend.

On the right-hand side of Figure 3.6, we display the number of data points exhibiting more than M features for various values of M across the three values of α. Unlike the previous plots in Figure 3.5 and Figure 3.6, there is no power-law behavior for the cases $\alpha > 0$, as predicted in Section 3.6. We also note that here the $\alpha = 0.3$ curve does not lie between the $\alpha = 0$ and $\alpha = 0.6$ curves. Such an occurrence is not unusual in this case since, as we saw in Eq. (3.31), the rate of decrease is modulated by the total mass of the feature probabilities drawn from the beta process, which is random and not necessarily smaller when α is smaller.

Finally, since our simulation involves generating the underlying feature probabilities from the beta process as well as the actual feature assignments from repeated draws from the Bernoulli process, we may examine the feature probabilities themselves; see Figure 3.7. As usual, the blue points represent the classic, two-parameter ($\alpha = 0$) beta process. Black points represent $\alpha = 0.3$ (center) and $\alpha = 0.6$ (upper). Perhaps due to the fact that there is only the beta process noise to contend with in this aspect of the simulation (and not the combined randomness due to the beta process and Bernoulli process), we see the most

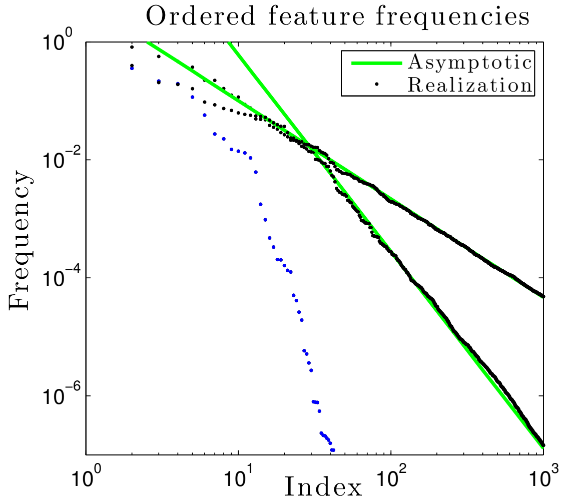


Figure 3.7: Feature probabilities from the beta process plotted in decreasing size order. Blue points represent probabilities from the $\alpha = 0$ case; center black points show $\alpha = 0.3$, and upper black points show $\alpha = 0.6$. The green lines show theoretical asymptotic behavior of the ranked probabilities (Eq. (3.33)).

striking demonstration of both power law behavior in the $\alpha > 0$ cases and faster decay in the $\alpha = 0$ case in this figure. The two $\alpha > 0$ cases clearly adhere to a power law that may be predicted from our results above and the Gnedin, Hansen, and Pitman (2007) results with C as in Eq. (3.28):

$$\#\{i : q_i \geq x\} \overset{a.s.}{\sim} Cx^{-\alpha} \quad x \downarrow 0. \tag{3.33}$$

Note that ranking the probabilities merely inverts the plot that would be created with x on the horizontal axis and $\{i : q_i \geq x\}$ on the vertical axis. The simulation demonstrates little noise about these power laws beyond the 100th ranked probability. The decay for $\alpha = 0$ is markedly faster than the other cases.

3.8 Experimental results

We have seen that the Poisson process formulation allows for an easy extension of the beta process to a three-parameter model. In this section we study this model empirically in the setting of the modeling of handwritten digits. Paisley, Zaas, et al. (2010) present results for this problem using a two-parameter beta process coupled with a discrete factor analysis model; we repeat those experiments with the three-parameter beta process. The data consists of 3,000 examples of handwritten digits, in particular 1,000 handwriting samples of each of the digits 3, 5, and 8 from the MNIST Handwritten Digits database (LeCun and Cortes, 1998; Roweis, 2007). Each handwritten digit is represented by a matrix of 28×28 pixels; we project these matrices into 50 dimensions using principal components analysis. Thus, our data takes the form $X \in \mathbb{R}^{50\times 3000}$, and we may apply the beta process factor model from Eq. (3.2) with $P = 50$ and $N = 3{,}000$ to discover latent structure in this data.

The generative model for X that we use is as follows (see Paisley, Zaas, et al., 2010):

$$\begin{aligned} X &= (W \circ Z)\Phi + E \\ Z &\sim \mathrm{BP} - \mathrm{BeP}(N, \gamma, \theta, \alpha) \\ \Phi_{k,p} &\overset{iid}{\sim} N(0, \rho_p) \\ W_{n,k} &\overset{iid}{\sim} N(0, \zeta) \\ E_{n,p} &\overset{iid}{\sim} N(0, \eta), \end{aligned} \tag{3.34}$$

with familiar beta process hyperparameters θ, α, and $\gamma = \mathbb{E}B_0$ and new (positive) variance hyperparameters $\{\rho_p\}_{p=1}^{P}, \zeta, \eta$. Recall from Eq. (3.2) that $X \in \mathbb{R}^{N\times P}$ is the data, $\Phi \in \mathbb{R}^{K\times P}$ is a matrix of factors, and $E \in \mathbb{R}^{N\times P}$ is an error matrix. Here, we introduce the weight matrix $W \in \mathbb{R}^{N\times K}$, which modulates the binary factor loadings $Z \in \mathbb{R}^{N\times K}$. In Eq. (3.34), $\circ$ denotes elementwise multiplication, and the indices have ranges $n \in \{1, \ldots, N\}, k \in \{1, \ldots, K\}, p \in \{1, \ldots, P\}$. Since we draw Z from a beta-Bernoulli process, the dimension K is theoretically infinite in the generative model notation of Eq. (3.34). However, we have seen that the number of columns of Z with nonzero entries is a.s. finite. We use K to denote this number.

We initialized both the two-parameter and the three-parameter models with the same number of latent features, $K = 200$, and the same values for all shared parameters (i.e., every variable except the new discount parameter α). We ran the experiment for 2,000 MCMC iterations, noting that the MCMC runs in both models seem to have reached equilibrium by 500 iterations (see Figures 3.8 and 3.9).

Figures 3.8 and 3.9 show the sampled values of various parameters as a function of MCMC iteration. In particular, we see how the number of features K (Figure 3.8), the concentration parameter θ, and the discount parameter α (Figure 3.9) change over time. All three graphs illustrate that the three-parameter model takes a longer time to reach equilibrium than the two-parameter model (approximately 500 iterations vs. approximatively 100 iterations). However, once at equilibrium, the sampling time series associated with the three-parameter iterations exhibit lower autocorrelation than the samples associated with the two-parameter iterations (Figure 3.10). In the implementation of both the original two-parameter model and the three-parameter model, the range for θ is considered to be bounded above by approximately 100 for computational reasons (in accordance with the original methodology of Paisley, Zaas, et al. (2010)). As shown in Figure 3.9, this bound affects sampling in the two-parameter experiment whereas, after burn-in, the effect is not noticeable in the three-parameter experiment. While the discount parameter α also comes close to the lower boundary of its discretization (Figure 3.9)—which cannot be exactly zero due to computational concerns—the samples nonetheless seem to explore the space well.

We can see from Figure 3.10 that the estimated value of the concentration parameter θ is much lower when the discount parameter α is also estimated. This behavior may be seen to result from the fact that the power law growth of the expected number of represented

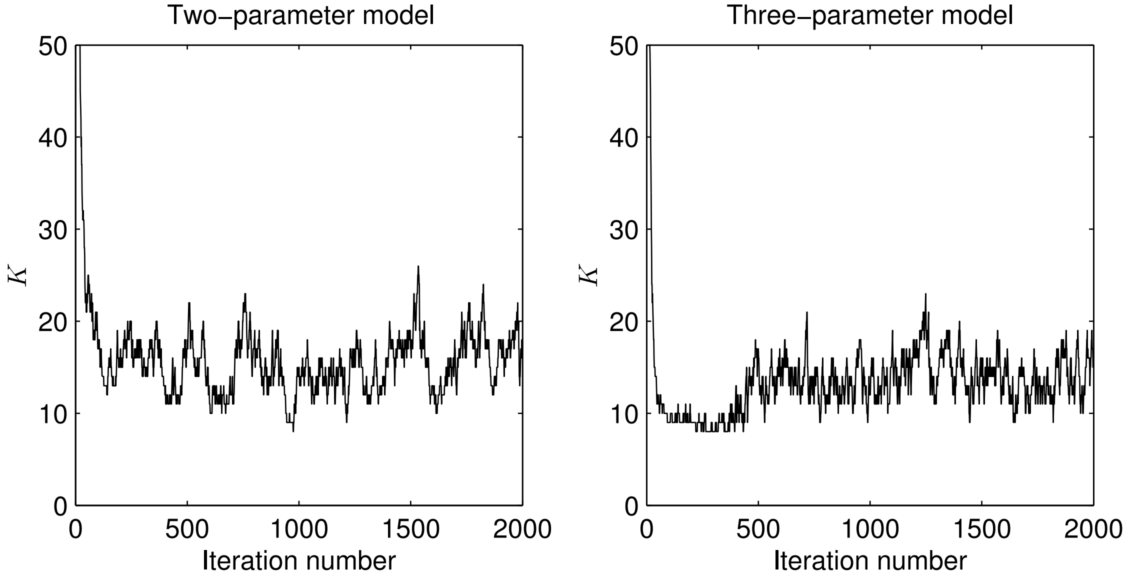


Figure 3.8: The number of latent features K as a function of the MCMC iteration. Results for the original, two-parameter model are represented on the *left*, and results for the new, three-parameter model are illustrated on the *right*.

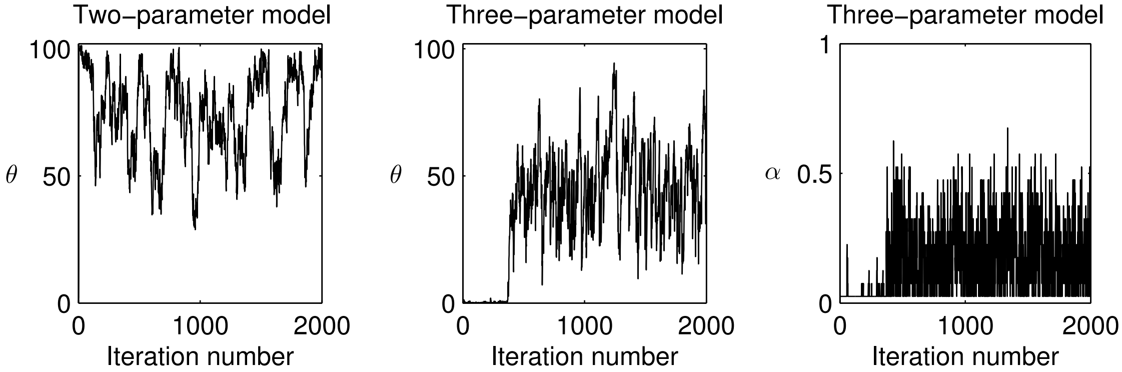


Figure 3.9: The random values drawn for the hyperparameters as a function of the MCMC iteration. Draws for the concentration parameter θ under the two-parameter model are shown on the *left*, and draws for θ under the three-parameter model are shown in the *middle*. On the *right* are draws of the new discount parameter α under the three-parameter model.

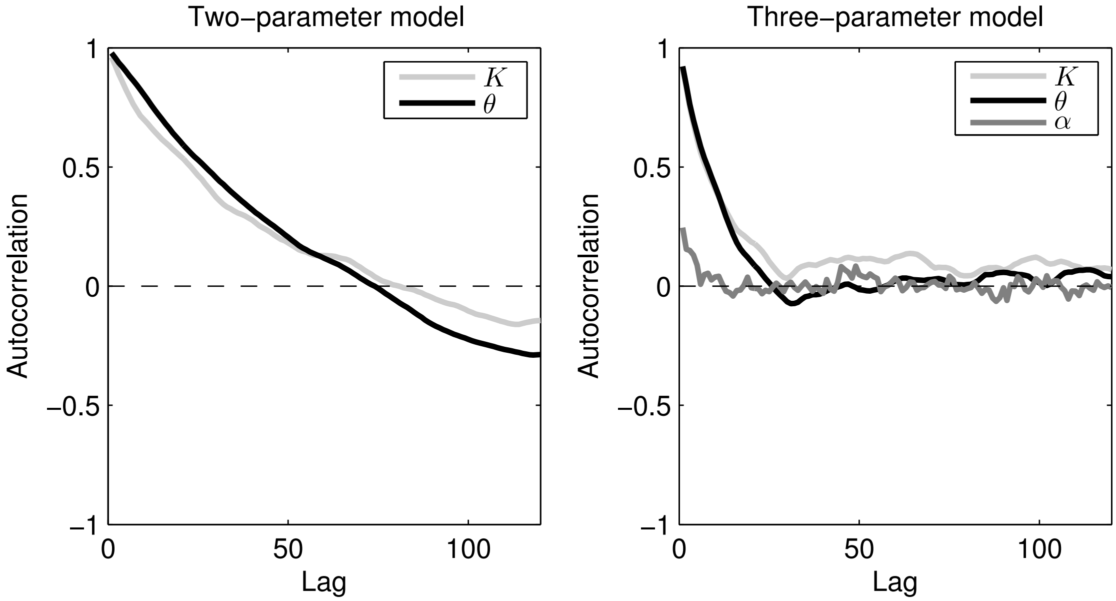


Figure 3.10: Autocorrelation of the number of factors K, concentration parameter θ, and discount parameter α for the MCMC samples after burn-in (where burn-in is taken to end at 500 iterations) under the two-parameter model (*left*) and three-parameter model (*right*).

features Φ_N in the $\alpha > 0$ case yields a generally higher expected number of features than in the $\alpha = 0$ case for a fixed concentration parameter θ. Further, we see from Eq. (3.32) that the expected number of features when $\alpha = 0$ is linear in θ. Therefore, if we instead fix the number of features, the $\alpha = 0$ model can compensate by increasing θ over the $\alpha > 0$ model. Indeed, we see in Figure 3.8 that the number of features discovered by both models is roughly equal; in order to achieve this number of features, the $\alpha = 0$ model seems to be compensating by overestimating the concentration parameter θ.

To get a sense of the actual output of the model, we can look at some of the learned features. In particular, we collected the set of features from the last MCMC iteration in each model. The kth feature is expressed or not for the nth data point according to whether Z_{nk} is one or zero. Therefore, we can find the most-expressed features across the data set using the set of features on this iteration as well as the sampled Z matrix on this iteration. We plot the nine most-expressed features under each model in Figure 3.11. In both models, we can see how the features have captured distinguishing features of the 3, 5, and 8 digits.

Finally, we note that the three-parameter version of the algorithm is competitive with the two-parameter version in running time once equilibrium is reached. After the burn-in regime of 500 iterations, the average running time per iteration under the three-parameter model is 14.5 seconds, compared with 11.7 seconds average running time per iteration under the two-parameter model.

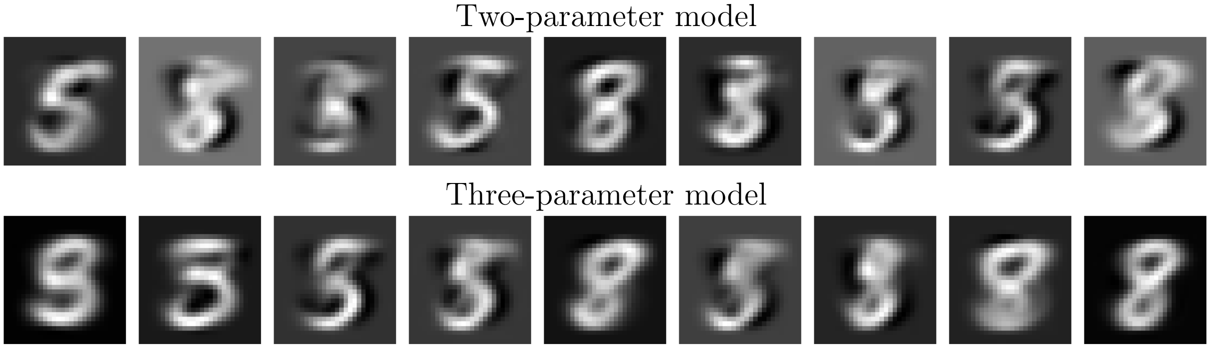


Figure 3.11: *Upper*: The top nine features by sampled representation across the data set on the final MCMC iteration for the original, two-parameter model. *Lower*: The top nine features determined in the same way for the new, three-parameter model.

3.9 Discussion

We have shown that the stick-breaking representation of the beta process due to Paisley, Zaas, et al. (2010) can be obtained directly from the representation of the beta process as a completely random measure. With this result in hand the set of connections between the beta process, stick-breaking, and the Indian buffet process are essentially as complete as those linking the Dirichlet process, stick-breaking, and the Chinese restaurant process.

We have also shown that this approach motivates a three-parameter generalization of the stick-breaking representation of Paisley, Zaas, et al. (2010), which is the analog of the Pitman-Yor generalization of the stick-breaking representation for the Dirichlet process. We have shown that Type I and Type II power laws follow from this three-parameter model. We have also shown that Type III power laws cannot be obtained within this framework. It is an open problem to discover useful classes of stochastic processes that provide such power laws.

3.A A Markov chain Monte Carlo algorithm

Posterior inference under the three-parameter model can be performed with a Markov chain Monte Carlo (MCMC) algorithm. Many conditionals have simple forms that allow Gibbs sampling although others require further approximation. Most of our sampling steps are as in Paisley, Zaas, et al. (2010) with the notable exceptions of a new sampling step for the discount parameter α and integration of the discount parameter α into the existing framework. We describe the full algorithm here.

Notation and auxiliary variables

Call the index i in Eq. (3.14) the *round*. Then introduce the round-indicator variables r_k such that $r_k = i$ exactly when the kth atom, where k indexes the sequence $(\psi_{1,1}, \ldots, \psi_{1,C_1}, \psi_{2,1}, \ldots, \psi_{2,C_2}, \ldots)$, occurs in round i. We may write

$$r_k := 1 + \sum_{i=1}^{\infty} \mathbb{1}\left\{ \sum_{j=1}^{i} C_j < k \right\}.$$

To recover the round lengths C from $r = (r_1, r_2, \ldots)$, note that

$$C_i = \sum_{k=1}^{\infty} \mathbb{1}(r_k = i). \tag{3.35}$$

With the definition of the round indicators r in hand, we can rewrite the beta process B as

$$B = \sum_{k=1}^{\infty} V_{k,r_k} \prod_{j=1}^{r_k - 1} (1 - V_{k,j}) \delta_{\psi_k},$$

where $V_{k,j} \overset{iid}{\sim} \text{Beta}(1-\alpha, \theta + i\alpha)$ and $\psi_k \overset{iid}{\sim} \gamma^{-1} B_0$ as usual although the indexing is not the same as in Eq. (3.14). It follows that the expression of the kth feature for the nth data point is given by

$$Z_{n,k} \sim \text{Bern}\left(\pi_k\right), \quad \pi_k := V_{k,r_k} \prod_{j=1}^{r_k - 1} (1 - V_{k,j}).$$

We also introduce notation for the number of data points in which the kth feature is, respectively, expressed and not expressed:

$$m_{1,k} := \sum_{n=1}^{N} \mathbb{1}(Z_{n,k} = 1), \quad m_{0,k} := \sum_{n=1}^{N} \mathbb{1}(Z_{n,k} = 0)$$

Finally, let K be the number of represented features; i.e., $K := \#\{k : m_{1,k} > 0\}$. Without loss of generality, we assume the represented features are the first K features in the index k. The new quantities $\{r_k\}$, $\{m_{1,k}\}$, $\{m_{0,k}\}$, and K will be used in describing the sampler steps below.

Latent indicators

First, we describe the sampling of the round indicators $\{r_k\}$ and the latent feature indicators $\{Z_{n,k}\}$. In these and other steps in the MCMC algorithm, we integrate out the stick-breaking proportions $\{V_i\}$.

Round indicator variables

We wish to sample the round indicator r_k for each feature k with $1 \leq k \leq K$. We can write the conditional for r_k as

$$\begin{aligned} & p(r_k = i | \{r_l\}_{l=1}^{k-1}, \{Z_{n,k}\}_{n=1}^{N}, \theta, \alpha, \gamma) \\ & \quad \propto \; p(\{Z_{n,k}\}_{n=1}^{N} | r_k = i, \theta, \alpha) p(r_k = i | \{r_l\}_{l=1}^{k-1}). \end{aligned} \tag{3.36}$$

It remains to calculate the two factors in the product.

For the first factor in Eq. (3.36), we write out the integration over stick-breaking proportions and approximate with a Monte Carlo integral:

$$\begin{aligned} p(\{Z_{n,k}\}_{n=1}^{N} | r_k = i, \theta, \alpha) &= \int_{[0,1]^i} \pi_k^{m_{1,k}} (1 - \pi_k)^{m_{0,k}} \, dV \\ &\approx \frac{1}{S} \sum_{s=1}^{S} (\pi_k^{(s)})^{m_{1,k}} (1 - \pi_k^{(s)})^{m_{0,k}}. \end{aligned} \tag{3.37}$$

Here, $\pi_k^{(s)} := V_{k,r_k}^{(s)} \prod_{j=1}^{r_k - 1} (1 - V_{k,j}^{(s)})$, and $V_{k,j}^{(s)} \overset{indep}{\sim} \text{Beta}(1 - \alpha, \theta + j\alpha)$. Also, S is the number of samples in the sum approximation. Note that the computational trick employed in Paisley, Zaas, et al. (2010) for sampling the $\{V_i\}$ more efficiently than directly using the approximation above relies on the first parameter of the beta distribution being equal to one; therefore, the sampling described above, without further tricks, is exactly the sampling that must be used in this more general parameterization.

For the second factor in Eq. (3.36), there is no dependence on the α parameter, so the draws are the same as in Paisley, Zaas, et al. (2010). For $R_k := \sum_{j=1}^{k} \mathbb{1}(r_j = r_k)$, we have

$$\begin{aligned} & p(r_k = r | \gamma, \{r_l\}_{l=1}^{k-1}) \\ & = \begin{cases} 0 & r < r_{k-1} \\ \frac{1 - \sum_{i=1}^{R_k - 1} \text{Poisson}(i|\gamma)}{1 - \sum_{i=1}^{R_{k-1} - 1} \text{Poisson}(i|\gamma)} & r = r_{k-1} \\ \left(1 - \frac{1 - \sum_{i=1}^{R_k - 1} \text{Poisson}(i|\gamma)}{1 - \sum_{i=1}^{R_{k-1} - 1} \text{Poisson}(i|\gamma)}\right) (1 - \text{Poisson}(0|\gamma)) \, \text{Poisson}(0|\gamma)^{h-1} & r = r_{k-1} + h \end{cases} \end{aligned}$$

for each $h \geq 1$. Note that these draws make the approximation that the first K features correspond to the first K tuples (i, j) in the double sum of Eq. (3.14); these orderings do not in general agree.

To complete the calculation of the posterior for r_k, we need to sum over all values of i to normalize $p(r_k = i | \{r_l\}_{l=1}^{k-1}, \{Z_{n,k}\}_{n=1}^{N}, \theta, \alpha, \gamma)$. Since this is not computationally feasible, an alternative method is to calculate Eq. (3.36) for increasing values of i until the result falls below a pre-determined threshold.

Factor indicators

In finding the posterior for the kth feature indicator in the nth latent factor, $Z_{n,k}$, we can integrate out both $\{V_i\}$ and the weight variables $\{W_{n,k}\}$. The conditional for $Z_{n,k}$ is

$$\begin{aligned} & p(Z_{n,k}|X_{n,\cdot}, \Phi, Z_{n,-k}, r, \theta, \alpha, \eta, \zeta) \\ &= p(X_{n,\cdot}|Z_{n,\cdot}, \Phi, \eta, \zeta) p(Z_{n,k}|r, \theta, \alpha, Z_{n,-k}). \end{aligned} \tag{3.38}$$

First, we consider the likelihood. For this factor, we integrate out W explicitly:

$$\begin{aligned} & p(X_{n,\cdot}|Z_{n,\cdot}, \Phi, \eta, \zeta) \\ &= \int_W p(X_{n,\cdot}|Z_{n,\cdot}, \Phi, W, \eta) p(W|\zeta) \\ &= \int_{W_{n,I}} N(X_{n,\cdot}|W_{n,I}\Phi_{I,\cdot}, \eta I_P) N(W_{n,I}|0_{|I|}, \zeta I_{|I|}) dW_{n,I} \\ & \quad \text{where } I = \{i : Z_{n,i} = 1\} \\ &= N\left(X_{n,\cdot}|0_P, \left[\eta^{-1} I_P - \eta^{-2}\Phi_{I,\cdot}\left(\eta^{-1}\Phi_{I,\cdot}^\top \Phi_{I,\cdot} + \zeta^{-1} I_{|I|}\right)^{-1} \Phi_{I,\cdot}^\top\right]^{-1}\right) \\ &= N\left(X_{n,\cdot}|0_P, \eta I_P + \zeta \Phi_{I,\cdot}\Phi_{I,\cdot}^\top\right), \end{aligned}$$

where the final step follows from the Sherman-Morrison-Woodbury lemma.

For the second factor in Eq. (3.38), we can write

$$p(Z_{n,k}|r, \theta, \alpha, Z_{n,-k}) = \frac{p(Z_n|r, \theta, \alpha)}{p(Z_{n,-k}|r, \theta, \alpha)},$$

and the numerator and denominator can both be estimated as integrals over V using the same Monte Carlo integration trick as in Eq. (3.37).

Hyperparameters

Next, we describe sampling for the three parameters of the beta process. The mass and concentration parameters are shared by the two-parameter process; the discount parameter is unique to the three-parameter beta process.

Mass parameter

With the round indicators $\{r_k\}$ in hand as from Appendix 3.A above, we can recover the round lengths $\{C_i\}$ with Eq. (3.35). Assuming an improper gamma prior on γ—with both shape and inverse scale parameters equal to zero—and recalling the iid Poisson generation of the $\{C_i\}$, the posterior for γ is

$$p(\gamma|r, Z, \theta, \alpha) = \text{Gamma}(\gamma|\sum_{i=1}^{r_K} C_i, r_K).$$

Note that it is necessary to sample γ since it occurs in, e.g., the conditional for the round indicator variables (Appendix 3.A).

Concentration parameter

The conditional for θ is

$$p(\theta|Z, r, \alpha) \propto p(\theta) \prod_{k=1}^{K} p(Z|r, \theta, \alpha).$$

Again, we calculate the likelihood factors $p(Z|r, \theta, \alpha)$ with a Monte Carlo approximation as in Eq. (3.37). In order to find the conditional over θ from the likelihood and prior, we further approximate the space of $\theta > 0$ by a discretization around the previous value of θ in the Monte Carlo sampler: $\{\theta_{prev} + t\Delta\theta\}_{t=S}^{t=T}$, where S and T are chosen so that all potential new θ values are nonnegative and so that the tails of the distribution fall below a pre-determined threshold. To complete the description, we choose the improper prior $p(\theta) \propto 1$.

Discount parameter

We sample the discount parameter α in a similar manner to θ. The conditional for α is

$$p(\alpha|Z, r, \theta) \propto p(\alpha) \prod_{k=1}^{K} p(Z|r, \theta, \alpha).$$

As usual, we calculate the likelihood factors $p(Z|r, \theta, \alpha)$ with a Monte Carlo approximation as in Eq. (3.37). While we discretize the sampling of α as we did for θ, note that sampling α is more straightforward since α must lie in $[0, 1]$. Therefore, the choice of $\Delta\alpha$ completely characterizes the discretization of the interval. In particular, to avoid endpoint behavior, we consider new values of α among $\{\Delta\alpha/2 + t\Delta\alpha\}_{t=0}^{(\Delta\alpha)^{-1}-1}$. Moreover, the choice of $p(\alpha) \propto 1$ is, in this case, a proper prior for α.

Factor analysis components

In order to use the beta process as a prior in the factor analysis model described in Eq. (3.2), we must also describe samplers for the feature matrix Φ and weight matrix W.

Feature matrix

The conditional for the feature matrix Φ is

$$\begin{aligned} p(\Phi_{\cdot,p}|X, W, Z, \eta, \rho_p) &\propto p(X_{\cdot,p}|\Phi_{\cdot,p}, W, Z, \eta I_N) p(\Phi_{\cdot,p}|\rho_p) \\ &= N(X_{\cdot,p}|(W \circ Z)\Phi_{\cdot,p}, \eta I_N) N(\Phi_{\cdot,p}|0_K, \rho_p I_K) \\ &\propto N\left(\Phi_{\cdot,p}|\mu, \Sigma\right), \end{aligned}$$

where, in the final line, the variance is defined as follows:

$$\Sigma := \left(\eta^{-1}(W \circ Z)^\top (W \circ Z) + \rho_p^{-1} I_K\right)^{-1},$$

and similarly for the mean:

$$\mu := \Sigma \eta^{-1} (W \circ Z)^\top X_{\cdot,p}.$$

Weight matrix

Let $I = \{i : Z_{n,i} = 1\}$. Then the conditional for the weight matrix W is

$$\begin{aligned} p(W_{n,I} | X, Z, \Phi, \eta) &\propto p(X_{n,\cdot} | \Phi_{I,\cdot}, W_{n,I}, \eta) p(W_{n,I} | \zeta) \\ &= N(X_{n,\cdot} | W_{n,I} \Phi_{I,\cdot}, \eta I_p) N(W_{n,I} | 0_{|I|}, \zeta I_{|I|}) \\ &\propto N(W_{n,I} | \tilde{\mu}, \tilde{\Sigma}), \end{aligned}$$

where, in the final line, the variance is defined as $\tilde{\Sigma} := \left(\eta^{-1} \Phi_{I,\cdot} \Phi_{I,\cdot}^\top + \zeta^{-1} I_{|I|}\right)^{-1}$, and the mean is defined as $\tilde{\mu} := \tilde{\Sigma} \eta^{-1} X_{n,\cdot} \Phi_{I,\cdot}^\top$.

Chapter 4

Combinatorial clustering and the beta negative binomial process

We develop a Bayesian nonparametric approach to a general family of latent class problems in which individuals can belong simultaneously to multiple classes and where each class can be exhibited multiple times by an individual. We introduce a combinatorial stochastic process known as the *negative binomial process* (NBP) as an infinite-dimensional prior appropriate for such problems. We show that the NBP is conjugate to the beta process, and we characterize the posterior distribution under the beta-negative binomial process (BNBP) and hierarchical models based on the BNBP (the HBNBP). We study the asymptotic properties of the BNBP and develop a three-parameter extension of the BNBP that exhibits power-law behavior. We derive MCMC algorithms for posterior inference under the HBNBP, and we present experiments using these algorithms in the domains of image segmentation, object recognition, and document analysis.

4.1 Introduction

In traditional clustering problems the goal is to induce a set of latent classes and to assign each data point to one and only one class. This problem has been approached within a model-based framework via the use of finite mixture models, where the mixture components characterize the distributions associated with the classes, and the mixing proportions capture the mutual exclusivity of the classes (Fraley and Raftery, 2002; McLachlan and Basford, 1988). In many domains in which the notion of latent classes is natural, however, it is unrealistic to assign each individual to a single class. For example, in genetics, while it may be reasonable to assume the existence of underlying ancestral populations that define distributions on observed alleles, each individual in an existing population is likely to be a blend of the patterns associated with the ancestral populations. Such a genetic blend is known as an *admixture* (Pritchard, Stephens, and Donnelly, 2000). A significant literature on model-based approaches to admixture has arisen in recent years (Blei, Ng, and Jordan, 2003;

Erosheva and Fienberg, 2005; Pritchard, Stephens, and Donnelly, 2000), with applications to a wide variety of domains in genetics and beyond, including document modeling and image analysis.[1]

Model-based approaches to admixture are generally built on the foundation of mixture modeling. The basic idea is to treat each individual as a collection of data, with an exchangeability assumption imposed for the data within an individual but not between individuals. For example, in the genetics domain the intra-individual data might be a set of genetic markers, with marker probabilities varying across ancestral populations. In the document domain the intra-individual data might be the set of words in a given document, with each document (the individual) obtained as a blend across a set of underlying "topics" that encode probabilities for the words. In the image domain, the intra-individual data might be visual characteristics like edges, hue, and location extracted from image patches. Each image is then a blend of object classes (e.g., grass, sky, or car), each defining a distinct distribution over visual characteristics. In general, this blending is achieved by making use of the probabilistic structure of a finite mixture but using a different sampling pattern. In particular, mixing proportions are treated as random effects that are drawn once per individual, and the data associated with that individual are obtained by repeated draws from a mixture model having that fixed set of mixing proportions. The overall model is a hierarchical model, in which mixture components are shared among individuals and mixing proportions are treated as random effects.

Although the literature has focused on using finite mixture models in this context, there has also been a growing literature on Bayesian nonparametric approaches to admixture models, notably the *hierarchical Dirichlet process* (HDP) (Teh, Jordan, et al., 2006), where the number of shared mixture components is infinite. Our focus in the current chapter is also on nonparametric methods, given the open-ended nature of the inferential objects with which real-world admixture modeling is generally concerned.

Although viewing an admixture as a set of repeated draws from a mixture model is natural in many situations, it is also natural to take a different perspective, akin to latent trait modeling, in which the individual (e.g., a document or a genotype) is characterized by the set of "traits" or "features" that it possesses, and where there is no assumption of mutual exclusivity. Here the focus is on the individual and not on the "data" associated with an individual. Indeed, under the exchangeability assumption alluded to above it is natural to reduce the repeated draws from a mixture model to the counts of the numbers of times that each mixture component is selected, and we may wish to model these counts directly. We may further wish to consider hierarchical models in which there is a linkage among the counts for different individuals.

This idea has been made explicit in a recent line of work based on the *beta process*. Originally developed for survival analysis, where an integrated form of the beta process was used as a model for random hazard functions (Hjort, 1990), more recently it has been observed

[1] While we refer to such models generically as "admixture models," we note that they are also often referred to as *topic models* or *mixed membership models*.

that the beta process also provides a natural framework for latent feature modeling (Thibaux and Jordan, 2007). In particular, as we discuss in detail in Section 4.2, a draw from the beta process yields an infinite collection of coin-tossing probabilities. Tossing these coins—a draw from a *Bernoulli process*—one obtains a set of binary features that can be viewed as a description of an admixed individual. A key advantage of this approach is the conjugacy between the beta and Bernoulli processes: this property allows for tractable inference, despite the countable infinitude of coin-tossing probabilities. A limitation of this approach, however, is its restriction to binary features; indeed, one of the virtues of the mixture-model-based approach is that a given mixture component can be selected more than once, with the total number of selections being random.

We develop a model for admixture that meets all of the desiderata outlined thus far. Unlike the Bernoulli process likelihood, our featural model allows each feature to be exhibited any non-negative integer number of times by an individual. Unlike admixture models based on the HDP, our model cohesively includes a random total number of features (e.g., words or traits) per individual (e.g., a document or genotype).

As inspiration, we note that in the setting of classical random variables, beta-Bernoulli conjugacy is not the only form of conjugacy involving the beta distribution—the negative binomial distribution is also conjugate to the beta. Anticipating the value of conjugacy in the setting of nonparametric models, we define and develop a stochastic process analogue of the negative binomial distribution, which we refer to as the *negative binomial process* (NBP),[2] and provide a rigorous proof of its conjugacy to the beta process. We use this process as part of a new model—the *hierarchical beta negative binomial process* (HBNBP)—based on the NBP and the hierarchical beta process (Thibaux and Jordan, 2007). Our theoretical and experimental development focus on the usefulness of the HBNBP in the admixture setting, where flexible modeling of feature totals can lead to improved inferential accuracy (see Figure 4.8 and the surrounding discussion). However, the utility of the HBNBP is not limited to the admixture setting and should extend readily to the modeling of latent factors and the identification of more general latent features. Moreover, the negative binomial component of our model offers addtional flexibility in the form of a new parameter unavailable in either the Bernoulli or multinomial likelihoods traditionally explored in Bayesian nonparametrics.

The remainder of the chapter is organized as follows. In Section 4.2 we present the framework of completely random measures that provides the formal underpinnings for our work. We discuss the Bernoulli process, introduce the NBP, and demonstrate the conjugacy of both to the beta process in Section 4.3. Section 4.4 focuses on the problem of modeling admixture and on general hierarchical modeling based on the negative binomial process. Section 4.5 and Section 4.6 are devoted to a study of the asymptotic behavior of the NBP with a beta process prior, which we call the beta-negative binomial process (BNBP). We describe algorithms for posterior inference in Section 4.7. Finally, we present experimental results. First, we use the BNBP to define a generative model for summaries of terrorist

[2] Zhou et al. (2012) have independently investigated negative binomial processes in the context of integer matrix factorization. We discuss their concurrent contributions in more detail in Section 4.4.

incidents with the goal of identifying the perpetrator of a given terrorist attack in Section 4.8. Second, we demonstrate the utility of a finite approximation to the BNBP in the domain of automatic image segmentation in Section 4.9. Section 4.10 presents our conclusions.

4.2 Completely random measures

In this section we review the notion of a completely random measure (CRM), a general construction that yields random measures that are closely tied to classical constructions involving sets of independent random variables. We present CRM-based constructions of several of the stochastic processes used in Bayesian nonparametrics, including the beta process, gamma process, and Dirichlet process. In the following section we build on the foundations presented here to consider additional stochastic processes.

Consider a probability space $(\Psi, \mathcal{F}, \mathbb{P})$. A *random measure* is a random element μ such that $\mu(A)$ is a non-negative random variable for any A in the sigma algebra $\mathcal{F}$. A *completely random measure* (CRM) μ is a random measure such that, for any disjoint, measurable sets $A, A' \in \mathcal{F}$, we have that $\mu(A)$ and $\mu(A')$ are independent random variables (Kingman, 1967). Completely random measures can be shown to be composed of at most three components:

1. A *deterministic measure*. For deterministic μ_{det}, it is trivially the case that $\mu_{det}(A)$ and $\mu_{det}(A')$ are independent for disjoint A, A'.
2. A *set of fixed atoms*. Let $(u_1, \ldots, u_L) \in \Psi^L$ be a collection of deterministic locations, and let $(\eta_1, \ldots, \eta_L) \in \mathbb{R}_+^L$ be a collection of independent random weights for the atoms. The collection may be countably infinite, in which case we say $L = \infty$. Then let $\mu_{fix} = \sum_{l=1}^{L} \eta_l \delta_{u_l}$. The independence of the η_l implies the complete randomness of the measure.
3. An *ordinary component*. Let ν_{PP} be a Poisson process intensity on the space $\Psi \times \mathbb{R}_+$. Let $\{(v_1, \xi_1), (v_2, \xi_2), \ldots\}$ be a draw from the Poisson process with intensity ν_{PP}. Then the ordinary component is the measure $\mu_{ord} = \sum_{j=1}^{\infty} \xi_j \delta_{v_j}$. Here, the complete randomness follows from properties of the Poisson process.

One observation from this componentwise breakdown of CRMs is that we can obtain a countably infinite collection of random variables, the ξ_j, from the Poisson process component if ν_{PP} has infinite total mass (but is still sigma-finite). Consider again the criterion that a CRM μ yield independent random variables when applied to disjoint sets. In light of the observation about the collection $\{\xi_j\}$, this criterion may now be seen as an extension of an independence assumption in the case of a finite set of random variables. We cover specific examples next.

Beta process

The *beta process* (Hjort, 1990; Kim, 1999a; Thibaux and Jordan, 2007) is an example of a CRM. It has the following parameters: a *mass parameter* $\gamma > 0$, a *concentration parameter*

$\theta > 0$, a purely atomic measure $H_{fix} = \sum_l \rho_l \delta_{u_l}$ with $\gamma\rho_l \in (0,1)$ for all l a.s., and a purely continuous probability measure H_{ord} on Ψ. Note that we have explicitly separated out the mass parameter γ so that, e.g., H_{ord} is a probability measure; in Thibaux and Jordan (2007), these two parameters are expressed as a single measure with total mass equal to γ. Typically, though, the normalized measure H_{ord} is used separately from the mass parameter γ (as we will see below), so the notational separation is convenient. Often the final two measure parameters are abbreviated as their sum: $H = H_{fix} + H_{ord}$.

Given these parameters, the beta process has the following description as a CRM:

1. The deterministic measure is uniformly zero.
2. The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, where L is potentially infinite though typically finite. Atom weight η_l has distribution

   $$\eta_l \overset{indep}{\sim} \text{Beta}\left(\theta\gamma\rho_l, \theta(1-\gamma\rho_l)\right), \tag{4.1}$$

   where the ρ_l parameters are the weights in the purely atomic measure H_{fix}.
3. The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure

   $$\nu(db) = \gamma\theta b^{-1}(1-b)^{\theta-1}\, db, \tag{4.2}$$

   which is sigma-finite with finite mean. It follows that the number of atoms in this component will be countably infinite, but the atom weights will have finite sum.

As in the original specification of Hjort (1990) and Kim (1999a), Eq. (4.2) can be generalized by allowing θ to depend on the Ψ coordinate. The homogeneous intensity in Eq. (4.2) seems to be used predominantly in practice (Thibaux and Jordan, 2007; Fox et al., 2009) though, and we focus on it here for ease of exposition. Nonetheless, we note that our results below extend easily to the non-homogeneous case.

The CRM is the sum of its components. Therefore, we may write a draw from the beta process as

$$B = \sum_{k=1}^{\infty} b_k \delta_{\psi_k} \triangleq \sum_{l=1}^{L} \eta_l \delta_{u_l} + \sum_{j=1}^{\infty} \xi_j \delta_{v_j}, \tag{4.3}$$

with atom locations equal to the union of the fixed atom and ordinary component atom locations $\{\psi_k\}_k = \{u_l\}_{l=1}^L \cup \{v_j\}_{j=1}^\infty$. Notably, B is a.s. discrete. We denote a draw from the beta process as $B \sim \text{BP}(\theta, \gamma, H)$. The provenance of the name “beta process” is now clear; each atom weight in the fixed atomic component is beta-distributed, and the Poisson process intensity generating the ordinary component is that of an improper beta distribution.

From the above description, the beta process provides a prior on a potentially infinite vector of weights, each in $(0,1)$ and each associated with a corresponding parameter $\psi \in \Psi$. The potential countable infinity comes from the Poisson process component. The weights in $(0,1)$ may be interpreted as probabilities, though not as a distribution across the indices as we note that they need not sum to one. We will see in Section 4.4 that the beta process is appropriate for feature modeling (Thibaux and Jordan, 2007; Griffiths and Ghahramani, 2006). In this context, each atom, indexed by k, of B corresponds to a feature. The atom weights $\{b_k\}$, which are each in $[0,1]$ a.s., can be viewed as representing the frequency with which each feature occurs in the dataset. The atom locations $\{\psi_k\}$ represent parameters associated with the features that can be used in forming a likelihood.

In Section 4.5, we will show that an extension to the beta process called the *three-parameter beta process* has certain desirable properties beyond the classic beta process, in particular its ability to generate power-law behavior (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012), which roughly says that the number of features grows as a power of the number of data points. In the three-parameter case, we introduce a *discount parameter* $\alpha \in (0,1)$ with $\theta > -\alpha$ and $\gamma > 0$ such that:

1. There is again no deterministic component.
2. The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, with L potentially infinite but typically finite. Atom weight η_l has distribution $\eta_l \overset{indep}{\sim} \text{Beta}\left(\theta\gamma\rho_l - \alpha, \theta(1-\gamma\rho_l) + \alpha\right)$, where the ρ_l parameters are the weights in the purely atomic measure H_{fix} and we now have the constraints $\theta\gamma\rho_l - \alpha, \theta(1-\gamma\rho_l) + \alpha \geq 0$.
3. The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure:

$$\nu(db) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} b^{-1-\alpha}(1-b)^{\theta+\alpha-1}\, db.$$

Again, we focus on the homogeneous intensity ν as in the beta process case though it is straightforward to allow θ to depend on coordinates in Ψ.

In this case, we again have the full process draw B as in Eq. (4.3), and we say $B \sim 3\text{BP}(\alpha, \theta, \gamma, H)$.

Full beta process

The specification that the atom parameters in the beta process be of the form $\theta\gamma\rho_l$ and $\theta(1-\gamma\rho_l)$ can be unnecessarily constraining; $\theta\gamma\rho_l - \alpha$ and $\theta(1-\gamma\rho_l)+\alpha$ are even more unwieldy in the power-law case. Indeed, the classical beta distribution has two free parameters. Yet, in the beta process as described above, θ and γ are determined as part of the Poisson process intensity, so there is essentially one free parameter for each of the beta-distributed weights associated with the atoms (Eq. (4.1)). A related problematic issue is that the beta process forces the two parameters in the beta distribution associated with each atom to sum to θ, which is constant across all of the atoms.

One way to remove these restrictions is to allow $\theta = \theta(\psi)$, a function of the position $\psi \in \Psi$ as mentioned above. However, we demonstrate in Appendix 4.A that there are reasons to prefer a fixed concentration parameter θ for the ordinary component; there is a fundamental relation between this parameter and similar parameters in other common CRMs (e.g., the Dirichlet process, which we describe in Section 4.2). Moreover, the concern here is entirely centered on the behavior of the fixed atoms of the process, and letting θ depend on ψ retains the unusual—from a classical parametric perspective—form of the beta distribution in Eq. (4.1). As an alternative, we provide a generalization of the beta process that more closely aligns with the classical perspective in which we allow two general beta parameters for each atom. As we will see, this generalization is natural, and indeed necessary, in considering conjugacy.

We thus define the *full beta process* (RBP) as having the following parameterization: a *mass parameter* $\gamma > 0$, a *concentration parameter* $\theta > 0$, a number of fixed atoms $L \in \{0, 1, 2, \ldots\} \cup \{\infty\}$ with locations $(u_1, \ldots, u_L) \in \Psi^L$, two sets of strictly positive atom weight parameters $\{\rho_l\}_{l=1}^L$ and $\{\sigma_l\}_{l=1}^L$, and a purely continuous measure H_{ord} on Ψ. In this case, the atom weight parameters satisfy the simple condition $\rho_l, \sigma_l > 0$ for all $l \in \{1, \ldots, L\}$. This specification is the same as the beta process specification introduced above with the sole exception of a more general parameterization for the fixed atoms. We obtain the following CRM:

1. There is no deterministic measure.
2. There are L fixed atoms with locations $(u_1, \ldots, u_L) \in \Psi^L$ and corresponding weights $\eta_l \overset{indep}{\sim} \text{Beta}(\rho_l, \sigma_l)$.
3. The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure $\nu(db) = \gamma\theta b^{-1}(1-b)^{\theta-1}\, db$.

As discussed above, we favor the homogeneous intensity ν in exposition but note the straightforward extension to allow θ to depend on Ψ location.

We denote this CRM by $B \sim \text{RBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$.

Gamma process

While the beta process provides a countably infinite vector of frequencies in $(0, 1]$ with associated parameters ψ_k, it is sometimes useful to have a countably infinite vector of positive, real-valued quantities that can be used as rates rather than frequencies for features. We can obtain such a prior with the *gamma process* (Ferguson, 1973), a CRM with the following parameters: a *concentration parameter* $\theta > 0$, a *scale parameter* $c > 0$, a purely atomic measure $H_{fix} = \sum_l \rho_l \delta_{u_l}$ with $\forall l, \rho_l > 0$, and a purely continuous measure H_{ord} with support on Ψ. Its description as a CRM is as follows (Thibaux, 2008):

1. There is no deterministic measure.
2. The fixed atoms have locations $(u_1, \ldots, u_L) \in \Psi^L$, where L is potentially infinite but typically finite. Atom weight η_l has distribution $\eta_l \overset{indep}{\sim} \text{Gamma}(\theta\rho_l, c)$, where we use the shape-inverse-scale parameterization of the gamma distribution and where the ρ_l parameters are the weights in the purely atomic measure H_{fix}.

3. The ordinary component has Poisson process intensity $H_{ord} \times \nu$, where ν is the measure:

$$\nu(d\tilde{g}) = \theta \tilde{g}^{-1} \exp\left(-c\tilde{g}\right)\ d\tilde{g}. \tag{4.4}$$

As in the case of the beta process, the gamma process can be expressed as the sum of its components: $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k} \triangleq \sum_{l=1}^{L} \eta_l \delta_{u_l} + \sum_j \xi_j \delta_{v_j}$. We denote this CRM as $\tilde{G} \sim \Gamma\mathrm{P}(\theta, c, H)$, for $H = H_{fix} + H_{ord}$.

Dirichlet process

While the beta process has been used as a prior in featural models, the Dirichlet process is the classic Bayesian nonparametric prior for clustering models (Ferguson, 1973; S. N. MacEachern and Müller, 1998; McCloskey, 1965; Neal, 2000; West, 1992). The Dirichlet process itself is not a CRM; its atom weights, which represent cluster frequencies, must sum to one and are therefore correlated. But it can be obtained by normalizing the gamma process (Ferguson, 1973).

In particular, using facts about the Poisson process (Kingman, 1993), one can check that, when there are finitely many fixed atoms, we have $\tilde{G}(\Psi) < \infty$ a.s.; that is, the total mass of the gamma process is almost surely finite despite having infinitely many atoms from the ordinary component. Therefore, normalizing the process by dividing its weights by its total mass is well-defined. We thus can define a *Dirichlet process* as

$$G = \sum_k g_k \delta_{\psi_k} \triangleq \tilde{G}/\tilde{G}(\Psi),$$

where $\tilde{G} \sim \Gamma\mathrm{P}(\theta, 1, H)$, and where there are two parameters: a *concentration parameter* θ and a *base measure* H with finitely many fixed atoms. Note that while we have chosen the scale parameter $c = 1$ in this construction, the choice is in fact arbitrary for $c > 0$ and does not affect the G distribution (Eq. 4.15 and p. 83 of Pitman (2006)).

From this construction, we see immediately that the Dirichlet process is almost surely atomic, a property inherited from the gamma process. Moreover, not only are the weights of the Dirichlet process all contained in $(0, 1)$ but they further sum to one. Thus, the Dirichlet process may be seen as providing a probability distribution on a countable set. In particular, this countable set is often viewed as a countable number of clusters, with cluster parameters ψ_k.

4.3 Conjugacy and combinatorial clustering

In Section 4.2, we introduced CRMs and showed how a number of classical Bayesian nonparametric priors can be derived from CRMs. These priors provide infinite-dimensional vectors of real values, which can be interpreted as feature frequencies, feature rates, or cluster frequencies. To flesh out such interpretations we need to couple these real-valued processes with discrete-valued processes that capture combinatorial structure. In particular, viewing the weights of the beta process as feature frequencies, it is natural to consider binomial and negative binomial models that transform these frequencies into binary values or nonnegative integer counts. In this section we describe stochastic processes that achieve such transformations, again relying on the CRM framework.

The use of a Bernoulli likelihood whose frequency parameter is obtained from the weights of the beta process has been explored in the context of survival models by Hjort (1990) and Kim (1999a) and in the context of feature modeling by Thibaux and Jordan (2007). After reviewing the latter construction, we discuss a similar construction based on the negative binomial process. Moreover, recalling that Thibaux and Jordan (2007), building on work of Hjort (1990) and Kim (1999a), have shown that the Bernoulli likelihood is conjugate to the beta process, we demonstrate an analogous conjugacy result for the negative binomial process.

Bernoulli process

One way to make use of the beta process is to couple it to a *Bernoulli process* (Thibaux and Jordan, 2007). The Bernoulli process, denoted $\mathrm{BeP}(\tilde{H})$, has a single parameter, a *base measure* $\tilde{H}$; $\tilde{H}$ is any discrete measure with atom weights in $(0, 1]$. Although our focus will be on models in which $\tilde{H}$ is a draw from a beta process, as a matter of the general definition of the Bernoulli process the base measure $\tilde{H}$ need not be a CRM or even random—just as the Poisson distribution is defined relative to a parameter that may or may not be random in general but which is sometimes given a gamma distribution prior. Since $\tilde{H}$ is discrete by assumption, we may write

$$\tilde{H} = \sum_{k=1}^{\infty} b_k \delta_{\psi_k} \tag{4.5}$$

with $b_k \in (0, 1]$. We say that the random measure I is drawn from a Bernoulli process, $I \sim \mathrm{BeP}(\tilde{H})$, if $I = \sum_{k=1}^{\infty} i_k \delta_{\psi_k}$ with $i_k \overset{indep}{\sim} \mathrm{Bern}(b_k)$ for $k = 1, 2, \ldots$. That is, to form the Bernoulli process, we simply make a Bernoulli random variable draw for every one of the (potentially countable) atoms of the base measure. This definition of the Bernoulli process was proposed by Thibaux and Jordan (2007); it differs from a precursor introduced by Hjort (1990) in the context of survival analysis.

One interpretation for this construction is that the atoms of the base measure $\tilde{H}$ represent potential features of an individual, with feature frequencies equal to the atom weights and feature characteristics defined by the atom locations. The Bernoulli process draw can be viewed as characterizing the individual by the set of features that have weights equal to one. Suppose $\tilde{H}$ is derived from a Poisson process as the ordinary component of a completely random measure and has finite mass; then the number of features exhibited by the Bernoulli process, i.e. the total mass of the Bernoulli process draw, is a.s. finite. Thus the Bernoulli process can be viewed as providing a Bayesian nonparametric model of sparse binary feature vectors.

Now suppose that the base measure parameter is a draw from a beta process with parameters $\theta > 0$, $\gamma > 0$, and base measure H. That is, $B \sim \mathrm{BP}(\theta, \gamma, H)$ and $I \sim \mathrm{BeP}(B)$. We refer to the overall process as the *beta-Bernoulli process* (BBeP). Suppose that the beta process B has a finite number of fixed atoms. Then we note that the finite mass of the ordinary component of B implies that I has support on a finite set. That is, even though B has a countable infinity of atoms, I has only a finite number of atoms. This observation is important since, in any practical model, we will want an individual to exhibit only finitely many features.

Hjort (1990) and Kim (1999a) originally established that the posterior distribution of B under a constrained form of the BBeP was also a beta process with known parameters. Thibaux and Jordan (2007) went on to extend this analysis to the full BBeP. We cite the result by Thibaux and Jordan, 2007 here, using the completely random measure notation established above.

Theorem 4.3.1 (The beta process prior is conjugate to the Bernoulli process likelihood)**.** *Let H be a measure with atomic component $H_{fix} = \sum_{l=1}^{L} \rho_l \delta_{u_l}$ and continuous component H_{ord}. Let θ and γ be strictly positive scalars. Consider N conditionally-independent draws from the Bernoulli process: $I_n = \sum_{l=1}^{L} i_{fix,n,l} \delta_{u_l} + \sum_{j=1}^{J} i_{ord,n,j} \delta_{v_j} \overset{iid}{\sim} \mathrm{BeP}(B)$, for $n = 1, \ldots, N$ with $B \sim \mathrm{BP}(\theta, \gamma, H)$. That is, the Bernoulli process draws have J atoms that are not located at the atoms of H_{fix}. Then, $B | I_1, \ldots, I_N \sim \mathrm{BP}(\theta_{post}, \gamma_{post}, H_{post})$ with $\theta_{post} = \theta + N$, $\gamma_{post} = \gamma \frac{\theta}{\theta + N}$, and $H_{post,ord} = H_{ord}$. Further, $H_{post,fix} = \sum_{l=1}^{L} \rho_{post,l} \delta_{u_l} + \sum_{j=1}^{J} \xi_{post,j} \delta_{v_j}$, where $\rho_{post,l} = \rho_l + (\theta_{post} \gamma_{post})^{-1} \sum_{n=1}^{N} i_{fix,n,l}$ and $\xi_{post,j} = (\theta_{post} \gamma_{post})^{-1} \sum_{n=1}^{N} i_{ord,n,j}$.*

Note that the posterior beta-distributed fixed atoms are well-defined since $\xi_{post,j} > 0$ follows from $\sum_{n=1}^{N} i_{ord,n,j} > 0$, which holds by construction. As shown by Thibaux and Jordan (2007), if the underlying beta process is integrated out in the BBeP, we recover the *Indian buffet process* of Griffiths and Ghahramani, 2006.

Since the RBP and BP differ only in the fixed atoms, where conjugacy reduces to the finite-dimensional case, Theorem 4.3.1 immediately implies the following.

Corollary 4.3.2 (The RBP prior is conjugate to the Bernoulli process likelihood)**.**
Assume the conditions of Theorem 4.3.1, and consider N conditionally-independent Bernoulli process draws: $I_n = \sum_{l=1}^{L} i_{fix,n,l} \delta_{u_l} + \sum_{j=1}^{J} i_{ord,n,j} \delta_{v_j} \overset{iid}{\sim} \mathrm{BeP}(B)$, for $n = 1, \ldots, N$ with $B \sim \mathrm{RBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$ and $\{\rho_l\}_{l=1}^{L}$ and $\{\sigma_l\}_{l=1}^{L}$ strictly positive scalars. Then, $B | I_1, \ldots, I_N \sim \mathrm{RBP}(\theta_{post}, \gamma_{post}, \mathbf{u}_{post}, \boldsymbol{\rho}_{post}, \boldsymbol{\sigma}_{post}, H_{post,ord})$, for $\theta_{post} = \theta + N$, $\gamma_{post} = \gamma \frac{\theta}{\theta + N}$, $H_{post,ord} = H_{ord}$, and $L + J$ fixed atoms, $\{u_{post,l'}\} = \{u_l\}_{l=1}^{L} \cup \{v_j\}_{j=1}^{J}$. The $\boldsymbol{\rho}_{post}$ and $\boldsymbol{\sigma}_{post}$ parameters satisfy $\rho_{post,l} = \rho_l + \sum_{n=1}^{N} i_{fix,n,l}$ and $\sigma_{post,l} = \sigma_l + N - \sum_{n=1}^{N} i_{fix,n,l}$ for $l \in \{1, \ldots, L\}$ and $\rho_{post,L+j} = \sum_{n=1}^{N} i_{ord,n,j}$ and $\sigma_{post,L+j} = \theta + N - \sum_{n=1}^{N} i_{ord,n,j}$ for $j \in \{1, \ldots, J\}$.

The usefulness of the RBP becomes apparent in the posterior parameterization; the distributions associated with the fixed atoms more closely mirror the classical parametric conjugacy between the Bernoulli distribution and the beta distribution. This is an issue of convenience in the case of the BBeP, but it is more significant in the case of the negative binomial process, as we show in the following section, where conjugacy is preserved only in the RBP case (and not for the traditional, more constrained BP).

Negative binomial process

The Bernoulli distribution is not the only distribution that yields conjugacy when coupled to the beta distribution in the classical parametric setting; conjugacy holds for the negative binomial distribution as well. As we show in this section, this result can be extended to stochastic processes via the CRM framework.

We define the *negative binomial process* as a CRM with two parameters: a shape parameter $r > 0$ and a discrete base measure $\tilde{H} = \sum_k b_k \delta_{\psi_k}$ whose weights b_k take values in $(0, 1]$. As in the case of the Bernoulli process, $\tilde{H}$ need not be random at this point. Since $\tilde{H}$ is discrete, we again have a representation for $\tilde{H}$ as in Eq. (4.5), and we say that the random measure I is drawn from a negative binomial process, $I \sim \mathrm{NBP}(r, \tilde{H})$, if $I = \sum_{k=1}^{\infty} i_k \delta_{\psi_k}$ with $i_k \overset{indep}{\sim} \mathrm{NegBin}(r, b_k)$ for $k = 1, 2, \ldots$. That is, the negative binomial process is formed by simply making a single draw from a negative binomial distribution at each of the (potentially countably infinite) atoms of $\tilde{H}$. This construction generalizes the geometric process studied by Thibaux (2008).

As a Bernoulli process draw can be interpreted as assigning a set of features to a data point, so can we interpret a draw from the negative binomial process as assigning a set of feature counts to a data point. In particular, as for the Bernoulli process, we assume that each data point has its own draw from the negative binomial process. Every atom with strictly positive mass in this draw corresponds to a feature that is exhibited by this data point. Moreover, the size of the atom, which is a positive integer by construction, dictates how many times the feature is exhibited by the data point. For example, if the data point is a document, and each feature represents a particular word, then the negative binomial process draw would tell us how many occurrences of each word there are in the document.

If the base measure for a negative binomial process is a beta process, we say that the combined process is a *beta-negative binomial process* (BNBP). If the base measure is a three-parameter beta process, we say that the combined process is a *three-parameter beta-negative binomial process* (3BNBP). When either the BP or 3BP has a finite number of fixed atoms, the ordinary component of the BP or 3BP still has an infinite number of atoms, but the number of atoms in the negative binomial process is a.s. finite. We prove this fact and more in Section 4.5.

We now suppose that the base measure for the negative binomial process is a draw B from an RBP with parameters $\theta > 0$, $\gamma > 0$, $\{u_l\}_{l=1}^{L}$, $\{\rho_l\}_{l=1}^{L}$, $\{\sigma_l\}_{l=1}^{L}$, and H_{ord}. The overall specification is $B \sim \mathrm{RBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$ and $I \sim \mathrm{NBP}(r, B)$. The following theorem characterizes the posterior distribution for this model. The proof is given in Appendix 4.E.

Theorem 4.3.3 (The RBP prior is conjugate to the negative binomial process likelihood)**.** *Let θ and γ be strictly positive scalars. Let $(u_1, \ldots, u_L) \in \Psi^L$. Let the members of $\{\rho_l\}_{l=1}^{L}$*

and $\{\sigma_l\}_{l=1}^L$ *be strictly positive scalars. Let* H_{ord} *be a continuous measure on* Ψ*. Consider the following model for* N *draws from a negative binomial process:* $I_n = \sum_{l=1}^L i_{fix,n,l}\delta_{u_l} + \sum_{j=1}^J i_{ord,n,j}\delta_{v_j} \overset{iid}{\sim} \mathrm{NBP}(B)$*, for* $n = 1, \ldots, N$ *with* $B \sim \mathrm{RBP}(\theta, \gamma, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\sigma}, H_{ord})$*. That is, the negative binomial process draws have* J *atoms that are not located at the atoms of* H_{fix}*. Then,* $B|I_1, \ldots, I_N \sim \mathrm{RBP}(\theta_{post}, \gamma_{post}, \mathbf{u}_{post}, \boldsymbol{\rho}_{post}, \boldsymbol{\sigma}_{post}, H_{post,ord})$ *for* $\theta_{post} = \theta + Nr$*,* $\gamma_{post} = \gamma\frac{\theta}{\theta+Nr}$*,* $H_{post,ord} = H_{ord}$*, and* $L + J$ *fixed atoms,* $\{u_{post,l}\} = \{u_l\}_{l=1}^L \cup \{v_j\}_{j=1}^J$*. The* $\boldsymbol{\rho}_{post}$ *and* $\boldsymbol{\sigma}_{post}$ *parameters satisfy* $\rho_{post,l} = \rho_l + \sum_{n=1}^N i_{fix,n,l}$ *and* $\sigma_{post,l} = \sigma_l + Nr$ *for* $l \in \{1, \ldots, L\}$ *and* $\rho_{post,L+j} = \sum_{n=1}^N i_{ord,n,j}$ *and* $\sigma_{post,L+j} = \theta + Nr$ *for* $j \in \{1, \ldots, J\}$*.*

For the posterior measure to be a BP, we must have $\rho_{post,k} + \sigma_{post,k} = \theta_{post}$ for all k, but this equality can fail to hold even when the prior is a BP. For instance, whenever there are new fixed atom locations in the posterior relative to the prior, this equality will fail. So the BP, by contrast to the RBP, is not conjugate to the negative binomial process likelihood.

4.4 Mixtures and admixtures

We now assemble the pieces that we have introduced and consider Bayesian nonparametric models of admixture. Recall that the basic idea of an admixture is that an individual (e.g., an organism, a document, or an image) can belong simultaneously to multiple classes. This can be represented by associating a binary-valued vector with each individual; the vector has value one in components corresponding to classes to which the individual belongs and zero in components corresponding to classes to which the individual does not belong. More generally, we wish to remove the restriction to binary values and consider a general notion of admixture in which an individual is represented by a nonnegative, integer-valued vector. We refer to such vectors as *feature vectors*, and view the components of such vectors as counts representing the number of times the corresponding feature is exhibited by a given individual. For example, a document may exhibit a given word zero or more times.

As we discussed in Section 4.1, the standard approach to modeling an admixture is to assume that there is an exchangeable set of data associated with each individual and to assume that these data are drawn from a finite mixture model with individual-specific mixing proportions. There is another way to view this process, however, that opens the door to a variety of extensions. Note that to draw a set of data from a mixture, we can first choose the number of data points to be associated with each mixture component (a vector of counts) and then draw the data point values independently from each selected mixture component. That is, we randomly draw nonnegative integers i_k for each mixture component (or *cluster*) k. Then, for each k and each $n = 1, \ldots, i_k$, we draw a data point $x_{k,n} \sim F(\psi_k)$, where ψ_k is the parameter associated with mixture component k. The overall collection of data for this individual is $\{x_{k,n}\}_{k,n}$, with $N = \sum_k i_k$ total points. One way to generate data according to this decomposition is to make use of the NBP. We draw $I = \sum_k i_k \delta_{\psi_k} \sim \mathrm{NBP}(r, B)$, where B is drawn from a beta process, $B \sim \mathrm{BP}(\theta, \gamma, H)$. The overall model is a BNBP mixture model for the counts, coupled to a conditionally independent set of draws for the individual's data points $\{x_{k,n}\}_{k,n}$.

An alternative approach in the same spirit is to make use of a gamma process (to obtain a set of rates) that is coupled to a Poisson likelihood process (PLP)[3] to convert the rates into counts (Titsias, 2008). In particular, given a base measure $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k}$, let $I \sim \text{PLP}(\tilde{G})$ denote $I = \sum_k i_k \delta_{\psi_k}$, with $i_k \sim \text{Poisson}(\tilde{g}_k)$. We then consider a *gamma Poisson likelihood process* (ΓPLP) as follows: $\tilde{G} \sim \Gamma\text{P}(\theta, c, H)$, $I = \sum_k i_k \delta_{\psi_k} \sim \text{PLP}(\tilde{G})$, and $x_{k,n} \sim F(\psi_k)$, for $n = 1, \ldots, i_k$ and each k.

Both the BNBP approach and the ΓPLP approach deliver a random measure, $I = \sum_k i_k \delta_{\psi_k}$, as a representation of an admixed individual.[4] While the atom locations, (ψ_k), are subsequently used to generate data points, the pattern of admixture inheres in the vector of weights (i_k). It is thus natural to view this vector as the representation of an admixed individual. Indeed, in some problems such a weight vector might itself be the observed data. In other problems, the weights may be used to generate data in some more complex way that does not simply involve conditionally i.i.d. draws.

This perspective on admixture—focusing on the vector of weights (i_k) rather than the data associated with an individual—is also natural when we consider multiple individuals. The main issue becomes that of linking these vectors among multiple individuals; this linking can readily be achieved in the Bayesian formalism via a hierarchical model. In the remainder of this section we consider examples of such hierarchies in the Bayesian nonparametric setting.

Let us first consider the standard approach to admixture in which an individual is represented by a set of draws from a mixture model. For each individual we need to draw a set of mixing proportions, and these mixing proportions need to be coupled among the individuals. This can be achieved via a prior known as the *hierarchical Dirichlet process* (HDP) (Teh, Jordan, et al., 2006):

$$\begin{aligned} &G_0 \sim \text{DP}(\theta, H) \\ &G_d = \sum_k g_{d,k} \delta_{\psi_k} \overset{indep}{\sim} \text{DP}(\theta_d, G_0), \quad d = 1, 2, \ldots, \end{aligned}$$

where the index d ranges over the individuals. Note that the global measure G_0 is a discrete random probability measure, given that it is drawn from a Dirichlet process. In drawing the individual-specific random measure G_d at the second level, we therefore resample from among the atoms of G_0 and do so according to the weights of these atoms in G_0. This shares atoms among the individuals and couples the individual-specific mixing proportions $g_{d,k}$. We complete the model specification as follows:

$$z_{d,n} \overset{iid}{\sim} (g_{d,k})_k \quad \text{for } n = 1, \ldots, N_d$$

[3] We use the terminology "Poisson likelihood process" to distinguish a particular process with Poisson distributions affixed to each atom of some base distribution from the more general Poisson point process of Kingman (1993).

[4] We elaborate on the parallels and deep connections between the BNBP and ΓPLP in Appendix 4.A.

$$x_{d,n} \overset{indep}{\sim} F(\psi_{z_{d,n}}),$$

which draws an index $z_{d,n}$ from the discrete distribution $(g_{d,k})_k$ and then draws a data point $x_{d,n}$ from a distribution indexed by $z_{d,n}$. For instance, $(g_{d,k})$ might represent topic proportions in document d; $\psi_{z_{d,n}}$ might represent a topic, i.e. a distribution over words; and $x_{d,n}$ might represent the nth word in the dth document.

In the HDP, N_d is known for each d and is part of the model specification. We propose to instead take the featural approach as follows; we draw an individual-specific set of counts from an appropriate stochastic process and then generate the appropriate number of data points for each individual. Then the number of data points for each individual is itself a random variable and potentially coupled across individuals. In particular, one might consider the following conditional independence hierarchy involving the NBP:

$$\begin{aligned} B_0 &\sim \mathrm{BP}(\theta, \gamma, H) \\ I_d &= \sum_k i_{d,k} \delta_{\psi_k} \overset{indep}{\sim} \mathrm{NBP}(r_d, B_0), \end{aligned} \tag{4.6}$$

where we first draw a random measure B_0 from the beta process and then draw multiple times from an NBP with base measure given by B_0.

Although this conditional independence hierarchy does couple count vectors across multiple individuals, it uses a single collection of mixing proportions, the atom weights of B_0, for all individuals. By contrast, the HDP draws individual-specific mixing proportions from an underlying set of population-wide mixing proportions—and then converts these mixing proportions into counts. We can model individual-specific, but coupled, mixing proportions within an NBP-based framework by simply extending the hierarchy by one level:

$$\begin{aligned} B_0 &\sim \mathrm{BP}(\theta, \gamma, H) \\ B_d &\overset{indep}{\sim} \mathrm{BP}(\theta_d, \gamma_d, B_0 / B_0(\Psi)) \\ I_d &= \sum_k i_{d,k} \delta_{\psi_k} \overset{indep}{\sim} \mathrm{NBP}(r_d, B_d). \end{aligned} \tag{4.7}$$

Since B_0 is almost surely an atomic measure, the atoms of each B_d will coincide with those of B_0 almost surely. The weights associated with these atoms can be viewed as individual-specific feature probability vectors. We refer to this prior as the *hierarchical beta-negative binomial process* (HBNBP).

We also note that it is possible to consider additional levels of structure in which a population is decomposed into subpopulations and further decomposed into subsubpopulations and so on, bottoming out in a set of individuals. This tree structure can be captured by repeated draws from a set of beta processes at each level of the tree, conditioning on the beta process at the next highest level of the tree. Hierarchies of this form have previously been explored for beta-Bernoulli processes by Thibaux and Jordan (2007).

Comparison with Zhou et al. (2012). Zhou et al. (2012) have independently proposed a (non-hierarchical) beta-negative binomial process prior

$$\begin{aligned} B_0 &= \sum_k b_k \delta_{r_k,\psi_k} \sim \mathrm{BP}(\theta, \gamma, R \times H) \\ I_d &= \sum_k i_{d,k} \delta_{\psi_k} \quad \text{where} \quad i_{d,k} \overset{indep}{\sim} \mathrm{NegBin}(r_k, b_k), \end{aligned}$$

where R is a continuous finite measure over $\mathbb{R}^+$ used to associate a distinct failure parameter r_k with each beta process atom. Note that each individual is restricted to use the same failure parameters and the same beta process weights under this model. In contrast, our BNBP formulation (4.6) offers the flexibility of differentiating individuals by assigning each its own failure parameter r_d. Our HBNBP formulation (4.7) further introduces heterogeneity in the individual-specific beta process weights by leveraging the hierarchical beta process. We will see that these modeling choices are particularly well-suited for admixture modeling in the coming sections.

Zhou et al. (2012) use their prior to develop a Poisson factor analysis model for integer matrix factorization, while our primary motivation is mixture and admixture modeling. Our differing models and motivating applications have led to different challenges and algorithms for posterior inference. While Zhou et al. (2012) develop an inexact inference scheme based on a finite approximation to the beta process, we develop both an exact Markov chain Monte Carlo sampler and a finite approximation sampler for posterior inference under the HBNBP (see Section 4.7). Finally, unlike Zhou et al. (2012), we provide an extensive theoretical analysis of our priors including a proof of the conjugacy of the full beta process and the NBP (given in Section 4.3) and an asymptotic analysis of the BNBP (see Section 4.5).

4.5 Asymptotics

An important component of choosing a Bayesian prior is verifying that its behavior aligns with our beliefs about the behavior of the data-generating mechanism. In models of clustering, a particular measure of interest is the *diversity*—the dependence of the number of clusters on the number of data points. In speaking of the diversity, we typically assume a finite number of fixed atoms in a process derived from a CRM, so that asymptotic behavior is dominated by the ordinary component.

It has been observed in a variety of different contexts that the number of clusters in a dataset grows as a *power law* of the size of the data; that is, the number of clusters is asymptotically proportional to the number of data points raised to some positive power (Gnedin, Hansen, and Pitman, 2007). Real-world examples of such behavior are provided by M. E. Newman (2005) and Mitzenmacher (2004).

The diversity has been characterized for the Dirichlet process (DP) and a two-parameter extension to the Dirichlet process known as the *Pitman-Yor process* (PYP) (Pitman and Yor, 1997), with extra parameter $\alpha \in (0, 1)$ and concentration parameter $\theta > -\alpha$. We will see that while the number of clusters generated according to a DP grows as a logarithm of the size of the data, the number of clusters generated according to a PYP grows as a power of the size of the data. Indeed, the popularity of the Pitman-Yor process—as an alternative prior to the Dirichlet process in the clustering domain—can be attributed to this power-law growth (Goldwater, Griffiths, and M. Johnson, 2006; Teh, 2006; Wood et al., 2009). In this section, we derive analogous asymptotic results for the BNBP treated as a clustering model.

We first highlight a subtle difference between our model and the Dirichlet process. For a Dirichlet process, the number of data points N is known a priori and fixed. An advantage of our model is that it models the number of data points N as a random variable and therefore has potentially more predictive power in modeling multiple populations. We note that a similar effect can be achieved for the Dirichlet process by using the gamma process for feature modeling as described in Section 4.4 rather than normalizing away the mass that determines the number of observations. However, there is no such unnormalized completely random measure for the PYP (Pitman and Yor, 1997). We thus treat N as fixed for the DP and PYP, in which case the number of clusters $K(N)$ is a function of N. On the other hand, the number of data points $N(r)$ depends on r in the case of the BNBP, and the number of clusters $K(r)$ does as well. We also define $K_j(N)$ to be the number of clusters with exactly j elements in the case of the DP and PYP, and we define $K_j(r)$ to be the number of clusters with exactly j elements in the BNBP case.

For the DP and PYP, $K(N)$ and $K_j(N)$ are random even though N is fixed, so it will be useful to also define their expectations:

$$\Phi(N) \triangleq \mathbb{E}[K(N)], \quad \Phi_j(N) \triangleq \mathbb{E}[K_j(N)]. \tag{4.8}$$

In the BNBP and 3BNBP cases, all of $K(r)$, $K_j(r)$, and $N(r)$ are random. So we further define

$$\Phi(r) \triangleq \mathbb{E}[K(r)], \quad \Phi_j(r) \triangleq \mathbb{E}[K_j(r)], \quad \xi(r) \triangleq \mathbb{E}[N(r)]. \tag{4.9}$$

We summarize the results that we establish in this section in Table 4.1, where we also include comparisons to existing results for the DP and PYP.[5] The full statements of our results, from which the table is derived, can be found in Appendix 4.C, and proofs are given in Appendix 4.D.

The table shows, for example, that for the DP, $\Phi(N) \sim \theta \log(N)$ as $N \to \infty$, and, for the BNBP, $\Phi_j(r) \sim \gamma\theta j^{-1}$ as $r \to \infty$ (i.e., constant in r). The result for the expected number of clusters for the DP can be found in Korwar and Hollander (1973); results for the expected number of clusters for both the DP and PYP can be found in Pitman (2006, Eq. 3.24 on p. 69 and Eq. 3.47 on p. 73). Note that in all cases the expected counts of clusters of size j are asymptotic expansions in terms of r for fixed j and should not be interpreted as asymptotic expansions in terms of j.

[5] The reader interested in power laws may also note that the generalized gamma process is a completely random measure that, when normalized, provides a probability measure for clusters that has asymptotic behavior similar to the PYP; in particular, the expected number of clusters grows almost surely as a power of the size of the data (Lijoi, Mena, and Prünster, 2007).

Table 4.1: Let N be the number of data points when this number is fixed and $\xi(r)$ be the expected number of data points when N is random. Let $\Phi(N)$, $\Phi_j(N)$, $\Phi(r)$, and $\Phi_j(r)$ be the expected number of clusters under various scenarios and defined as in Eqs. (4.8) and (4.9). The upper part of the table gives the asymptotic behavior of Φ up to a multiplicative constant, and the bottom part of the table gives the multiplicative constants. For the DP, $\theta > 0$. For the PYP, $\alpha \in (0,1)$ and $\theta > -\alpha$. For the BNBP, $\theta > 1$. For the 3BNBP, $\alpha \in (0,1)$ and $\theta > 1 - \alpha$.

Process	Expected number of clusters	Expected number of clusters of size j
	Function of N or $\xi(r)$	
DP	$\log(N)$	1
PYP	N^{α}	N^{α}
BNBP	$\log(\xi(r))$	1
3BNBP	$(\xi(r))^{\alpha}$	$(\xi(r))^{\alpha}$
	Constants	
DP	θ	θj^{-1}
PYP	$\frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}$	$\frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)}$
BNBP	$\gamma\theta$	$\gamma\theta j^{-1}$
3BNBP	$\frac{\gamma^{1-\alpha}}{\alpha} \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \left(\frac{\theta+\alpha-1}{\theta}\right)^{\alpha}$	$\gamma^{1-\alpha} \frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)} \left(\frac{\theta+\alpha-1}{\theta}\right)^{\alpha}$

We conclude that, just as for the Dirichlet process, the BNBP can achieve both logarithmic cluster number growth in the basic model and power law cluster number growth in the expanded, three-parameter model.

4.6 Simulation

Our theoretical results in Section 4.5 are supported by simulation results, summarized in Figure 4.1; in particular, our simulation corroborated the existence of power laws in the three-parameter beta process case examined in Section 4.5. The simulation was performed as follows. For values of the negative binomial parameter r evenly spaced between 1 and 1,001, we generated beta process weights according to a beta process (or three-parameter beta process) using a stick-breaking representation (Paisley, Zaas, et al., 2010; Broderick, Jordan, and Pitman, 2012). For each of the resulting atoms, we simulated negative binomial draws to arrive at a sample from a BNBP. For each such BNBP, we can count the resulting total number of data points N and total number of clusters K. Thus, each r gives us an (r, N, K) triple.

In the simulation, we set the mass parameter $\gamma = 3$. We set the concentration parameter $\theta = 3$; in particular, we note that the analysis in Section 4.5 implies that we should always have $\theta > 1$. Finally, we ran the simulation for both the $\alpha = 0$ case, where we expect no power law behavior, and the $\alpha = 0.5$ case, where we do expect power law behavior. The results are shown in Figure 4.1. Is this figure, we scatter plot the (r, K) tuples from the generated (r, N, K) triples on the left and plot the (N, K) tuples on the right.

In the left plot, the upper black points represent the simulation with $\alpha = 0.5$, and the lower blue data points represent the $\alpha = 0$ case. The lower red line illustrates the asymptotic theoretical result corresponding to the $\alpha = 0$ case (Lemma 4.C.6 in Appendix 4.C), and we can see that the anticipated logarithmic growth behavior agrees with our simulation. The upper red line illustrates the theoretical result for the $\alpha = 0.5$ case (Lemma 4.C.7 in Appendix 4.C). The agreement between simulation and theory here demonstrates that, in contrast to the $\alpha = 0$ case, the $\alpha = 0.5$ case exhibits power law growth in the number of clusters K as a function of the negative binomial parameter r.

Our simulations also bear out that the expectation of the random number of data points N increases linearly with r (Lemmas 4.C.4 and 4.C.5 in Appendix 4.C). We see, then, on the right side of Figure 4.1 the behavior of the number of clusters K now plotted as a function of N. As expected given the asymptotics of the expected value of N, the behavior in the right plot largely mirrors the behavior in the left plot. Just as in the left plot, the lower red line (Theorem 4.C.10 in Appendix 4.C) shows the anticipated logarithmic growth of K and N when $\alpha = 0$. And the upper red line (Theorem 4.C.11 in Appendix 4.C) shows the anticipated power law growth of K and N when $\alpha = 0.5$.

We can see the parallels with the DP and PYP here. Clusters generated from the Dirichlet process (i.e., Pitman-Yor process with $\alpha = 0$) exhibit logarithmic growth of the expected number of clusters K as the (deterministic) number of data points N grows. And clusters generated from the Pitman-Yor process with $\alpha \in (0, 1)$ exhibit power law behavior in the expectation of K as a function of (fixed) N. So too do we see that the BNBP, when applied to clustering problems, yields asymptotic growth similar to the DP and that the 3BNBP yields asymptotic growth similar to the PYP.

4.7 Posterior inference

In this section we present posterior inference algorithms for the HBNBP. We focus on the setting in which, for each individual d, there is an associated exchangeable sequence of observations $(x_{d,n})_{n=1}^{N_d}$. We seek to infer both the admixture component responsible for each observation and the parameter ψ_k associated with each component. Hereafter, we let $z_{d,n}$ denote the unknown component index associated with $x_{d,n}$, so that $x_{d,n} \sim F(\psi_{z_{d,n}})$.

Under the HBNBP admixture model introduced in Section 4.4, the posterior over component indices and parameters has the form

$$p(\mathbf{z}_{\cdot,\cdot}, \boldsymbol{\psi}_{\cdot} \mid \mathbf{x}_{\cdot,\cdot}, \Theta) \propto p(\mathbf{z}_{\cdot,\cdot}, \boldsymbol{\psi}_{\cdot}, \mathbf{b}_{0,\cdot}, \mathbf{b}_{\cdot,\cdot} \mid \mathbf{x}_{\cdot,\cdot}, \Theta),$$

where $\Theta \triangleq (F, H, \gamma_0, \theta_0, \boldsymbol{\gamma}_{\cdot}, \boldsymbol{\theta}_{\cdot}, \mathbf{r}_{\cdot})$ is the collection of all fixed hyperparameters. As is the case with HDP admixtures (Teh, Jordan, et al., 2006) and earlier hierarchical beta process featural models (Thibaux and Jordan, 2007), the posterior of the HBNBP admixture cannot be obtained in analytical form due to complex couplings in the marginal $p(\mathbf{x}_{\cdot,\cdot} \mid \Theta)$. We

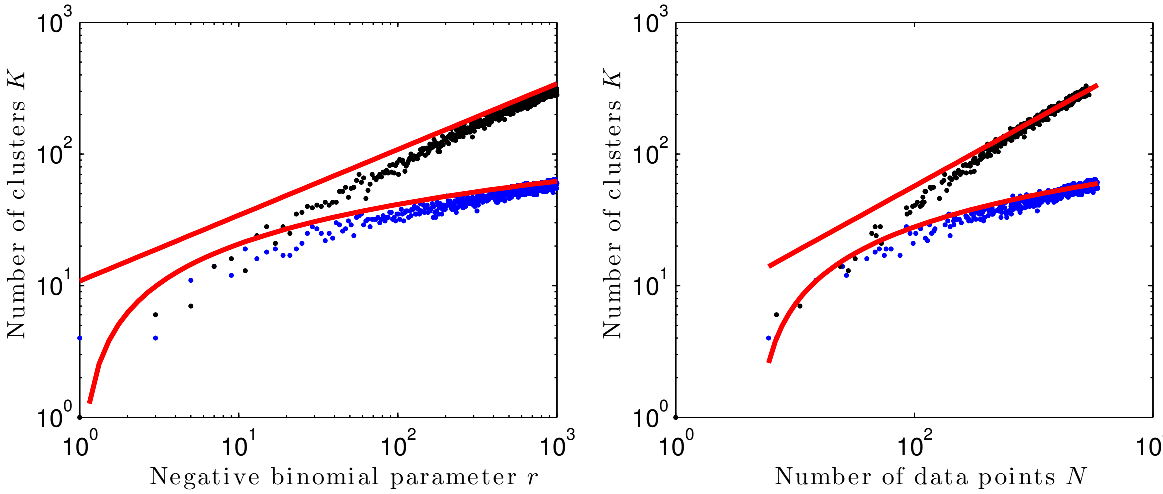


Figure 4.1: For each r evenly spaced between 1 and 1,001, we simulate (random) values of the number of data points N and number of clusters K from the BNBP and 3BNBP. In both plots, we have mass parameter $\gamma = 3$ and concentration parameter $\theta = 3$. On the *left*, we see the number of clusters K as a function of the negative binomial parameter r (see Lemma 4.C.6 and Lemma 4.C.7 in Appendix 4.C); on the *right*, we see the number of clusters K as a function of the (random) number of data points N (see Theorem 4.C.10 and Theorem 4.C.11 in Appendix 4.C). In both plots, the upper black points show simulation results for the case $\alpha = 0.5$, and the lower blue points show $\alpha = 0$. Red lines indicate the theoretical asymptotic mean behavior we expect from Section 4.5.

therefore develop Gibbs sampling algorithms (S. Geman and D. Geman, 1984) to draw samples of the relevant latent variables from their joint posterior.

A challenging aspect of inference in the nonparametric setting is the countable infinitude of component parameters and the countably infinite support of the component indices. We develop two sampling algorithms that cope with this issue in different ways. In Section 4.7, we use slice sampling to control the number of components that need be considered on a given round of sampling and thereby derive an exact Gibbs sampler for posterior inference under the HBNBP admixture model. In Section 4.7, we describe an efficient alternative sampler that makes use of a finite approximation to the beta process. Throughout we assume that the base measure H is continuous. We note that neither procedure requires conjugacy between the base distribution H and the data-generating distribution F.

Exact Gibbs slice sampler

Slice sampling (Damien, Wakefield, and Walker, 1999; Neal, 2003) has been successfully employed in several Bayesian nonparametric contexts, including Dirichlet process mixture modeling (Walker, 2007; Papaspiliopoulos, 2008; Kalli, Griffin, and Walker, 2011) and beta process feature modeling (Teh, Görür, and Ghahramani, 2007). The key to its success lies in the introduction of one or more auxiliary variables that serve as adaptive truncation levels for an infinite sum representation of the stochastic process.

This adaptive truncation procedure proceeds as follows. For each observation associated with individual d, we introduce an auxiliary variable $u_{d,n}$ with conditional distribution

$$u_{d,n} \sim \mathrm{Unif}(0, \zeta_{d,z_{d,n}}),$$

where $(\zeta_{d,k})_{k=1}^{\infty}$ is a fixed positive sequence with $\lim_{k\to\infty} \zeta_{d,k} = 0$. To sample the component indices, we recall that a negative binomial draw $i_{d,k} \sim \mathrm{NegBin}(r_d, b_{d,k})$ may be represented as a gamma-Poisson mixture:

$$\begin{aligned} \lambda_{d,k} &\sim \mathrm{Gamma}\left(r_d, \frac{1-b_{d,k}}{b_{d,k}}\right) \\ i_{d,k} &\sim \mathrm{Poisson}(\lambda_{d,k}). \end{aligned}$$

We first sample $\lambda_{d,k}$ from its full conditional. By gamma-Poisson conjugacy, this has the simple form

$$\lambda_{d,k} \sim \mathrm{Gamma}\left(r_d + i_{d,k}, 1/b_{d,k}\right).$$

We next note that, given $\boldsymbol{\lambda}_{d,\cdot}$ and the total number of observations associated with individual d, the cluster sizes $i_{d,k}$ may be constructed by sampling each $z_{d,n}$ independently from $\boldsymbol{\lambda}_{d,\cdot}/\sum_k \lambda_{d,k}$ and setting $i_{d,k} = \sum_n \mathbb{I}(z_{d,n} = k)$. Hence, conditioned on the number of data points N_d, the component parameters ψ_k, the auxiliary variables $\lambda_{d,k}$, and the slice-sampling variable $u_{d,n}$, we sample the index $z_{d,n}$ from a discrete distribution with

$$\mathbb{P}(z_{d,n} = k) \propto F(dx_{d,n} \mid \psi_k) \frac{\mathbb{I}(u_{d,n} \leq \zeta_{d,k})}{\zeta_{d,k}} \lambda_{d,k}$$

so that only the finite set of component indices $\{k : \zeta_{d,k} \geq u_{d,n}\}$ need be considered when sampling $z_{d,n}$.

Let $K_d \triangleq \max\{k : \exists n \text{ s.t. } \zeta_{d,k} \geq u_{d,n}\}$ and $K \triangleq \max_d K_d$. Then, on a given round of sampling, we need only explicitly represent $\lambda_{d,k}$ and $b_{d,k}$ for $k \leq K_d$ and ψ_k and $b_{0,k}$ for $k \leq K$. The simple Gibbs conditionals for $b_{d,k}$ and ψ_k can be found in Appendix 4.F. To sample the shared beta process weights $b_{0,k}$, we leverage the size-biased construction of the beta process introduced by Thibaux and Jordan, 2007:

$$B_0 = \sum_{m=0}^{\infty} \sum_{i=1}^{C_m} b_{0,m,i} \delta_{\psi_{m,i,\cdot}},$$

where

$$C_m \overset{indep}{\sim} \mathrm{Poisson}\left(\frac{\theta_0 \gamma_0}{\theta_0 + m}\right), \quad b_{0,m,i} \overset{indep}{\sim} \mathrm{Beta}(1, \theta_0 + m), \quad \text{and} \quad \psi_{m,i,\cdot} \overset{iid}{\sim} H,$$

and we develop a Gibbs slice sampler for generating samples from its posterior. The details are deferred to Appendix 4.F.

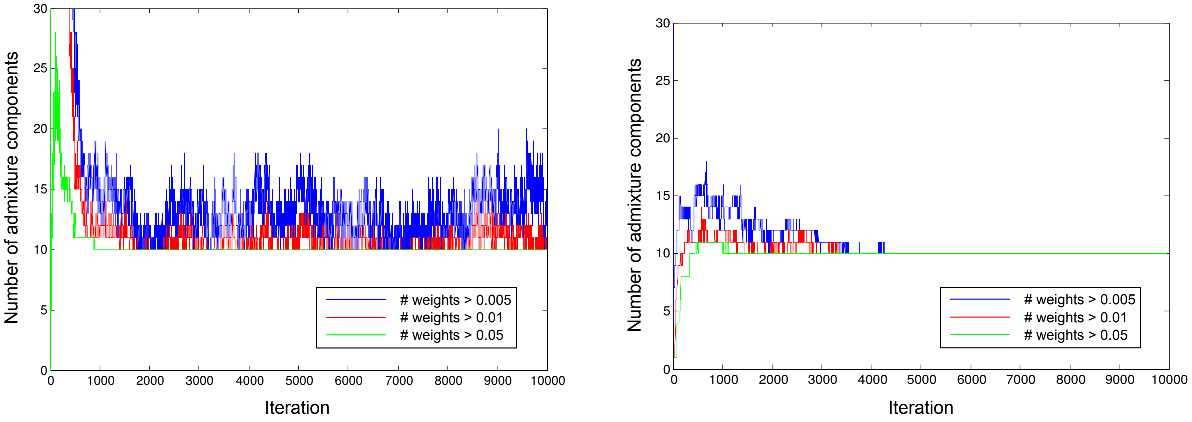


Figure 4.2: Number of admixture components used by the finite approximation sampler with $K = 100$ (*left*) and the exact Gibbs slice sampler (*right*) on each iteration of HBNBP admixture model posterior inference. We use a standard “toy bars” dataset with ten underlying admixture components (cf. Griffiths and Steyvers, 2004). We declare a component to be used by a sample if the sampled beta process weight, $b_{0,k}$, exceeds a small threshold. Both the exact and the finite approximation sampler find the correct underlying structure, while the finite sampler attempts to innovate more because of the larger number of proposal components available to the data in each iteration.

Finite approximation Gibbs sampler

An alternative to the size-biased construction of B_0 is a finite approximation to the beta process with a fixed number of components, K:

$$b_{0,k} \overset{iid}{\sim} \text{Beta}(\theta_0\gamma_0/K, \theta_0(1-\gamma_0/K)), \quad \psi_k \overset{iid}{\sim} H, \quad k \in \{1,\ldots,K\}. \tag{4.10}$$

It is known that, when H is continuous, the distribution of $\sum_{k=1}^{K} b_{0,k}\delta_{\psi_k}$ converges to $\text{BP}(\theta_0, \gamma_0, H)$ as the number of components $K \to \infty$ (see the proof of Theorem 3.1 by Hjort (1990) with the choice $A_0(t) = \gamma$). Hence, we may leverage the beta process approximation (4.10) to develop an approximate posterior sampler for the HBNBP admixture model with an approximation level K that trades off between computational efficiency and fidelity to the true posterior. We defer the detailed conditionals of the resulting Gibbs sampler to Appendix 4.F and briefly compare the behavior of the finite and exact samplers on a toy dataset in Figure Figure 4.2. We note finally that the beta process approximation in Eq. (4.10) also gives rise to a new finite admixture model that may be of interest in its own right; we explore the utility of this HBNBP approximation in Section 4.9.

4.8 Document topic modeling

In the next two sections, we show how the HBNBP admixture model and its finite approximation can be used as practical building blocks for more complex supervised and unsupervised inferential tasks.

We first consider the unsupervised task of *document topic modeling*, in which each individual d is a document containing N_d observations (words) and each word $x_{d,n}$ belongs to a vocabulary of size V. The topic modeling framework is an instance of admixture modeling in which we assume that each word of each document is generated from a latent admixture component or *topic*, and our goal is to infer the topic underlying each word.

In our experiments, we let H_{ord}, the Ψ dimension of the ordinary component intensity measure, be a Dirichlet distribution with parameter $\eta\mathbf{1}$ for $\eta = 0.1$ and $\mathbf{1}$ a V-dimensional vector of ones and let $F(\psi_k)$ be Multinomial$(1, \psi_k)$. We use the setting $(\gamma_0, \theta_0, \gamma_d, \theta_d) = (3, 3, 1, 10)$ for the global and document-specific mass and concentration parameters and set the document-specific negative binomial shape parameter according to the heuristic $r_d = N_d(\theta_0 - 1)/(\theta_0\gamma_0)$. We arrive at this heuristic by matching N_d to its expectation under a non-hierarchical BNBP model and solving for r_d:

$$\mathbb{E}[N_d] = r_d\mathbb{E}\left[\sum\nolimits_{k=1}^{\infty} b_{d,k}/(1 - b_{d,k})\right] = \gamma_0\theta_0/(\theta_0 - 1).$$

When applying the exact Gibbs slice sampler, we let the slice sampling decay sequence follow the same pattern across all documents: $\zeta_{d,k} = 1.5^{-k}$.

Worldwide Incidents Tracking System

We report results on the Worldwide Incidents Tracking System (WITS) dataset.[6] This dataset consists of reports on 79,754 terrorist attacks from the years 2004 through 2010. Each event contains a written summary of the incident, location information, victim statistics, and various binary fields such as “assassination,” “IED,” and “suicide.” We transformed each incident into a text document by concatenating the summary and location fields and then adding further words to account for other, categorical fields: e.g., an incident with seven hostages would have the word “hostage” added to the document seven times. We used a vocabulary size of $V = 1{,}048$ words.

Perpetrator Identification. Our experiment assesses the ability of the HBNBP admixture model to discriminate among incidents perpetrated by different organizations. We first grouped documents according to the organization claiming responsibility for the reported incident. We considered 5,390 claimed documents in total distributed across the ten organizations listed in Table 4.2. We removed all organization identifiers from all documents and randomly set aside 10% of the documents in each group as test data. Next, for each group, we trained an independent, organization-specific HBNBP model on the remaining documents in that group by drawing 10,000 MCMC samples. We proceeded to classify each

[6] `https://wits.nctc.gov`

Table 4.2: The number of incidents claimed by each organization in the WITS perpetrator identification experiment.

Group ID	Perpetrator	# Claimed Incidents
1	taliban	2647
2	al-aqsa	417
3	farc	76
4	izz al-din al-qassam	478
5	hizballah	89
6	al-shabaab al-islamiya	426
7	al-quds	505
8	abu ali mustafa	249
9	al-nasser salah al-din	212
10	communist party of nepal (maoist)	291

test document by measuring the likelihood of the document under each trained HBNBP model and assigning the label associated with the largest likelihood. The resulting confusion matrix across the ten candidate organizations is displayed in Table 4.3. Results are reported for the exact Gibbs slice sampler; performance under the finite approximation sampler is nearly identical.

For comparison, we carried out the same experiment using the more standard HDP admixture model in place of the HBNBP. For posterior inference, we used the HDP block sampler code of Yee Whye Teh[7] and initialized the sampler with 100 topics and topic hyperparameter $\eta = 0.1$ (all remaining parameters were set to their default values). For each organization, we drew 250,000 MCMC samples and kept every twenty-fifth sample for evaluation. The confusion matrix obtained through HDP modeling is displayed in Table 4.3. We see that, overall, HBNBP modeling leads to more accurate identification of perpetrators than its HDP counterpart. Most notably, the HDP wrongly attributes more than half of all documents from group 1 (taliban) to group 3 (farc) or group 6 (al-shabaab al-islamiya). We hypothesize that the HBNBP's superior discriminative power stems from its ability to distinguish between documents both on the basis of word frequency and on the basis of document length.

We would expect the HBNBP to have greatest difficulty discriminating among perpetrators when both word usage frequencies and document length distributions are similar across groups. To evaluate the extent to which this occurs in our perpetrator identification experiment, for each organization, we plotted the density histogram of document lengths in Figure 4.8 and the heat map displaying word usage frequency across all associated documents in Figure 4.8. We find that the word frequency patterns are nearly identical across groups 2, 7, 8, and 9 (al-aqsa, al-quds, abu ali mustafa, and al-nasser salah al-din, respectively) and that the document length distributions of these four groups are all well aligned. As expected,

[7] `http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/npbayes-r1.tgz`

Table 4.3: Confusion matrices for WITS perpetrator identification. See Table 4.2 for the organization names matching each group ID.

[HBNBP Confusion Matrix]

	Predicted Groups									
Actual Groups	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.38	0.00	0.02	0.00	0.00	0.29	0.29	0.02	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.54	0.00	0.00	0.15	0.27	0.04	0.00
5	0.11	0.33	0.00	0.11	0.44	0.00	0.00	0.00	0.00	0.00
6	0.02	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
7	0.00	0.10	0.00	0.06	0.02	0.00	0.48	0.30	0.04	0.00
8	0.00	0.04	0.00	0.00	0.00	0.00	0.16	0.76	0.04	0.00
9	0.00	0.10	0.00	0.05	0.10	0.00	0.29	0.43	0.05	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

[HDP Confusion Matrix]

	Predicted Groups									
Actual Groups	1	2	3	4	5	6	7	8	9	10
1	0.46	0.00	0.26	0.00	0.03	0.23	0.00	0.00	0.00	0.01
2	0.00	0.31	0.02	0.02	0.00	0.00	0.29	0.36	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.52	0.04	0.00	0.06	0.31	0.06	0.00
5	0.11	0.00	0.00	0.00	0.44	0.00	0.11	0.11	0.11	0.11
6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
7	0.00	0.10	0.00	0.04	0.00	0.00	0.38	0.42	0.06	0.00
8	0.00	0.04	0.00	0.00	0.00	0.00	0.08	0.84	0.04	0.00
9	0.00	0.05	0.00	0.10	0.00	0.00	0.24	0.62	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

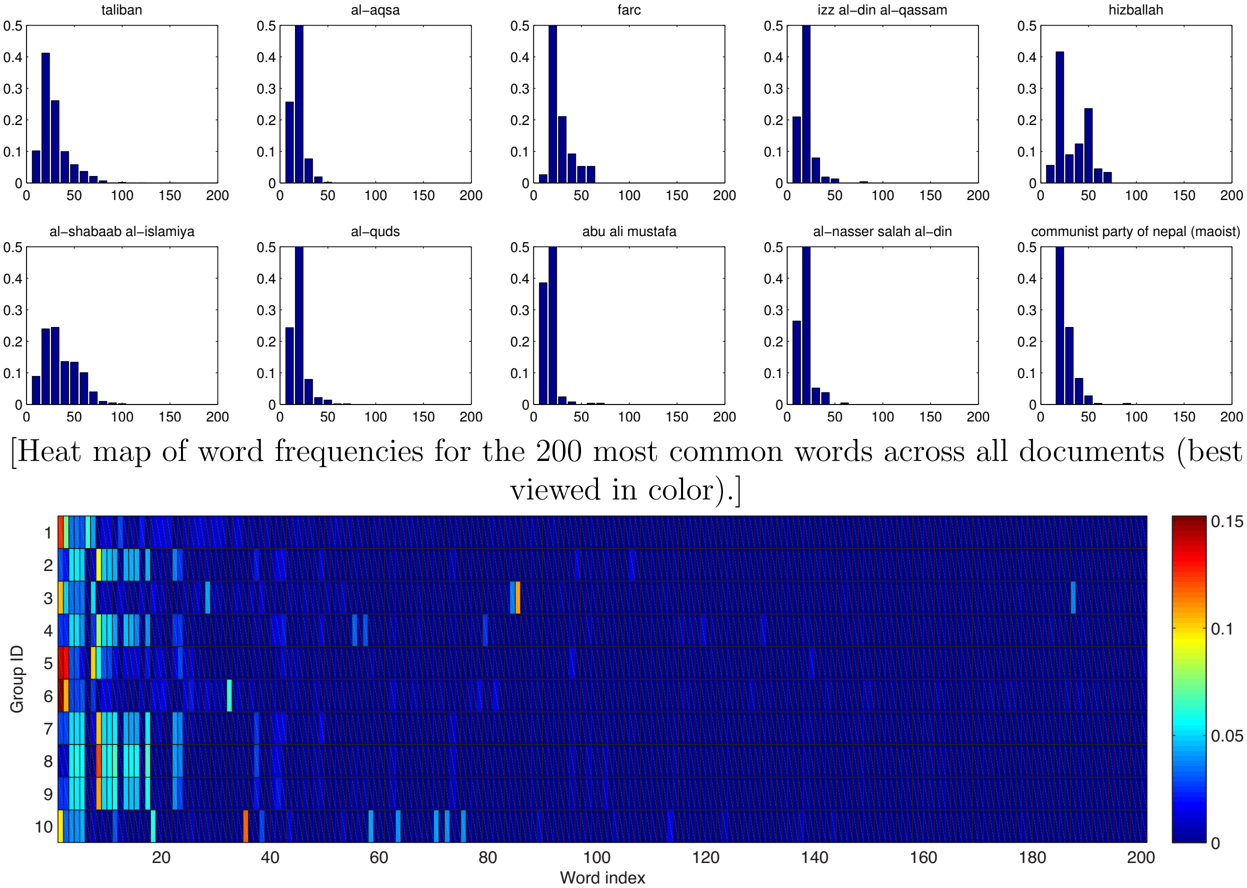
the majority of classification errors made by our HBNBP models result from misattribution among these same four groups. The same group similarity structure is evidenced in a display of the ten most probable words from the most probable HBNBP topic for each group, Table 4.4. There, we also find an intuitive summary of the salient regional and methodological vocabulary associated with each organization.

4.9 Image segmentation and object recognition

Two problems of enduring interest in the computer vision community are *image segmentation*, dividing an image into its distinct, semantically meaningful regions, and *object recognition*, labeling the regions of images according to their semantic object classes. Solutions to these problems are at the core of applications such as content-based image retrieval, video surveying, and object tracking. Here we will take an admixture modeling approach to jointly recognizing and localizing objects within images (Cao and F. Li, 2007; Russell et al., 2006; Sivic et al., 2005; Verbeek and Bill Triggs, 2007). Each individual d is an image comprised

Table 4.4: The ten most probable words from the most probable topic in the final MCMC sample of each group in the WITS perpetrator identification experiment. The topic probability is given in parentheses. See Table 4.2 for the organization names matching each group ID.

HBNBP: Top topic per organization	
group 1 (0.29)	afghanistan, assailants, claimed, responsibility, armedattack, fired, police, victims, armed, upon
group 2 (0.77)	israel, assailants, armedattack, responsibility, fired, claimed, district, causing, southern, damage
group 3 (0.95)	colombia, victims, facility, wounded, armed, claimed, forces, revolutionary, responsibility, assailants
group 4 (0.87)	israel, fired, responsibility, claimed, armedattack, causing, injuries, district, southern, assailants
group 5 (0.95)	victims, wounded, facility, israel, responsibility, claimed, armedattack, fired, rockets, katyusha
group 6 (0.54)	wounded, victims, somalia, civilians, wounding, facility, killing, mortars, armedattack, several
group 7 (0.83)	israel, district, southern, responsibility, claimed, fired, armedattack, assailants, causing, injuries
group 8 (0.94)	israel, district, southern, armedattack, claimed, fired, responsibility, assailants, causing, injuries
group 9 (0.88)	israel, district, southern, fired, responsibility, claimed, armedattack, assailants, causing, injuries
group 10 (0.80)	nepal, victims, hostage, assailants, party, communist, claimed, front, maoist/united, responsibility



[Density histograms of document lengths.]

[Heat map of word frequencies for the 200 most common words across all documents (best viewed in color).]

Figure 4.3: Document length distributions and word frequencies for each organization in the WITS perpetrator identification experiment.

of N_d image patches (observations), and each patch $\mathbf{x}_{d,n}$ is assumed to be generated by an unknown object class (a latent component of the admixture). Given a series of training images with image patches labeled, the problem of recognizing and localizing objects in a new image reduces to inferring the latent class associated with each new image patch. Since the number of object classes is typically known *a priori*, we will tackle this inferential task with the finite approximation to the HBNBP admixture model given in Section 4.7 and compare its performance with that of a more standard model of admixture, latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003).

Representing an Image Patch

We will represent each image patch as a vector of visual descriptors drawn from multiple modalities. Verbeek and Bill Triggs (2007) suggest three complementary modalities: texture, hue, and location. Here, we introduce a fourth: opponent angle. To describe hue, we use the robust hue descriptor of Van De Weijer and Schmid, 2006, which grants invariance to illuminant variations, lighting geometry, and specularities. For texture description we use "dense SIFT" features (Lowe, 2004; Dalal and B. Triggs, 2005), histograms of oriented gradients computed not at local keypoints but rather at a single scale over each patch. To describe coarse location, we cover each image with a regular c x c grid of cells (for a total of $V^{\text{loc}} = c^2$ cells) and assign each patch the index of the covering cell. The opponent angle descriptor of Van De Weijer and Schmid, 2006 captures a second characterization of image patch color. These features are invariant to specularities, illuminant variations, and diffuse lighting conditions.

To build a discrete visual vocabulary from these raw descriptors, we vector quantize the dense SIFT, hue, and opponent angle descriptors using k-means, producing V^{sift}, V^{hue}, and V^{opp} clusters respectively. Finally, we form the observation associated with a patch by concatenating the four modality components into a single vector, $\mathbf{x}_{d,n} = (x^{\text{sift}}_{d,n}, x^{\text{hue}}_{d,n}, x^{\text{loc}}_{d,n}, x^{\text{opp}}_{d,n})$. As in Verbeek and Bill Triggs, 2007, we assume that the descriptors from disparate modalities are conditionally independent given the latent object class of the patch. Hence, we define our data generating distribution and our base distribution over parameters $\boldsymbol{\psi}_k = (\psi^{\text{sift}}_k, \psi^{\text{hue}}_k, \psi^{\text{loc}}_k, \psi^{\text{opp}}_k)$ via

$$\begin{aligned} \psi^m_k &\overset{indep}{\sim} \text{Dirichlet}(\eta \mathbf{1}_{V^m}) && \text{for } m \in \{\text{sift}, \text{hue}, \text{loc}, \text{opp}\} \\ x^m_{d,n} \mid z_{d,n}, \boldsymbol{\psi}_{\cdot} &\overset{indep}{\sim} \text{Multinomial}(1, \psi^m_{z_{d,n}}) && \text{for } m \in \{\text{sift}, \text{hue}, \text{loc}, \text{opp}\} \end{aligned}$$

for a hyperparameter $\eta \in \mathbb{R}$ and $\mathbf{1}_{V^m}$ a V^m-dimensional vector of ones.

Experimental Setup

We use the Microsoft Research Cambridge pixel-wise labeled image database v1 in our experiments.[8] The dataset consists of 240 images, each of size 213 x 320 pixels. Each image

[8] http://research.microsoft.com/vision/cambridge/recognition/

has an associated pixel-wise ground truth labeling, with each pixel labeled as belonging to one of 13 semantic classes or to the *void* class. Pixels have a ground truth label of *void* when they do not belong to any semantic class or when they lie on the boundaries between classes in an image. The dataset provider notes that there are insufficiently many instances of *horse*, *mountain*, *sheep*, or *water* to learn these classes, so, as in Verbeek and Bill Triggs, 2007, we treat these ground truth labels as *void* as well. Thus, our general task is to learn and segment the remaining nine semantic object classes.

From each image, we extract 20 x 20 pixel patches spaced at 10 pixel intervals across the image. We choose the visual vocabulary sizes $(V^{\text{sift}}, V^{\text{hue}}, V^{\text{loc}}, V^{\text{opp}}) = (1000, 100, 100, 100)$ and fix the hyperparameter $\eta = 0.1$. As in Verbeek and Bill Triggs, 2007, we assign each patch a ground truth label $z_{d,n}$ representing the most frequent pixel label within the patch. When performing posterior inference, we divide the dataset into training and test images. We allow the inference algorithm to observe the labels of the training image patches, and we evaluate the algorithm's ability to correctly infer the label associated with each test image patch.

Since the number of object classes is known *a priori*, we employ the HBNBP finite approximation Gibbs sampler of Section 4.7 to conduct posterior inference. We again use the hyperparameters $(\gamma_0, \theta_0, \gamma_d, \theta_d) = (3, 3, 1, 10)$ for all documents d and set r_d according to the heuristic $r_d = N_d(\theta_0 - 1)/(\theta_0 \gamma_0)$. We draw 10,000 samples and, for each test patch, predict the label with the highest posterior probability across the samples. We compare HBNBP performance with that of LDA using the standard variational inference algorithm of Blei, Ng, and Jordan, 2003 and *maximum a posteriori* prediction of patch labels. For each model, we set $K = 10$, allowing for the nine semantic classes plus *void*, and, following Verbeek and Bill Triggs, 2007, we ensure that the *void* class remains generic by fixing $\psi_{10}^{m} = (\frac{1}{V^m}, \cdots, \frac{1}{V^m})$ for each modality m.

Results

Figure 4.4 displays sample test image segmentations obtained using the HBNBP admixture model. Each pixel is given the predicted label of its closest patch center. Test patch classification accuracies for the HBNBP admixture model and LDA are reported in Tables 4.5 and 4.5 respectively. All results are averaged over twenty randomly generated 90% training / 10% test divisions of the dataset. The two methods perform comparably, with the HBNBP admixture model outperforming LDA in the prediction of every object class save *building*. Indeed, the mean object class accuracy is 0.79 for the HBNBP model versus 0.76 for LDA, showing that the HBNBP provides a viable alternative to more classical approaches to admixture.

Parameter Sensitivity

To test the sensitivity of the HBNBP admixture model to misspecification of the mass, concentration, and likelihood hyperparameters, we measure the fluctuation in test set per-

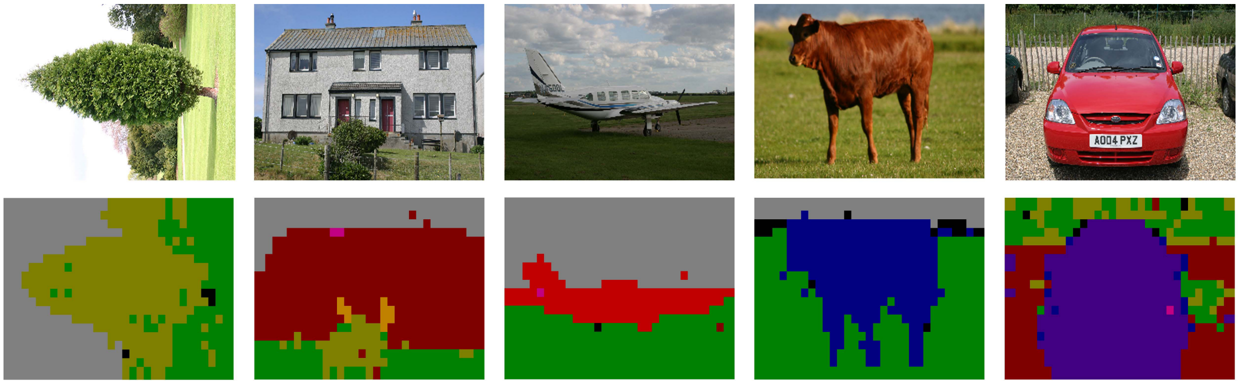


Figure 4.4: .

MSRC-v1 test image segmentations inferred by the HBNBP admixture model (best viewed in color).

Table 4.5: Confusion matrices for patch-level image segmentation and object recognition on the MSRC-v1 database. We report test image patch inference accuracy averaged over twenty randomly generated 90% training / 10% test divisions.

[HBNBP Confusion Matrix]

Predicted Class Label

Actual Class Label	building	grass	tree	cow	sky	aeroplane	face	car	bicycle
building	0.66	0.01	0.05	0.00	0.03	0.09	0.01	0.03	0.09
grass	0.00	0.89	0.06	0.02	0.00	0.01	0.00	0.00	0.00
tree	0.01	0.08	0.75	0.01	0.04	0.03	0.00	0.00	0.07
cow	0.00	0.10	0.04	0.72	0.00	0.00	0.05	0.01	0.01
sky	0.04	0.00	0.01	0.00	0.93	0.01	0.00	0.00	0.00
aeroplane	0.10	0.04	0.01	0.00	0.02	0.81	0.00	0.02	0.00
face	0.04	0.00	0.01	0.04	0.00	0.00	0.84	0.00	0.00
car	0.20	0.00	0.01	0.00	0.01	0.01	0.00	0.73	0.02
bicycle	0.16	0.00	0.04	0.00	0.00	0.00	0.00	0.02	0.73

[LDA Confusion Matrix]

Predicted Groups

Actual Groups	building	grass	tree	cow	sky	aeroplane	face	car	bicycle
building	0.69	0.01	0.04	0.01	0.03	0.07	0.01	0.03	0.08
grass	0.00	0.88	0.05	0.02	0.00	0.01	0.00	0.00	0.00
tree	0.02	0.08	0.75	0.01	0.04	0.02	0.00	0.00	0.05
cow	0.00	0.10	0.03	0.70	0.00	0.00	0.05	0.01	0.01
sky	0.05	0.00	0.02	0.00	0.91	0.01	0.00	0.00	0.00
aeroplane	0.12	0.04	0.01	0.00	0.02	0.75	0.00	0.03	0.00
face	0.04	0.00	0.01	0.05	0.00	0.00	0.80	0.00	0.00
car	0.19	0.00	0.01	0.00	0.01	0.01	0.00	0.71	0.03
bicycle	0.19	0.00	0.04	0.01	0.00	0.00	0.00	0.02	0.68

Table 4.6: Sensitivity of HBNBP admixture model to hyperparameter specification for joint image segmentation and object recognition on the MSRC-v1 database. Each hyperparameter is varied across the specified range while the remaining parameters are held fixed to the default values reported in Section 4.9. We report test patch inference accuracy averaged across object classes and over twenty randomly generated 90% training / 10% test divisions. For each test patch, we predict the label with the highest posterior probability across 2,000 samples.

Hyperparameter	Parameter range	Minimum accuracy	Maximum accuracy
γ_0	$[0.3, 30]$	0.786	0.787
θ_0	$[1.5, 30]$	0.786	0.786
η	$[2 \times 10^{-16}, 1]$	0.778	0.788

formance as each hyperparameter deviates from its default value (with the remainder held fixed). The results of this study are summarized in Table 4.6. We find that the HBNBP model is rather robust to changes in the hyperparameters and maintains nearly constant predictive performance, even as the parameters vary over several orders of magnitude.

4.10 Discussion

Motivated by problems of admixture, in which individuals are represented multiple times in multiple latent classes, we introduced the negative binomial process, an infinite-dimensional prior for vectors of counts. We developed new nonparametric admixture models based on the NBP and its conjugate prior, the beta process, and characterized the relationship between the BNBP and preexisting models for admixture. We also analyzed the asymptotics of our new priors, derived MCMC procedures for posterior inference, and demonstrated the effectiveness of our models in the domains of image segmentation and document analysis.

There are many other problem domains in which latent vectors of counts provide a natural modeling framework and where we believe that the HBNBP can prove useful. These include the computer vision task of *multiple object recognition*, where one aims to discover which and how many objects are present in a given image (Titsias, 2008), and the problem of modeling *copy number variation* in genomic regions, where one seeks to infer the underlying events responsible for large repetitions or deletions in segments of DNA (H. Chen, Xing, and Zhang, 2011).

4.A Connections

In Section 4.4 we noted that both the beta-negative binomial process (BNBP) and the gamma Poisson likelihood process (ΓPLP) provide nonparametric models for the count vectors arising in admixture models. In this section, we will elucidate some of the deeper connections

Table 4.7: A comparison of two Bayesian nonparametric constructions of clusterings such that the clusters have conditionally independent, random sizes; hence the dataset size itself is random. PP indicates a Poisson point process draw with the given intensity.

Beta negative binomial process	Gamma Poisson likelihood process
$\nu(d\psi, db) = \gamma\theta b^{-1}(1-b)^{\theta-1}\, db\; H(d\psi)$	$\nu(d\psi, d\tilde{g}) = \theta\tilde{g}^{-1}e^{-c\tilde{g}}\; d\tilde{g}\; H(d\psi)$
$(\psi_k, b_k) \sim \mathrm{PP}(\nu(d\psi, db))$	$(\psi_k, \tilde{g}_k) \sim \mathrm{PP}(\nu(d\psi, d\tilde{g}))$
$B = \sum_k b_k \delta_{\psi_k}$	$\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k}$
$\lambda_k \overset{indep}{\sim} \mathrm{Gamma}(r, \frac{1-b_k}{b_k})$	
$i_k \overset{indep}{\sim} \mathrm{Poisson}(\lambda_k)$	$i_k \overset{indep}{\sim} \mathrm{Poisson}(\tilde{g}_k)$

between these two stochastic processes. We will see that understanding these connections can not only inspire new stochastic process constructions but also lead to novel inference algorithms.

We are motivated by Table 4.7, which indicates a strong parallel between the BNBP and ΓPLP constructions for clusterings where the size of each cluster is independent and random conditioned on some underlying process. The former requires an additional random stage consisting of a draw from a gamma distribution. Here, we use the representation of the negative binomial distribution, $i \sim \mathrm{NegBin}(r, b)$, as a gamma mixture of Poisson distributions: $\tilde{b} \sim \mathrm{Gamma}(r, (1-b)/b)$ and $i \sim \mathrm{Poisson}(\tilde{b})$. However, this table mostly highlights the parallel on the level of the likelihood process and therefore on the level of classic, one-dimensional distributions. The relations between such distributions are well-studied.

Noting that many classic, one-dimensional distributions are easily obtained from each other by a simple change of variables, we aim to find new, analogous transformations in the stochastic process setting. In particular, all of our results in this section, which apply to nonparametric Bayesian priors derived from Poisson point processes, have direct analogues in the setting of one-dimensional distributions. We start by reviewing these known distributional relations. First, consider a beta distributed random variable $x \sim \mathrm{Beta}(a, b)$. Then the variable $x/(1-x)$ has a *beta prime distribution* with parameters a and b; specifically, $\beta'(a, b)$ denotes the beta prime distribution with density

$$\beta'(z \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1+z)^{-a-b}.$$

The beta prime distribution can alternatively be derived from a gamma distribution. Namely, if $x \sim \mathrm{Gamma}(a, c)$ and $y \sim \mathrm{Gamma}(b, c)$ are independent, then $x/y \sim \beta'(a, b)$. This connection is not the only one between the beta and gamma distributions though. Let

$$x \sim \mathrm{Gamma}(a, c), \quad y \sim \mathrm{Gamma}(b, c). \tag{4.11}$$

Then

$$x/(x+y) \sim \mathrm{Beta}(a, b). \tag{4.12}$$

In the rest of this section, we present similar results but now for the process case—the beta process, gamma process, and a new process we call the *beta prime process.* The proofs of these results appear in Appendix 4.B.

We start by defining a new completely random measure with nonnegative, real-valued feature weights. First, we note that, as for the processes defined in Section 4.2, there is no deterministic measure. Second, we specify that the fixed atoms have distribution

$$\eta_l \overset{indep}{\sim} \beta'(\theta\gamma\rho_l, \theta(1-\gamma\rho_l))$$

at locations (u_l). Here, $\theta > 0$, $\gamma > 0$, $(\rho_l)_{l=1}^{\infty}$, and (u_l) are parameters. As usual, while the number of fixed atoms L may be countably infinite, it is typically finite. Finally, the ordinary component has Poisson process intensity $H_{ord} \times \nu$, where

$$\nu(d\tilde{b}) = \gamma\theta\tilde{b}^{-1}(1+\tilde{b})^{-\theta}\, d\tilde{b}, \tag{4.13}$$

which we note is sigma-finite with finite mean, guaranteeing that the number of atoms generated from the ordinary component will be countably infinite with finite sum.

We abbreviate by defining $H = \sum_{l=1}^{L} \rho_l \delta_{u_l} + H_{ord}$ and say that the resulting CRM $\tilde{B} \triangleq \sum_k \tilde{b}_k \delta_{\psi_k}$ is a draw from a *beta prime process* (BPP) with base distribution H: $\tilde{B} \sim \mathrm{BPP}(\theta, \gamma, H)$. The name "beta prime process" reflects the fact that the underlying intensity is an improper beta prime distribution as well as the beta prime distribution of the fixed atoms.

With this definition in hand, we can find the stochastic process analogues of the distributional results above (with proofs in Appendix 4.B). Just as a beta prime distribution can be derived from a beta random variable, we have the following result that a similar transformation of the atom weights of a beta process yields a beta prime process.

Proposition 4.A.1. *Suppose* $B = \sum_k b_k \delta_{\psi_k} \sim \mathrm{BP}(\theta, \gamma, H)$*. Then* $\sum_k \frac{b_k}{1-b_k} \delta_{\psi_k} \sim \mathrm{BPP}(\theta, \gamma, H)$*.*

Further, just as a beta prime random variable can be derived as the ratio of gamma random variables, we find that the atoms of the beta prime process can be constructed by taking ratios of gamma random variables and the atoms of a gamma process.

Proposition 4.A.2. *Suppose* $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k} \sim \Gamma\mathrm{P}(\gamma\theta, c, H)$ *and* $\tau_k \sim \mathrm{Gamma}(\theta(1 - \gamma H(\{\psi_k\})), c)$ *independently for each* k*. Then* $\sum_k \frac{\tilde{g}_k}{\tau_k} \delta_{\psi_k} \sim \mathrm{BPP}(\theta, \gamma, H)$*.*

And, finally, the analogue to constructing a beta random variable from two gamma random variables is the construction of a beta process from a gamma process and an infinite vector of independent gamma random variables.

Proposition 4.A.3. *Suppose* $\tilde{G} = \sum_k \tilde{g}_k \delta_{\psi_k} \sim \Gamma\mathrm{P}(\gamma\theta, c, H)$ *and* $\tau_k \sim \mathrm{Gamma}(\theta(1 - \gamma H(\{\psi_k\})), c)$ *independently for each* k*. Then* $\sum_k \frac{\tilde{g}_k}{\tau_k + \tilde{g}_k} \delta_{\psi_k} \sim \mathrm{BP}(\theta, \gamma, H)$*.*

The key to the manipulations above is the Poisson process framework of the ordinary component, with the BPP providing a convenient stepping stone between the BP and ΓP. We discover, in particular, that the BP can be derived from the ΓP, elucidating a new parallel, at the prior level, between the BNBP (which we see may be thought of as a ΓPLP augmented with two additional Gamma draws per atom) and the ΓPLP. Moreover, Proposition 4.A.3 reduces sampling from a BP to sampling from a ΓP, thus allowing us to leverage any sampler for the ΓP in carrying out BNBP and HBNBP posterior inference. Inference algorithms built upon existing mature and efficient ΓP samplers see, e.g., Thibaux, 2008 could provide promising alternatives to the methods derived in Section 4.7.

4.B Proofs for Appendix 4.A

Proof of Proposition 4.A.1: First, consider the ordinary component of a beta process. The Mapping Theorem of Kingman (1993) tells us that if the collection of tuples (ψ_k, b_k) come from a Poisson process with intensity $H_{ord} \times \nu_{beta}$, where ν_{beta} is the beta process intensity of Eq. (4.2), then the collection of tuples $(\psi_k, b_k/(1-b_k))$ are draws from a Poisson process with intensity $H_{ord} \times \nu$, where we apply a change of variables to find:

$$\begin{aligned} \nu(d\tilde{b}) &= \gamma\theta \left(\frac{\tilde{b}}{1+\tilde{b}}\right)^{-1} \left(1 - \frac{\tilde{b}}{1+\tilde{b}}\right)^{\theta-1} \frac{1}{(1+\tilde{b})^2}\, d\tilde{b} \\ \nu(d\tilde{b}) &= \gamma\theta\tilde{b}^{-1}(1+\tilde{b})^{-\theta}\, d\tilde{b}, \end{aligned}$$

which matches Eq. (4.13).

For any particular atom where $b_k \sim \text{Beta}(\theta\gamma\rho_k, \theta(1-\gamma\rho_k))$ and $\rho_k = H(\{\psi_k\}) > 0$, we simply quote the well-known, one-dimensional change of variables $b_k/(1-b_k) \sim \beta'(\theta\gamma\rho_k, \theta(1-\gamma\rho_k))$.

Since there is no deterministic component, we have considered all components of the completely random measure. □

Proof of Proposition 4.A.2: We again start with the ordinary component of a completely random measure. In particular, we assume the collection of tuples $(\psi_k, \tilde{g}_k)$ is generated according to a Poisson process with intensity $H_{ord} \times \nu_{gamma}$, where ν_{gamma} is the gamma process intensity of Eq. (4.4).

Consider a random variable $\tau_k \sim \text{Gamma}(\theta, c)$ associated with each such tuple. Then $1/\tau_k \sim \text{IG}(\theta, c)$. We consider a marked Poisson process with mark $\tilde{b}_k \triangleq \tilde{g}_k/\tau_k$ at tuple $(\psi_k, \tilde{g}_k)$ of the original process. By the scaling property of the inverse gamma distribution, we note $\tilde{b}_k \sim \text{IG}(\theta, c\tilde{g}_k)$ given $\tilde{g}_k$. So the Marking Theorem (Kingman, 1993) implies that the collection of tuples $(\psi_k, \tilde{b}_k)$ is itself a draw from a Poisson point process with intensity $H_{ord} \times \nu$, where

$$\nu(d\tilde{b}) = \int p(\tilde{b} \mid \theta, c, \tilde{g})\, \nu(d\tilde{g})\, d\tilde{g}\, d\tilde{b}$$

$$
\begin{aligned}
&= d\tilde{b} \int \frac{1}{\Gamma(\theta)} (c\tilde{g})^{\theta} \tilde{b}^{-\theta-1} \exp(-c\tilde{g}/\tilde{b}) \cdot \gamma\theta\tilde{g}^{-1} \exp(-c\tilde{g}) \, d\tilde{g} \\
&= \gamma\theta c^{\theta} \frac{1}{\Gamma(\theta)} \tilde{b}^{-\theta-1} d\tilde{b} \int \tilde{g}^{\theta-1} \exp(-\tilde{g}c(1+\tilde{b})/\tilde{b}) \, d\tilde{g} \\
&= \gamma\theta c^{\theta} \frac{1}{\Gamma(\theta)} \tilde{b}^{-\theta-1} \Gamma(\theta) \left(\frac{\tilde{b}}{c(1+\tilde{b})} \right)^{\theta} \, d\tilde{b} \\
&= \gamma\theta\tilde{b}^{-1} \left(1+\tilde{b}\right)^{-\theta} \, d\tilde{b},
\end{aligned}
$$

which matches the beta prime process ordinary component intensity of Eq. (4.13).

For any particular atom of the gamma process, $\tilde{g}_k \sim \text{Gamma}(\theta\gamma\rho_k, c)$ with $\rho_k = H(\{\psi_k\}) > 0$, we have $\tau_k \sim \text{Gamma}(\theta(1-\gamma\rho_k), c)$ by construction. Then it is well known that $\tilde{g}_k/\tau_k$ has the $\beta'(\theta\gamma\rho_k, \theta(1-\gamma\rho_k))$ distribution, as desired.

There is no deterministic component of the gamma process. □

Proof of Proposition 4.A.3: Proposition 4.A.3 follows from Proposition 4.A.2, once we reverse the relationship established in Proposition 4.A.1. For completeness, we provide a more direct, self-contained proof paralleling that of Propositions 4.A.1 and 4.A.2 above. We first note that Proposition 4.A.3 can be derived from Proposition 4.A.2 and an inverse change of variables from that in Proposition 4.A.1. We begin with the ordinary component of the gamma process so that the collection of tuples $(\psi_k, \tilde{g}_k)$ is generated according to a Poisson process with intensity $H_{ord} \times \nu_{gamma}$, where ν_{gamma} is the gamma process intensity of Eq. (4.4). The Marking Theorem (Kingman, 1993) tells us that the marked Poisson process with points $(\psi_k, \tilde{g}_k, \tau_k)$ has intensity $H_{ord} \times \nu$, where

$$\nu(d\tilde{g}, d\tau) = \gamma\theta\tilde{g}^{-1} e^{-c\tilde{g}} \cdot (\Gamma(\theta))^{-1} \tau^{\theta-1} \exp(-c\tau) \, c^{\theta} \, d\tilde{g} \, d\tau.$$

Now consider the change of variables $u = \tilde{g}/(\tilde{g}+\tau), v = \tilde{g}+\tau$. The reverse transformation is $\tilde{g} = uv, \tau = (1-u)v$ with Jacobian v. Then the Poisson point process with points (ψ_k, u_k, v_k) has intensity $H_{ord} \times \nu$, where

$$\nu(d\psi, du, dv) = (\Gamma(\theta))^{-1} \gamma\theta c^{\theta} u^{-1} v^{-1} (1-u)^{\theta-1} v^{\theta-1} e^{-cv} \cdot v \, du \, dv.$$

So the Poisson point process with points (ψ_k, u_k) has intensity $H_{ord} \times \nu$, with

$$
\begin{aligned}
\nu(d\psi, du) &= \int_v \mu(d\psi, du, dv) \\
&= \int_v (\Gamma(\theta))^{-1} \gamma\theta c^{\theta} u^{-1} (1-u)^{\theta-1} v^{\theta-1} e^{-cv} \, du \, dv \\
&= (\Gamma(\theta))^{-1} \gamma\theta c^{\theta} u^{-1} (1-u)^{\theta-1} \Gamma(\theta) c^{-\theta} \, du \\
&= \gamma\theta u^{-1} (1-u)^{\theta-1} \, du,
\end{aligned}
$$

which is the known beta process intensity.

In the discrete case with $H(\{\psi_k\}) = \rho_k > 0$, we have by construction

$$\tilde{g}_k \sim \text{Gamma}(\theta\gamma\rho_k, c)$$

and

$$\tau_k \sim \text{Gamma}(\theta(1-\gamma\rho_k), c).$$

From classic finite distributional results, we have

$$\frac{\tilde{g}_k}{\tau_k + \tilde{g}_k} \sim \text{Beta}(\theta\gamma\rho_k, \theta(1-\gamma\rho_k)),$$

exactly as in the case of the beta process.

As the gamma process and beta process each have no deterministic components, this completes the proof. □

4.C Full results for Section 4.5

In order to fill in Table 4.1, we start by briefly establishing the results for the expected number of clusters of size j for the DP and PYP; the results for the expected total number of clusters are cited in the main text. We then move on to full results for the BNBP and 3BNBP. Proofs for all results in this section appear in Appendix 4.D.

Theorem 4.C.1. *Assume that the concentration parameter for the* DP *satisfies $\theta > 0$. Then the expected number of data clusters of size j, $\Phi_j(N)$, has asymptotic growth*

$$\Phi_j(N) \sim \theta j^{-1}, \quad N \to \infty.$$

Theorem 4.C.2. *Assume that the discount parameter for the* PYP *satisfies $\alpha \in (0,1)$ and the concentration parameter satisfies $\theta > 1-\alpha$. Then the expected number of data clusters of size j, $\Phi_j(N)$, has asymptotic growth*

$$\Phi_j(N) \sim \frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)} N^{\alpha}, \quad N \to \infty.$$

Next we establish how the expected number of data points, $\xi(r)$, grows asymptotically with r in the BNBP case (in Lemma 4.C.4) and the 3BNBP case (in Lemma 4.C.5). We begin by showing that the expected number of data points is infinite for the concentration parameter range $\theta \leq 1-\alpha$ in both the BNBP ($\alpha = 0$) and 3BNBP models.

Lemma 4.C.3. *Assume that the discount parameter for the three-parameter beta process satisfies $\alpha \in [0,1)$ (the beta process is the special case when $\alpha = 0$), the concentration parameter satisfies $\theta \leq 1-\alpha$, and the mass parameter satisfies $\gamma > 0$. Then the expected number of data points, $\xi(r) = \mathbb{E}[\sum_k i_k]$, from a* BNBP *or* 3BNBP*, as appropriate, is infinite.*

Lemma 4.C.4. *Assume that the concentration parameter for the beta process satisfies $\theta > 1$ and the mass parameter satisifies $\gamma > 0$. Then the expected number of data points $\xi(r) = \mathbb{E}[\sum_k i_k]$ from a* BNBP *has asymptotic growth*

$$\xi(r) \sim \gamma \frac{\theta}{\theta - 1} r, \quad r \to \infty.$$

Lemma 4.C.5. *Assume that a three-parameter beta process has discount parameter $\alpha \in (0, 1)$ and concentration parameter $\theta > 1 - \alpha$. Then the expected number of data points $\xi(r) = \mathbb{E}[\sum_k i_k]$ from a* 3BNBP *has asymptotic growth*

$$\xi(r) \sim \gamma \frac{\theta}{\theta + \alpha - 1} r, \quad r \to \infty.$$

Next, we establish how the expected number of clusters, $\Phi(r)$, grows asymptotically as $r \to \infty$ in the BNBP case (in Lemma 4.C.6) and in the 3BNBP case (in Lemma 4.C.7).

Lemma 4.C.6. *Let $\theta > 0$. Then the expected number of clusters $\Phi(r) = \mathbb{E}[\sum_k \mathbb{1}\{i_k > 0\}]$ from a* BNBP *has asymptotic growth*

$$\Phi(r) \sim \gamma\theta \log r, \quad r \to \infty.$$

Lemma 4.C.7. *Consider a three-parameter beta process. Let the discount parameter satisfy $\alpha > 0$ and the concentration parameter satisfy $\theta > -\alpha$. Then the number of clusters $K(r) \sum_k \mathbb{1}\{i_k > 0\}$ from a* 3BNBP *has almost sure asymptotic growth*

$$K(r) \overset{a.s.}{\sim} \frac{\gamma}{\alpha} \frac{\Gamma(\theta + 1)}{\Gamma(\theta + \alpha)} r^{\alpha}, \quad r \to \infty.$$

We are also interested in how the expected number of clusters of size j, $\Phi_j(r)$, grows as $r \to \infty$. To that end, we establish this asymptotic growth in the BNBP case in Lemma 4.C.8 and in the 3BNBP case in Lemma 4.C.9 below.

Lemma 4.C.8. *Let $\theta > 0$. Then the expected number of clusters of size j, $\Phi_j(r) = \mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}]$, from a* BNBP *has asymptotic growth*

$$\Phi_j(r) \sim \gamma\theta j^{-1}, \quad r \to \infty.$$

That is, the number is asymptotically constant in r.

Lemma 4.C.9. *Let $\theta > -\alpha$ and $\alpha \in (0, 1)$. Then the expected number of clusters of size j, $\Phi_j(r) = \mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}]$, from a* 3BNBP *has asymptotic growth*

$$\Phi_j(r) \sim \gamma \frac{\Gamma(1 + \theta)}{\Gamma(1 - \alpha)\Gamma(\theta + \alpha)} \frac{\Gamma(j - \alpha)}{\Gamma(j + 1)} r^{\alpha}, \quad r \to \infty.$$

Finally, we wish to combine these results to establish asymptotic results for the diversity, i.e., the expected number of clusters (or clusters of size j) as the expected number of data points varies. We find the asymptotic growth in the number of clusters for the BNBP in Theorem 4.C.10 and for the 3BNBP in Theorem 4.C.11. We find the asymptotic growth in the number of clusters of size j for the BNBP (in fact, the result has already been shown in Lemma 4.C.8) and for the 3BNBP in Theorem 4.C.12.

Theorem 4.C.10. *Let $\theta > 1$. Then the expected number of clusters Φ grows asymptotically as the log of the expected number of data points ξ:*

$$\Phi(r) \sim \gamma\theta \log(\xi(r)), \quad r \to \infty.$$

Theorem 4.C.11. *Let $\theta + \alpha > 1$ and $\alpha \in (0, 1)$. Then the number of clusters K grows asymptotically as a power of the expected number of data points ξ:*

$$K(r) \overset{a.s.}{\sim} \frac{\gamma^{1-\alpha}}{\alpha} \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \left(\frac{\theta+\alpha-1}{\theta}\right)^{\alpha} (\xi(r))^{\alpha}, \quad r \to \infty.$$

Theorem 4.C.12. *Let $\theta + \alpha > 1$ and $\alpha \in (0, 1)$. Then the expected number of clusters of size j, Φ_j, grows asymptotically as a power of the expected number of data points ξ:*

$$\Phi_j(r) \sim \gamma^{1-\alpha} \frac{\Gamma(\theta+1)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(j-\alpha)}{\Gamma(j+1)} \left(\frac{\theta+\alpha-1}{\theta}\right)^{\alpha} (\xi(r))^{\alpha}, \quad r \to \infty.$$

4.D Proofs for Appendix 4.C

Proof of Theorem 4.C.1: When cluster proportions are generated according to a Dirichlet process and cluster belonging is generated according to draws from the resulting random measure, the joint distribution of $(K_1(N), \ldots, K_N(N))$ is described by the *Ewens sampling formula*, which appears as Eq. 2.9 in (Watterson, 1974). It follows that Eq. 2.22 in (Watterson, 1974) gives $\Phi_j(N) = \mathbb{E}[K_j(N)]$:

$$\Phi_j(N) = \frac{\theta}{j} \binom{\theta+N-j-1}{N-j} \cdot \binom{\theta+N-1}{N}^{-1}.$$

Therefore,

$$\begin{aligned}
\Phi_j(N) &= \frac{\theta}{j} \frac{\Gamma(\theta+N-j)}{\Gamma(N-j+1)\Gamma(\theta)} \cdot \frac{\Gamma(N+1)\Gamma(\theta)}{\Gamma(N+\theta)} \\
&= \frac{\theta}{j} \cdot \frac{\Gamma(N+\theta-j)}{\Gamma(N+\theta)} \cdot \frac{\Gamma(N+1)}{\Gamma(N+1-j)} \\
&\sim \frac{\theta}{j} \cdot (N+\theta)^{-j} \cdot (N+1)^{j}, \quad N \to \infty
\end{aligned}$$

$$\sim \frac{\theta}{j}, \quad N \to \infty,$$

where the asymptotics for the ratios of gamma functions follow from Tricomi and Erdélyi (1951). □

Proof of Theorem 4.C.2: Pitman (2006) establishes that, for the PYP with parameters θ and α given in the result statement, we have $\Phi(N) \sim \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)} N^{\alpha}$ as $N \to \infty$.

Note that $\Phi(N)$ is in the form of Eq. 48 on p. 167 of (Gnedin, Hansen, and Pitman, 2007). The desired result follows by applying Eq. 51 on p. 167 of (Gnedin, Hansen, and Pitman, 2007). □

Proof of Lemma 4.C.3: In this case, we have

$$\begin{aligned}
\mathbb{E}[\sum_k i_k] &= \mathbb{E}\left[\mathbb{E}[\sum_k i_k | \mathbf{b}.]\right] \\
&\text{by the tower property} \\
&= \mathbb{E}\left[\sum_k \mathbb{E}[i_k | \mathbf{b}.]\right] \\
&\text{by monotonicity} \\
&= \mathbb{E}\left[\sum_k \frac{b_k r}{(1-b_k)}\right] \\
&\text{using the mean of the negative binomial distribution} \\
&= \int_0^1 \frac{br}{(1-b)}\, \nu(db) \\
&\text{by Campbell's Theorem (Kingman, 1993)} \\
&= r\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \int_0^1 b^{-\alpha}(1-b)^{\theta+\alpha-2}\, db.
\end{aligned}$$

The final line is finite iff

$$1-\alpha > 0, \quad \text{and} \quad \theta+\alpha-1 > 0.$$

Equivalently, the final line is finite iff

$$\alpha < 1 \quad \text{and} \quad \theta > 1-\alpha.$$

□

Proof of Lemma 4.C.4: Let $B = \sum_k b_k \psi_k$ be beta process distributed. Let $i_k \overset{iid}{\sim} \text{NegBin}(r, b_k)$. By the Marking theorem (Kingman, 1993), the Poisson process $\{(\psi_k, b_k, i_k)\}$ has intensity

$$\nu(d\psi, db, i) = \gamma\theta b^{-1}(1-b)^{\theta-1}\binom{i+r-1}{i}(1-b)^r b^i \; db \; H_{ord}(d\psi). \tag{4.14}$$

So the Poisson process $\{i_k\}$ has intensity

$$\nu(i) = \gamma\theta \frac{\Gamma(i+r)}{\Gamma(i+1)\Gamma(r)} \frac{\Gamma(i)\Gamma(r+\theta)}{\Gamma(i+r+\theta)}.$$

Thus, by Campbell's theorem (Kingman, 1993),

$$\mathbb{E}[\sum_k i_k] = \sum_{i=1}^{\infty} i\nu(i) = \gamma\theta \frac{\Gamma(r+\theta)}{\Gamma(r)} \sum_{i=1}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)}.$$

To evaluate the sum $\sum_{i=1}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)}$, we appeal to a result from Tricomi and Erdélyi (1951):

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} = x^{a-b}\left[1 + \frac{(a-b)(a+b-1)}{2x} + O(x^{-2})\right], \quad x \to \infty. \tag{4.15}$$

In particular,

$$\frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \leq (i+r)^{-\theta}\left[1 - \frac{\theta(\theta-1)}{2(i+r)} + C(i+r)^{-2}\right] \qquad \text{for some constant } C$$

and

$$\frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \geq (i+r)^{-\theta}\left[1 - \frac{\theta(\theta-1)}{2(i+r)} - C'(i+r)^{-2}\right] \qquad \text{for some constant } C'.$$

Before proceeding, we establish for $a > 1$,

$$\sum_{i=1}^{\infty}(i+r)^{-a} \leq \int_{x=0}^{\infty}(x+r)^{-a} \; dx = (a-1)^{-1}r^{1-a}$$

and

$$\sum_{i=1}^{\infty}(i+r)^{-a} \geq \int_{x=1}^{\infty}(x+r)^{-a} \; dx = (\alpha-1)^{-1}(r+1)^{1-a}.$$

So

$$\sum_{i=1}^{\infty}\frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \leq (\theta-1)^{-1}r^{1-\theta} - \frac{\theta-1}{2}(r+1)^{-\theta} + C(\theta+1)^{-1}r^{-\theta-1}$$

and

$$\sum_{i=1}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \geq (\theta-1)^{-1}(r+1)^{1-\theta} - \frac{\theta-1}{2} r^{-\theta} - C(\theta+1)^{-1}(r+1)^{-\theta-1}.$$

Since, for $\theta > 1$, we have

$$\frac{r^{1-\theta}}{(r+1)^{1-\theta}} \to 1, \quad r \to \infty, \tag{4.16}$$

it follows that

$$\sum_{i=1}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \sim (\theta-1)^{-1} r^{1-\theta}. \tag{4.17}$$

From Eq. (4.15), we also have $\frac{\Gamma(r+\theta)}{\Gamma(r)} \sim r^{\theta}$ as $r \to \infty$. So we conclude that

$$\mathbb{E}[\sum_k i_k] \sim \gamma \frac{\theta}{\theta-1} r, \quad r \to \infty,$$

as desired. □

Proof of Lemma 4.C.5: The proof proceeds as above. In this case, we have that the Poisson process $\{(\psi_k, b_k, i_k)\}$ has intensity

$$\nu(d\psi, db, i) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} b^{-1-\alpha}(1-b)^{\theta+\alpha-1} \frac{\Gamma(i+r)}{\Gamma(i+1)\Gamma(r)} (1-b)^r b^i \, db \, H(d\psi).$$

So the Poisson process $\{i_k\}$ has intensity

$$\nu(i) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(i+r)}{\Gamma(i+1)\Gamma(r)} \frac{\Gamma(i-\alpha)\Gamma(r+\theta+\alpha)}{\Gamma(i+r+\theta)}.$$

By Campbell's theorem,

$$\mathbb{E}[\sum_k i_k] = \sum_{i=1}^{\infty} i\nu(i) = \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(r+\theta+\alpha)}{\Gamma(r)} \sum_{i=1}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \frac{\Gamma(i-\alpha)}{\Gamma(i)}.$$

We will find the following inequalities, with $i \geq 1$ and $\alpha \in (0,1)$, useful (cf. Eq. 2.8 in Qi and Losonczi, 2010):

$$(i-\alpha)^{-\alpha} \leq \frac{\Gamma(i-\alpha)}{\Gamma(i)} \leq (i-1)^{-\alpha}. \tag{4.18}$$

We will also find the following integrals useful. Let $a > 1$.

$$\sum_{i=2}^{\infty} (i+r)^{-a}(i-\alpha)^{-\alpha} \leq \sum_{i=2}^{\infty} (i+r)^{-a}(i-1)^{-\alpha}$$

$$\begin{aligned}
&\leq \int_{x=0}^{\infty} (x+r)^{-a} x^{-\alpha}\, dx \\
&= r^{-a-\alpha+1} \int_{y=0}^{\infty} (y+1)^{-a} y^{-\alpha}\, dy \\
&= r^{-a-\alpha+1} \frac{\Gamma(1-\alpha)\Gamma(a+\alpha-1)}{\Gamma(a)}. \qquad (4.19)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{i=2}^{\infty} (i+r)^{-a}(i-1)^{-\alpha} &\geq \sum_{i=2}^{\infty} (i+r)^{-a}(i-\alpha)^{-\alpha} \\
&\geq \int_{x=2}^{\infty} (x+r)^{-a} x^{-\alpha}\, dx \\
&= \int_{x=0}^{\infty} (x+r)^{-a} x^{-\alpha}\, dx - \int_{0}^{2} (x+r)^{-a} x^{-\alpha}\, dx \\
&\geq r^{-a-\alpha+1} \frac{\Gamma(1-\alpha)\Gamma(a+\alpha-1)}{\Gamma(a)} - r^{-a}(1-\alpha)^{-1} 2^{1-\alpha}. \qquad (4.20)
\end{aligned}$$

First, we consider an upper bound. To that end,

$$\begin{aligned}
\sum_{i=2}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \frac{\Gamma(i-\alpha)}{\Gamma(i)} &\leq \sum_{i=2}^{\infty} (i+r)^{-\theta} \left(1 - \frac{\theta(\theta+1)}{2(i+r)} + C(i+r)^{-2}\right)(i-1)^{-\alpha} \\
&\quad \text{for some constant } C \\
&\leq r^{-\theta-\alpha+1} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha-1)}{\Gamma(\theta)} \\
&\quad - \frac{\theta(\theta+1)}{2} r^{-\theta-\alpha} \frac{\Gamma(1-\alpha)\Gamma(\theta+1+\alpha-1)}{\Gamma(\theta+1)} - r^{-\theta-1}(1-\alpha)^{-1} 2^{1-\alpha} \\
&\quad + C r^{-\theta-\alpha-1} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha+1)}{\Gamma(\theta+2)}.
\end{aligned}$$

For the lower bound,

$$\begin{aligned}
\sum_{i=2}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \frac{\Gamma(i-\alpha)}{\Gamma(i)} &\geq \sum_{i=2}^{\infty} (i+r)^{-\theta} \left(1 - \frac{\theta(\theta+1)}{2(i+r)} - C'(i+r)^{-2}\right)(i-\alpha)^{-\alpha} \\
&\quad \text{for some constant } C' \\
&\geq r^{-\theta-\alpha+1} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha-1)}{\Gamma(\theta)} - r^{-\theta}(1-\alpha)^{-1} 2^{1-\alpha} \\
&\quad - r^{-\theta-\alpha} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}{\Gamma(\theta+1)}
\end{aligned}$$

$$- C' r^{-\theta-\alpha-1} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha+1)}{\Gamma(\theta+2)}.$$

It follows from the two bounds above that

$$\sum_{i=2}^{\infty} \frac{\Gamma(i+r)}{\Gamma(i+r+\theta)} \frac{\Gamma(i-\alpha)}{\Gamma(i)} \sim \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha-1)}{\Gamma(\theta)} r^{-\theta-\alpha+1}.$$

Since

$$\frac{\Gamma(r+\theta+\alpha)}{\Gamma(r)} \sim r^{\theta+\alpha},$$

it follows that

$$\mathbb{E}[\sum_k i_k] \sim \gamma \frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)} \frac{\Gamma(1-\alpha)\Gamma(\theta+\alpha-1)}{\Gamma(\theta)} r = \gamma \frac{\theta}{\theta+\alpha-1} r,$$

as was to be shown. □

Proof of Lemma 4.C.6: Given an atom b_k of the beta process, the probability that the associated negative binomial count i_k is non-zero is $1-(1-b_k)^r$. It follows that

$$\mathbb{E}[\sum_k \mathbb{1}\{i_k > 0\}] = \mathbb{E}[\mathbb{E}[\sum_k \mathbb{1}\{i_k > 0\} | b_k]] = \mathbb{E}[\sum_k 1 - (1-b_k)^r] = \int_b (1-(1-b)^r) \nu_{\mathrm{BP}}(db),$$

where ν_{BP} is the intensity of beta process atoms $\{b_k\}$. For integer r, this integral was calculated by Broderick, Jordan, and Pitman (2012) to be $\sim \gamma\theta\log(r)$.

Note that, in applying the result of Broderick, Jordan, and Pitman (2012), we are using the form of the negative binomial distribution to reinterpret the desired expectation as the expected number of features represented in a beta-Bernoulli process with r draws from the same underlying base measure.

Now consider general $r > 1$. Let $r^{(0)} = \lfloor r \rfloor$ and $r^{(1)} = \lceil r \rceil$. Then

$$\frac{\int_b (1-(1-b)^{r^{(0)}}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r^{(1)})} \leq \frac{\int_b (1-(1-b)^{r}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r)} \leq \frac{\int_b (1-(1-b)^{r^{(1)}}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r^{(0)})} \tag{4.21}$$

by monotonicity. Moreover,

$$\begin{aligned} \frac{\int_b (1-(1-b)^{r^{(0)}}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r^{(1)})} &= \frac{\int_b (1-(1-b)^{r^{(0)}}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r^{(0)})} \cdot \frac{\gamma\theta\log(r^{(0)})}{\gamma\theta\log(r^{(1)})} \\ &\to 1, \quad r \to \infty. \end{aligned}$$

Similarly,

$$\frac{\int_b (1-(1-b)^{r^{(1)}}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r^{(0)})} \to 1, \quad r \to \infty \qquad \text{and hence} \qquad \frac{\int_b (1-(1-b)^{r}) \nu_{\mathrm{BP}}(db)}{\gamma\theta\log(r)} \to 1, \quad r \to \infty.$$

as was to be shown. □

Proof of Lemma 4.C.7: By the discussion in the previous proposition, this result follows from the results in Broderick, Jordan, and Pitman (2012). □

Proof of Lemma 4.C.8: Given an atom b_k of the beta process, the probability that the associated negative binomial count i_k is equal to j is $\mathrm{NegBin}(j|r, b_k)$. It follows that

$$\mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}] = \mathbb{E}[\mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}|\mathbf{b}.]] = \mathbb{E}[\sum_k \mathrm{NegBin}(j|r, b_k)] = \nu(j) = \gamma\theta\frac{\Gamma(j+r)}{\Gamma(j+1)\Gamma(r)}\frac{\Gamma(j)\Gamma(r+\theta)}{\Gamma(j+r+\theta)}$$

as above. Now we use $\frac{\Gamma(r+\theta)}{\Gamma(r)} \sim r^{\theta}$ and $\frac{\Gamma(j+r)}{\Gamma(j+r+\theta)} \sim r^{-\theta}$ to obtain $\mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}] \sim \gamma\theta j^{-1}$. □

Proof of Lemma 4.C.9: As in the BNBP case, we have

$$\mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}] = \nu(j) = \gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\frac{\Gamma(j+r)}{\Gamma(j+1)\Gamma(r)}\frac{\Gamma(j-\alpha)\Gamma(r+\theta+\alpha)}{\Gamma(j+r+\theta)}$$

Now we use $\frac{\Gamma(r+\theta+\alpha)}{\Gamma(r)} \sim r^{\theta+\alpha}$ and $\frac{\Gamma(j+r)}{\Gamma(j+r+\theta)} \sim r^{-\theta}$ to obtain $\mathbb{E}[\sum_k \mathbb{1}\{i_k = j\}] \sim \gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\frac{\Gamma(j-\alpha)}{\Gamma(j+1)}r^{\alpha}$. □

Proof of Theorem 4.C.10: Assume $\theta > 1$. We have from the previous discussion that $\lim_{r\to\infty}\frac{\xi(r)}{\gamma\frac{\theta}{\theta-1}r} = 1$. So

$$\lim_{r\to\infty} \log(\xi(r)) - \log(r) = -\log\left(\gamma\frac{\theta}{\theta-1}\right).$$

Hence $\lim_{r\to\infty}\frac{\log(\xi(r))}{\log(r)} = 1$ since $\log(r) \to \infty$ as $r \to \infty$.

From Lemma 4.C.6, we also have $\lim_{r\to\infty}\frac{\Phi(r)}{\gamma\theta\log(r)} = 1$. Finally, then,

$$\lim_{r\to\infty}\frac{\Phi(r)}{\gamma\theta\log(\xi(r))} = 1.$$

□

Proof of Theorem 4.C.11: From above, we have

$$\lim_{r\to\infty} \frac{\xi(r)}{\gamma\frac{\theta}{\theta+\alpha-1}r} = 1 \qquad \text{and hence} \qquad \lim_{r\to\infty} \frac{(\xi(r))^\alpha}{\left(\gamma\frac{\theta}{\theta+\alpha-1}r\right)^\alpha} = 1.$$

From Lemma 4.C.7, we also have

$$\lim_{r\to\infty} \frac{K(r)}{\frac{\gamma}{\alpha}\frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}r^\alpha} \overset{a.s.}{=} 1 \qquad \text{and hence} \qquad \lim_{r\to\infty} \frac{(\xi(r))^\alpha\frac{\gamma}{\alpha}\frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}}{\left(\gamma\frac{\theta}{\theta+\alpha-1}\right)^\alpha K(r)} \overset{a.s.}{=} 1.$$

□

Proof of Theorem 4.C.12: As above, we have from Lemma 4.C.9 that

$$\lim_{r\to\infty} \frac{\Phi_j(r)}{\gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\frac{\Gamma(j-\alpha)}{\Gamma(j+1)}r^\alpha} = 1 \qquad \text{and hence} \qquad \lim_{r\to\infty} \frac{(\xi(r))^\alpha\gamma\frac{\Gamma(1+\theta)}{\Gamma(1-\alpha)\Gamma(\theta+\alpha)}\frac{\Gamma(j-\alpha)}{\Gamma(j+1)}}{\left(\gamma\frac{\theta}{\theta+\alpha-1}\right)^\alpha \Phi_j(r)} = 1,$$

yielding the desired result. □

4.E Conjugacy proofs

Full beta process and negative binomial process

Theorem 4.3.3 in the main text is a corollary of Theorems 4.E.1 and 4.E.2 below. In particular, Theorems 4.E.1 and 4.E.2 give us the form of the posterior process when we have a general CRM prior with a Poisson process intensity with finite mean. Choosing the particular Poisson process intensity for the RBP and choosing the distributions of the prior fixed weights yields the result.

Finite Poisson process intensity

Theorem 4.E.1. *Let B_{prior} be a discrete, completely random measure on $[0,1]$ with atom locations in $[0,1]$. Suppose it has the following components.*

- *The ordinary component is generated from a Poisson point process with intensity $\nu(db)\, d\psi$ such that ν is continuous and $\nu[0,1] < \infty$. In particular, the weights are in the b axis, and the atom locations are in the ψ axis.*
- *There are L fixed atoms at locations $u_1, \ldots, u_L \in [0,1]$. The weight of the lth fixed atom is a random variable with distribution h_l.*

- *There is no deterministic measure component.*

Draw a negative binomial process I with shape parameter r and input measure B_{prior}. Let K be the number of (nonzero) atoms of I. Let $\Pi = \{(i_k, s_k)\}_{k=1}^{K}$ be the pairs of observed nonzero counts and corresponding atom locations.

Then the posterior process for the input measure to the negative binomial process given I is a completely random measure B_{post} with the following components.

- *The ordinary component is generated from a Poisson point process with intensity*

$$(1-b)^r \nu(db) \, d\psi.$$

- *There are three sets of fixed atoms.*

  1. *There are the old, repeated fixed atoms. If $u_l = s_k$ for some k, there is a fixed atom at u_l with weight density*

     $$c_{or}^{-1} (1-b)^r b^{i_k} h_l(db),$$

     where c_{or} is the normalizing constant:

     $$c_{or} = \int_{b=0}^{1} (1-b)^r b^{i_k} h_l(db).$$

  2. *There are the old, unrepeated fixed atoms. If $u_l \notin \{s_1, \ldots, s_K\}$, there is a fixed atom at u_l with weight density*

     $$c_{ou}^{-1} (1-b)^r h_l(db),$$

     where c_{or} is the normalizing constant:

     $$c_{ou} = \int_{b=0}^{1} (1-b)^r h_l(db).$$

  3. *There are the new fixed atoms. If $s_k \notin \{u_1, \ldots, u_L\}$, there is a fixed atom at s_k with weight density*

     $$c_{new}^{-1} (1-b)^r b^{i_k} \nu(db),$$

     where c_{new} is the normalizing constant:

     $$c_{new} = \int_{b=0}^{1} (1-b)^r b^{i_k} \nu(db).$$

- *There is no deterministic measure component.*

Proof of Theorem 4.E.1: Our proof follows the proof of beta-Bernoulli process conjugacy of Kim (1999a). Let $(\mathfrak{M}, \Sigma_{\mathfrak{M}})$ be the set of completely random measures on $[0, 1]$ with weights in $[0, 1]$ and its associated sigma algebra. Let $(\mathfrak{G}, \Sigma_{\mathfrak{G}})$ be the set of completely random measures on $[0, 1]$ with atom weights in $\{1, 2, \ldots\}$ and its associated sigma algebra. For any sets $M \in \Sigma_{\mathfrak{M}}$ and $G \in \Sigma_{\mathfrak{G}}$, let $\mathbb{P}_{prior}(M \times G)$ be the probability distribution induced on such sets by the construction of the prior measure B_{prior} and the negative binomial process I. Let $\mathbb{Q}(M : G)$ be the probability distribution induced on measures in $\mathfrak{M}$ by the proposed posterior distribution. Finally, let $\mathbb{P}_{marg}(G)$ be the prior marginal distribution on counting measures in $\mathfrak{G}$. To prove the theorem, it is enough to show that, for any such sets M and G, we have

$$\mathbb{P}_{prior}(M \times G) = \int_{I \in G} \mathbb{Q}(M : I) \, \mathbb{P}_{marg}(I). \tag{4.22}$$

The remainder of the proof will proceed as follows. We start by introducing some further notation. Then we will note that it is enough to prove Eq. (4.22) for certain, restricted forms of the sets M and G. Next, we will in turn find the form of each of (1) the prior distribution $\mathbb{P}_{prior}$, (2) the proposed posterior distribution $\mathbb{Q}$, and (3) the marginal count process distribution $\mathbb{P}_{marg}$ for our special sets of interest. Finally, we will show that we can integrate out the posterior with respect to the marginal in order to obtain the prior, as in Eq. (4.22).

Start by noting that we can write B_{prior} as

$$B_{prior}(d\psi) = \sum_{j=1}^{J} \xi_j \delta_{v_j}(d\psi) + \sum_{l=1}^{L} \eta_l \delta_{u_l}(d\psi). \tag{4.23}$$

Here, J is the number of atoms in the ordinary component of B_{prior}. So the total number of atoms in B_{prior} is $J + L$, and the total number of atoms in the counting measure with parameter B_{prior} is $K \leq J + L$. The atom locations of the ordinary component are $\{v_j\}$, and the fixed atom locations are at $\{u_l\}$. We will assume these location collections are each respectively in increasing order: $v_1 \leq v_2 \leq \cdots v_J$ and $u_1 \leq u_2 \leq \cdots u_L$. Note that the v_j order is well-defined since the density of the v_j is continuous. Together, we have that the full set of atoms of the counting measure is some subset of the disjoint union of the two types of prior atoms: $\{s_k\}_{k=1}^K \subseteq \{v_j\}_{j=1}^J \cup \{u_l\}_{l=1}^L$. The atom weight at the fixed u_l location is η_l, and the atom weight at the ordinary component location v_j is ξ_j.

Let $\lambda = \nu[0, 1]$, which we know to be finite by assumption. Then the number of atoms in the ordinary component is Poisson-distributed:

$$J \sim \text{Poisson}(\lambda).$$

The $\{\xi_j\}_{j=1}^J$ are independent and identically distributed random variables with values in $[0, 1]$ such that each has density $\nu(db)/\lambda$.

Next, we note that instead of general sets M and G, we can restrict to sets of the form

$$M' = \{J = \hat{J}\} \cap \bigcap_{j=1}^{\hat{J}} \{v_j \leq \hat{v}_j, \xi_j \leq \hat{\xi}_j\}_{j=1}^{\hat{J}} \cap \bigcap_{l=1}^{L} \{\eta_l \leq \hat{\eta}_l\}. \tag{4.24}$$

$$G' = \{K = 1\} \cap \{i_1 = \hat{i}_1, s_1 \leq \hat{s}_1\}. \tag{4.25}$$

That is, in the random measure B_{prior} case, we consider a set with a fixed number $\hat{J}$ of ordinary component atoms and with fixed upper bounds $\hat{v}_j$, $\hat{\xi}_j$, or $\hat{\eta}_l$ on, respectively, the location of the jth ordinary component atom, the weight of the jth ordinary component atom, and the weight of the lth fixed atom. In the counting measure I case, we can restrict to a single atom with location bounded by $\hat{s}_1$ and count equal to $\hat{i}_1 \in \{1, 2, \ldots\}$.

With this notation and restriction in hand, we proceed to compute the prior, marginal, and posterior so that we may check whether Eq. (4.22) holds.

Prior. We first calculate the prior measure of set M'. Recall that the number of atoms is Poisson-distributed:

$$\mathbb{P}_{prior}(J = \hat{J}) = \frac{\lambda^{\hat{J}}}{\hat{J}!} e^{-\lambda}. \tag{4.26}$$

Also, the locations of these atoms, given their number, are distributed as

$$\mathbb{P}_{prior}(\bigcap_{j=1}^{\hat{J}} \{v_j \leq \hat{v}_j\} | J = \hat{J}) = \hat{J}! \int_{\psi_1=0}^{\hat{v}_1} \int_{\psi_2=\psi_1}^{\hat{v}_2} \cdots \int_{\psi_{\hat{J}}=\psi_{\hat{J}-1}}^{\hat{v}_{\hat{J}}} \left(\prod_{j=1}^{\hat{J}} d\psi_j \right). \tag{4.27}$$

The $\hat{J}!$ term results from the fact that the v_j are, by construction, the order statistics of a collection of uniformly distributed random variables. Finally, the sizes of the atoms, given their location and number, have the distribution

$$\mathbb{P}_{prior}\left(\bigcap_{j=1}^{J} \{\xi_j \leq \hat{\xi}_j\} \cap \bigcap_{l=1}^{L} \{\eta_l \leq \hat{\eta}_l\} | J = \hat{J}, \bigcap_{j=1}^{\hat{J}} \{v_j \leq \hat{v}_j\} \right) = \left[\prod_{j=1}^{\hat{J}} \int_{b=0}^{\hat{\xi}_j} \frac{\nu(db)}{\lambda} \right] \cdot \left[\prod_{l=1}^{L} \int_{b=0}^{\hat{\eta}_l} h_l(db) \right]. \tag{4.28}$$

Together, Eqs. (4.26), (4.27), and (4.28) yield the prior probability of the set M' (Eq. (4.24)) describing the random measure B_{prior}.

Next, we turn to the prior probability of the set G' describing the counting measure I. In this case, we condition on a particular measure $\mu \in M'$. Now, in G', each counting measure I has exactly one atom. This atom can occur either at an atom in the ordinary component of μ, located at one of $\{v_j\}_{j=1}^{J}$, or at a fixed atom of μ, located at one of $\{u_l\}_{l=1}^{L}$. We take advantage of the fact that the u_l are unique by assumption and that the v_j are a.s. unique and distinct from the u_l by the assumption that the distribution on locations is continuous. We also note that on the set $\{s_1 \leq \hat{s}_1\}$, we need only consider those atoms with locations at most $\hat{s}_1$. Thus, we break into these two special cases as follows:

$$\mathbb{P}_{prior}(K = 1, i_1 = \hat{i}_1, s_1 \leq \hat{s}_1 | \mu) = \sum_{j=1}^{J} \mathbb{P}_{prior}(K = 1, i_1 = \hat{i}_1, s_1 = v_j | \mu) \mathbb{1}\{v_j \leq \hat{s}_1\}$$

$$+\sum_{l=1}^{L} \mathbb{P}_{prior}(K=1, i_1=\hat{i}_1, s_1=u_l|\mu)\mathbb{1}\{u_l \leq \hat{s}_1\}.$$

The probability that the single nonzero count occurs at a particular atom is the probability that a nonzero count appears at this atom and zero counts appear at all other atoms. To express this probability, we first define a new function:

$$\begin{aligned}\Phi(J, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, i_1, s) = &\left\{\prod_{j=1}^{J} [\text{NegBin}(0|r, \xi_j)]^{\mathbb{1}\{v_j \neq s\}} [\text{NegBin}(i_1|r, \xi_j)]^{\mathbb{1}\{v_j = s\}}\right\} \\ &\cdot \left\{\prod_{l=1}^{L} [\text{NegBin}(0|r, \eta_j)]^{\mathbb{1}\{u_l \neq s\}} [\text{NegBin}(i_1|r, \eta_l)]^{\mathbb{1}\{u_l = s\}}\right\}.\end{aligned}$$

Here, $\text{NegBin}(x|a, b)$ is the negative binomial density. A notable special case is $\text{NegBin}(0|a, b) = (1-b)^a$. We can write the single-atom probabilities with the Φ notation:

$$\begin{aligned}\mathbb{P}_{prior}(K=1, i_1=\hat{i}_1, s_1=v_j|\mu) &= \Phi(J, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, i_1, v_j) \\ \mathbb{P}_{prior}(K=1, i_1=\hat{i}_1, s_1=u_l|\mu) &= \Phi(J, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, i_1, u_l).\end{aligned}$$

We can combine the likelihood of the counting process I given the random measure B_{prior} with the prior of the random measure B_{prior} to find the joint prior probability of the set $M' \times G'$. If we use the following notation to express the sets over which we will integrate,

$$\begin{aligned}R(\hat{\mathbf{v}}, J) &\triangleq \{\boldsymbol{\psi} : \boldsymbol{\psi} \in [0,1]^J, \psi_1 \leq \cdots \leq \psi_J\} \cap \bigcap_{j=1}^{J} \{\psi : \psi_j \leq \hat{v}_j\} \\ r(\mathbf{T} = (t_1, \ldots, t_J), J) &\triangleq [0, t_1] \times \cdots \times [0, t_J],\end{aligned}$$

then we may write

$$\begin{aligned}\mathbb{P}_{prior}(M' \times G') &= \int_{B \in M'} \mathbb{P}_{prior}(G'|B)\, d\mathbb{P}_{prior}(B) \\ &= e^{-\lambda} \left\{ \sum_{j=1}^{\hat{J}} \left[\int_{\mathbf{v} \in R(\hat{\mathbf{v}}, \hat{J}), \boldsymbol{\xi} \in r(\hat{\boldsymbol{\xi}}, \hat{J}), \boldsymbol{\eta} \in r(\hat{\boldsymbol{\eta}}, L)} \mathbb{1}\{v_j \leq \hat{s}_1\} \right.\right. \\ &\left. \cdot \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, i_1, v_j) \cdot \left(\prod_{j=1}^{\hat{J}} dv_j\right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi)\right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l)\right)\right] \\ &+ \sum_{l=1}^{L} \left[\int_{\mathbf{v} \in R(\hat{\mathbf{v}}, \hat{J}), \boldsymbol{\xi} \in r(\hat{\boldsymbol{\xi}}, \hat{J}), \boldsymbol{\eta} \in r(\hat{\boldsymbol{\eta}}, L)} \mathbb{1}\{u_l \leq \hat{s}_1\} \right.\end{aligned}$$

$$\cdot \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, \hat{i}_1, u_l) \cdot \left(\prod_{j=1}^{\hat{J}} dv_j \right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi) \right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l) \right) \Bigg] \Bigg\} . \tag{4.29}$$

This equation completes our prior calculation for now. We will return to it when we evaluate Eq. (4.22) for sets M' and G'.

Proposed posterior. Next we consider the proposed posterior distribution $\mathbb{Q}$. Just as we calculated the probability of $M' \times G'$ under the measure induced by our prior generative model, we can analogously calculate the quantity $\mathbb{Q}(M' : I)$ for some $I \in G'$ according to the definition of $\mathbb{Q}$.

In the theorem statement, we specified a construction of a completely random measure to induce the proposed posterior. In this case, the completely random measure has an ordinary component and a set of fixed atoms. Given the specific set G' we are considering (Eq. (4.25)), the set of locations of the fixed atoms is $\{u_1, \ldots, u_L\} \cup \{\hat{s}_1\}$, where the union is not necessarily disjoint. So there are two cases we must examine: either the counting process atom is at the same location as a fixed atom of the prior random measure ($\hat{s}_1 = u_l$ for some $l \in \{1, \ldots, L\}$), or it is at a different location ($\hat{s}_1 \notin \{u_1, \ldots, u_L\}$).

First, we consider the case where the counting process atom location $\hat{s}_1$ is the same as that of a fixed atom of the prior random measure, say u_{l^*}. As before, the number of atoms in the ordinary component is Poisson-distributed with mean equal to the total Poisson point process mass

$$\lambda_{post} \triangleq \int_{b=0}^{1} (1-b)^r \nu(db).$$

So we have (cf. Eq. (4.26))

$$\mathbb{Q}(J = \hat{J} : K = 1, i_1 = \hat{i}_1, s_1 = u_{l^*}) = \frac{\lambda_{post}^{\hat{J}}}{\hat{J}!} e^{-\lambda_{post}}. \tag{4.30}$$

Also, as in the case of Eq. (4.27), we can calculate the distribution of the locations of the ordinary component atoms:

$$\mathbb{Q}(\bigcap_{j=1}^{\hat{J}} \{v_j \le \hat{v}_j\} | J = \hat{J} : K = 1, i_1 = \hat{i}_1, s_1 = u_{l^*}) = \hat{J}! \int_{\psi_1=0}^{\hat{v}_1} \int_{\psi_2=\psi_1}^{\hat{v}_2} \cdots \int_{\psi_{\hat{J}}=\psi_{\hat{J}-1}}^{\hat{v}_{\hat{J}}} \left(\prod_{j=1}^{\hat{J}} d\psi_j \right) . \tag{4.31}$$

And again, as in Eq. (4.28), the sizes of the atoms, given their location and number, have the distribution

$$\mathbb{Q} \left(\bigcap_{j=1}^{J} \{\xi_j \le \hat{\xi}_j\} \cap \bigcap_{l=1}^{L} \{\eta_l \le \hat{\eta}_l\} | J = \hat{J}, \bigcap_{j=1}^{\hat{J}} \{v_j \le \hat{v}_j\} : K = 1, i_1 = \hat{i}_1, s_1 = u_{l^*} \right)$$

$$= \left[\prod_{j=1}^{\hat{J}} \int_{b=0}^{\hat{\xi}_j} \frac{\mathrm{NegBin}(0|r,b)\nu(db)}{\lambda_{post}} \right] \left[\prod_{l=1}^{L} \frac{\int_{b=0}^{\hat{\eta}_l} [\mathrm{NegBin}(\hat{i}_1|r,b)]^{\mathbb{1}\{l=l^*\}} [\mathrm{NegBin}(0|r,b)]^{\mathbb{1}\{l\neq l^*\}} h_l(db)}{\int_{b=0}^{1} [\mathrm{NegBin}(\hat{i}_1|r,b)]^{\mathbb{1}\{l=l^*\}} [\mathrm{NegBin}(0|r,b)]^{\mathbb{1}\{l\neq l^*\}} h_l(db)} \right]. \tag{4.32}$$

Putting together Eqs. (4.30), (4.31), and (4.32), we can find the proposed measure of the set M' given $I \in G'$ for the case $\hat{s}_1 = u_{l^*}$:

$$\mathbb{Q}(M' : I) = \mathbb{Q}\left(J = \hat{J}, \bigcap_{j=1}^{\hat{J}} \{v_j \leq \hat{v}_j\}, \bigcap_{j=1}^{J} \{\xi_j \leq \hat{\xi}_j\} \cap \bigcap_{l=1}^{L} \{\eta_l \leq \hat{\eta}_l\} : K = 1, i_1 = \hat{i}_1, s_1 = u_{l^*} \right)$$

$$= C_{fixed,l^*}^{-1} e^{-\lambda_{post}} \int_{\mathbf{v} \in R(\hat{\mathbf{v}},\hat{J}), \boldsymbol{\xi} \in r(\hat{\boldsymbol{\xi}},\hat{J}), \boldsymbol{\eta} \in r(\hat{\boldsymbol{\eta}},L)} \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, i_1, u_{l^*}) \tag{4.33}$$

$$\cdot \left(\prod_{j=1}^{\hat{J}} dv_j \right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi) \right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l) \right), \tag{4.34}$$

where

$$C_{fixed,l^*} \triangleq \prod_{l=1}^{L} \int_{b=0}^{1} [\mathrm{NegBin}(\hat{i}_1|r,b)]^{\mathbb{1}\{l=l^*\}} [\mathrm{NegBin}(0|r,b)]^{\mathbb{1}\{l\neq l^*\}} h_l(db).$$

Second, we consider the case $\hat{s}_1 \notin \{u_1, \ldots, u_L\}$. Then $\hat{s}_1 = v_{j^*}$ for some $j^* \in \{1, \ldots, J\}$. Recall that v_{j^*} is the j^*th smallest element of $\{v_1, \ldots, v_J\}$. We proceed as above and start by noting that the number of atoms on either side of the location v_{j^*} is Poisson-distributed:

$$\mathbb{Q}\left(J = \hat{J} : K = 1, i_1 = \hat{i}_1, s_1 = v_{j^*} \right)$$

$$= \frac{(\lambda_{post} v_{j^*})^{j^*-1}}{(j^*-1)!} e^{-\left(\lambda_{post} v_{j^*}\right)} \cdot \frac{(\lambda_{post}(1-v_{j^*}))^{(\hat{J}-j^*)}}{(\hat{J}-j^*)!} e^{-\left(\lambda_{post}(1-v_{j^*})\right)}. \tag{4.35}$$

Further, we have the usual distribution for the atom locations on either side of v_{j^*}:

$$\mathbb{Q}\left(\bigcap_{j=1}^{\hat{J}} \{v_j \leq \hat{v}_j\} | J = \hat{J} : K = 1, i_1 = \hat{i}_1, s_1 = v_{j^*} \right)$$

$$= (j^*-1)! \int_{\psi_1=0}^{\hat{v}_1} \int_{\psi_2=\psi_1}^{\hat{v}_2} \cdots \int_{\psi_{j^*}=\psi_{j^*-1}}^{\hat{v}_{j^*}} \left(\prod_{j=1}^{j^*-1} \frac{d\psi_j}{v_{j^*}} \right)$$

$$\cdot (\hat{J}-j^*)! \int_{\psi_{j^*+1}=\hat{v}_{j^*}}^{\hat{v}_{j^*+1}} \cdots \int_{\psi_{\hat{J}}=\psi_{\hat{J}-1}}^{\hat{v}_{\hat{J}}} \left(\prod_{j=j^*+1}^{\hat{J}} \frac{d\psi_j}{1-v_{j^*}} \right). \tag{4.36}$$

As usual, the third step identifies the conditional distribution of the atom weights:

$$\mathbb{Q}\left(\bigcap_{j=1}^{J}\{\xi_j \leq \hat{\xi}_j\} \cap \bigcap_{l=1}^{L}\{\eta_l \leq \hat{\eta}_l\} | J = \hat{J}, \bigcap_{j=1}^{\hat{J}}\{v_j \leq \hat{v}_j\} : K = 1, i_1 = \hat{i}_1, s_1 = v_{j^*}\right) \tag{4.37}$$

$$= \left[\prod_{j=1}^{\hat{J}} \frac{\int_{b=0}^{\hat{\xi}_j} \left[\mathrm{NegBin}(\hat{i}_1|r,b)\right]^{\mathbb{1}\{j=j^*\}} [\mathrm{NegBin}(0|r,b)]^{\mathbb{1}\{j\neq j^*\}} \nu(db)}{\int_{b=0}^{1} \left[\mathrm{NegBin}(\hat{i}_1|r,b)\right]^{\mathbb{1}\{j=j^*\}} [\mathrm{NegBin}(0|r,b)]^{\mathbb{1}\{j\neq j^*\}} \nu(db)}\right] \left[\prod_{l=1}^{L} \frac{\int_{b=0}^{\hat{\eta}_l} \mathrm{NegBin}(0|r,b) h_l(db)}{\int_{b=0}^{1} \mathrm{NegBin}(0|r,b) h_l(db)}\right]. \tag{4.38}$$

So, combining Eqs. (4.35), (4.36), and (4.38), we find that the proposed posterior distribution in the case $\hat{s}_1 = v_{j^*}$ is

$$\begin{aligned}\mathbb{Q}(M' : I) &= \mathbb{Q}\left(J = \hat{J}, \bigcap_{j=1}^{\hat{J}}\{v_j \leq \hat{v}_j\}, \bigcap_{j=1}^{J}\{\xi_j \leq \hat{\xi}_j\} \cap \bigcap_{l=1}^{L}\{\eta_l \leq \hat{\eta}_l\} : K = 1, N_1 = n_1, S_1 = \xi_{j^*}\right) \\ &= C_{ord}^{-1} e^{-\lambda_{post}} \int_{\mathbf{v} \in R(\hat{\mathbf{v}},\hat{J}), \xi \in r(\hat{\boldsymbol{\xi}},\hat{J}), \eta \in r(\hat{\boldsymbol{\eta}},L)} \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, \hat{i}_1, v_{j^*}) \\ &\quad \cdot \left(\prod_{j=1}^{\hat{J}} dv_j\right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi)\right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l)\right),\end{aligned} \tag{4.39}$$

where

$$C_{ord} \triangleq \left(\int_{b=0}^{1} \mathrm{NegBin}(\hat{i}_1|r,b)\nu(db)\right) \cdot \left(\prod_{l=1}^{L} \int_{b=0}^{1} \mathrm{NegBin}(0|r,b) h_l(db)\right).$$

Putting together the cases $\hat{s}_1 = u_{l^*}$ for some l^* (Eq. (4.34)) and $\hat{s}_1 \notin \{u_1, \ldots, u_L\}$ (Eq. (4.39)), we obtain the full proposed posterior distribution:

$$\begin{aligned}\mathbb{Q}(M' : I) &= \mathbb{Q}\left(J = \hat{J}, \bigcap_{j=1}^{\hat{J}}\{v_j \leq \hat{v}_j\}, \bigcap_{j=1}^{J}\{\xi_j \leq \hat{\xi}_j\} \cap \bigcap_{l=1}^{L}\{\eta_l \leq \hat{\eta}_l\} : K = 1, i_1 = \hat{i}_1, s_1 = \hat{s}_1\right) \\ &= \sum_{l^*=1}^{L} \mathbb{1}\{\hat{s}_1 = u_{l^*}\}\, C_{fixed,l^*}^{-1} e^{-\lambda_{post}} \int_{\mathbf{v} \in R(\hat{\mathbf{v}},\hat{J}), \boldsymbol{\xi} \in r(\hat{\boldsymbol{\xi}},\hat{J}), \boldsymbol{\eta} \in r(\hat{\boldsymbol{\eta}},L)} \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, \hat{i}_1, u_{l^*}) \\ &\quad \cdot \left(\prod_{j=1}^{\hat{J}} dv_j\right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi)\right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l)\right) \\ &\quad + \mathbb{1}\left\{\hat{s}_1 \notin \{u_1, \ldots, u_L\}\right\}\, C_{ord}^{-1} e^{-\lambda_{post}} \int_{\mathbf{v} \in R(\hat{\mathbf{v}},\hat{J}), \boldsymbol{\xi} \in r(\hat{\boldsymbol{\xi}},\hat{J}), \eta \in r(\hat{\boldsymbol{\eta}},L)} \Phi(\hat{J}, L, \mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\eta}, \hat{i}_1, v_{j^*}) \\ &\quad \cdot \left(\prod_{j=1}^{\hat{J}} dv_j\right) \cdot \left(\prod_{j=1}^{\hat{J}} \nu(d\xi)\right) \cdot \left(\prod_{l=1}^{L} h_l(d\eta_l)\right).\end{aligned} \tag{4.40}$$

Counting process marginal. With the prior and proposed posterior in hand, it remains to calculate the marginal distribution of the counting process. Then we may integrate out the proposed posterior with respect to the counting process marginal in order to obtain the prior (Eq. (4.22)). Since we are focusing on counting process sets G' of the form in Eq. (4.25), we aim to calculate

$$\mathbb{P}_{marg}(K = 1, i_1 = \hat{i}_1, s_1 \leq \hat{s}_1).$$

In our calculations above, we also worked with a set of prior measure $\mu \in M'$ and therefore worked with a set of locations for the ordinary component atoms. In this case, we will need to calculate the probability of zero counts in an interval where the number and location of the ordinary component atoms is integrated out. Let $I'\{\psi\}$ be the counting process that includes exactly those counts at ordinary component atoms and not the counts at fixed atoms; we can see, e.g., that $I'\{\psi\} \leq I\{\psi\}$ at all ψ. Further, similar to Eq. (4.23), let B_{ord} be the random measure composed only of those atoms in the ordinary component of B_{prior}:

$$B_{ord} = \sum_{j=1}^{J} \xi_j \delta_{v_j}.$$

Then we are interested in the quantity:

$$\mathbb{E}\left[\mathbb{1}\{\forall t \in (\psi_1, \psi_2), I'\{t\} = 0\}\right] = \mathbb{E}\left[\prod_{t \in (\psi_1, \psi_2)} (1 - B_{ord}\{t\})^r\right] = \prod_{t \in (\psi_1, \psi_2)} \left(1 - \mathbb{E}\left[1 - (1 - B_{ord}\{t\})^r\right]\right),$$

where the last equality follows from the independence of B_{prior} across increments.

Now define a new process $B' \triangleq 1 - (1 - B_{ord})^r$. This process has intensity ν', which can be obtained by a change of variables from the Poisson process intensity ν of B_{ord}. We will find it notationally useful to refer to ν' though we do not calculate it here. Also, let $\bar{B}'$ be the mean process of B': $\bar{B}'(d\psi) \triangleq \mathbb{E}\left[B'(d\psi)\right]$. With this notation in hand, we can write

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{\forall t \in (\psi_1, \psi_2), I'\{t\} = 0\}\right] \\
&= \prod_{t \in (\psi_1, \psi_2)} \left(1 - \bar{B}'\{t\}\right) = \exp\left\{-\int_{t=\psi_1}^{\psi_2} \bar{B}'\{t\}\right\} = \exp\left\{-\int_{t=\psi_1}^{\psi_2} \int_{b=0}^{1} b' \, \nu'(db')\right\} \\
&= \exp\left\{-(\psi_2 - \psi_1) \int_{b=0}^{1} \left(1 - (1 - b)^r\right) \nu(db)\right\}.
\end{aligned}$$

As usual, we consider two separate cases. First, suppose $s_1 = u_{l^*}$ for some $l^* \in \{1, \ldots, L\}$. Then using the result above we find

$$\mathbb{P}_{marg}(K = 1, i_1 = \hat{i}_1, s_1 = u_{l^*}) = \mathbb{P}_{marg}(I\{u_{l^*}\} = \hat{i}_1) \; \mathbb{P}_{marg}(\forall l \neq l^*, I\{u_l\} = 0) \; \mathbb{P}_{marg}(\forall t \in (0, 1), I'\{t\} = 0)$$

$$= \left(\prod_{l=1}^{L} \int_{b=0}^{1} [\text{NegBin}(0|r,b)]^{\mathbb{1}\{l \neq l^*\}} \left[\text{NegBin}(\hat{i}_1|r,b)\right]^{\mathbb{1}\{l=l^*\}} h_l(db) \right)$$
$$\cdot \exp\left\{ -(1-0) \int_{b=0}^{1} (1-(1-b)^r)\, \nu(db) \right\}$$
$$= e^{-\lambda+\lambda_{post}} C_{fixed,l^*}. \tag{4.41}$$

Next, suppose $s_1 \notin \{u_1, \ldots, u_L\}$. Then

$$\mathbb{P}_{marg}(K=1, i_1=\hat{i}_1, s_1 \notin \{u_1,\ldots,u_L\})$$
$$= \mathbb{P}_{marg}(\forall l, I(u_l)=0) \cdot \mathbb{P}_{marg}(\exists \psi : I'\{\psi\} = \hat{i}_1 \text{ and } \forall t \in (0,1)\backslash\{\psi\}, I'\{t\}=0)$$
$$= \left[\prod_{l=1}^{L} \int_{b=0}^{1} \text{NegBin}(0|r,b) h_l(db) \right] \cdot e^{-(\lambda-\lambda_{post})} \frac{(\lambda-\lambda_{post})^1}{1!} \cdot \frac{\int_{\psi=0}^{1} \left(\int_{b=0}^{1} \text{NegBin}(\hat{i}_1|r,b)\nu(db) \right) d\psi}{\int_{\psi=0}^{1} \left(\int_{b=0}^{1} \sum_{i=1}^{\infty} \text{NegBin}(i|r,b)\nu(db) \right) d\psi}$$
$$= \left[\prod_{l=1}^{L} \int_{b=0}^{1} \text{NegBin}(0|r,b) h_l(db) \right] \cdot e^{-(\lambda-\lambda_{post})} \cdot \left(\int_{b=0}^{1} \text{NegBin}(\hat{i}_1|r,b)\nu(db) \right) = e^{-\lambda+\lambda_{post}} C_{ord}. \tag{4.42}$$

Checking integration. The final step is to note that we may integrate out the proposed posterior in Eq. (4.40) with respect to the marginal described by Eqs. (4.41) and (4.42) to obtain the joint prior in Eq. (4.29). This integration is exactly the one we desired from Eq. (4.22) in the special case of sets of the form M' in Eq. (4.24) and G' in Eq. (4.25), as was to be shown. □

Infinite Poisson process intensity

Theorem 4.E.2. *Theorem 4.E.1 still applies when the intensity measure ν does not necessarily have a finite integral $\nu[0,1]$ but satisfies the (weaker) condition*

$$\int_{b=0}^{1} b\, \nu(db) < \infty. \tag{4.43}$$

Proof of Theorem 4.E.2: Note that Eq. (4.43) implies

$$\forall \epsilon > 0, \nu[\epsilon, \infty) < \infty. \tag{4.44}$$

The main idea behind the proof of Theorem 4.E.2 is to take advantage of the finiteness condition in Eq. (4.44) to construct a sequence of finite intensity measures tending to the true intensity measure of the process. We will use the known form of the posterior in the finite case from Theorem 4.E.1 to deduce the form of the posterior in the case where ν merely satisfies the weaker condition in Eq. (4.44).

We therefore start by defining the sequence of (finite) measures ν_n by

$$\nu_n(A) \triangleq \int_{b \in A} \mathbb{1}\{b > 1/n\} \nu(db), \quad \text{for all measurable } A \subset [0,1]. \tag{4.45}$$

Further, we may generate a random measure $B_{prior,n}$ as described by the prior in Theorem 4.E.1 with Poisson point process intensity ν_n. And we may generate a counting process I_n with parameters r and $B_{prior,n}$ as described in Theorem 4.E.1.

As before, let $\mathbb{P}_{prior}$ be the prior distribution on the prior random measure B_{prior} and the counting process I. Let $\mathbb{E}_{prior}$ denote the expectation with respect to this distribution. Further, let $\mathbb{P}_{marg}$ represent the marginal distribution on the counting process from $\mathbb{P}_{prior}$. And let $\mathbb{Q}(M : G)$ represent the proposed posterior distribution on sets $M \in \mathfrak{M}$ given any set $G \in \Sigma_{\mathfrak{G}}$. We use the same notation, but with n subscripts, to denote the case with finite intensity ν_n.

Our proof will take advantage of Laplacian-style characterizations of distributions. In particular, we note that in order to prove Theorem 4.E.2, it is enough to show that, for arbitrary continuous and nonnegative functions f and g (i.e., $f, g \in C^+[0,1]$), we have

$$\begin{aligned} &\int_{B \in \mathfrak{M}} \int_{I \in \mathfrak{G}} \exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\}) \right\} \, d\mathbb{Q}(B : I) \, d\mathbb{P}_{marg}(I) \\ &= \mathbb{E}_{prior}\left[\exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\}) \right\} \right]. \end{aligned} \tag{4.46}$$

By Lemma 4.E.3, we have the following limit for all $f, g \in C^+[0,1]$ as $n \to \infty$:

$$\begin{aligned} &\mathbb{E}_{prior,n}\left[\exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\})) \right\} \right] \\ &\to \mathbb{E}_{prior}\left[\exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\}) \right\} \right]. \end{aligned}$$

Therefore, by Eq. (4.46) and the observation that Theorem 4.E.1 holds under the finite intensity ν_n, we see that it is enough to show that

$$\begin{aligned} &\int_{B \in \mathfrak{M}_n} \int_{I \in \mathfrak{G}} \exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\}) \right\} \, d\mathbb{Q}_n(B : I) \, d\mathbb{P}_{marg,n}(I) \\ &\to \int_{B \in \mathfrak{M}} \int_{I \in \mathfrak{G}} \exp\left\{ -\int_{\psi=0}^{1} (g(\psi)B\{\psi\} + f(\psi)I\{\psi\}) \right\} \, d\mathbb{Q}(B : I) \, d\mathbb{P}_{marg}(I), \quad n \to \infty. \end{aligned} \tag{4.47}$$

Define

$$\Psi_n(I) \triangleq \int_{B \in \mathfrak{M}_n} \exp\left\{ -\int_{\psi=0}^{1} g(\psi)B\{\psi\} \right\} d\mathbb{Q}_n(B : I) \tag{4.48}$$

$$\Psi(I) \triangleq \int_{B \in \mathfrak{M}} \exp\left\{ - \int_{\psi=0}^{1} g(\psi) B\{\psi\} \right\} d\mathbb{Q}(B : I). \tag{4.49}$$

By Lemma 4.E.4, we have

$$\int_{I \in \mathfrak{G}} \exp\left\{ - \int_{\psi=0}^{1} f(\psi) I\{\psi\} \right\} (\Psi_n(I) - \Psi(I)) d\mathbb{P}_{marg,n}(I) \to 0. \tag{4.50}$$

And Lemma 4.E.3 together with the fact that $\exp\left\{ - \int_{\psi=0}^{1} f(\psi) I\{\psi\} \right\} \Psi(I)$ is a bounded function of I yields

$$\int_{I \in \mathfrak{G}} \exp\left\{ - \int_{\psi=0}^{1} f(\psi) I\{\psi\} \right\} \Psi(I) (d\mathbb{P}_{marg,n}(I) - d\mathbb{P}_{marg}(I)) \to 0. \tag{4.51}$$

Combining Eqs. (4.50) and (4.51) yields the desired limit in Eq. (4.47).

□

Lemma 4.E.3. *Let $B_{prior,n}$ be a completely random measure with a finite set of fixed atoms in $[0,1]$ and with the Poisson process intensity ν_n in Eq. (4.45), where ν satisfies Eq. (4.43). Let I_n be drawn as a negative binomial process with parameters r and $B_{prior,n}$. Similarly, let B_{prior} be a completely random measure with Poisson process intensity ν, and let I be drawn as a negative binomial process with parameters r and B_{prior}. Then*

$$(B_{prior,n}, I_n) \overset{d}{\to} (B_{prior}, I)$$

Proof of Lemma 4.E.3: It is enough to show that, for all $f, g \in C^{+}[0,1]$, we have

$$\begin{aligned} &\mathbb{E}_{prior,n}\left[\exp\left\{ - \int_{\psi=0}^{1} (g(\psi) B_{prior,n}\{\psi\} + f(\psi) I_n\{\psi\}) \right\} \right] \\ &\to \mathbb{E}_{prior}\left[\exp\left\{ - \int_{\psi=0}^{1} (g(\psi) B_{prior}\{\psi\} + f(\psi) I\{\psi\}) \right\} \right], \quad n \to \infty. \end{aligned}$$

We can construct a new completely random measure, $\hat{B}_n$, by keeping only those jumps from B_{prior} (generated with intensity ν) that are either at the fixed atom locations or have height at least $1/n$. Then $\hat{B}_n \overset{d}{=} B_{prior,n}$ for $B_{prior,n}$ generated with intensity ν_n. Let $\hat{I}_n$ be the counting process generated with parameters r and $\hat{B}_n$. Then it is enough to show

$$\begin{aligned} &\mathbb{E}_{prior}\left[\exp\left\{ - \int_{\psi=0}^{1} (g(\psi) \hat{B}_n\{\psi\} + f(\psi) \hat{I}_n\{\psi\}) \right\} \right] \\ &\to \mathbb{E}_{prior}\left[\exp\left\{ - \int_{\psi=0}^{1} (g(\psi) B_{prior}\{\psi\} + f(\psi) I\{\psi\}) \right\} \right], \quad n \to \infty. \end{aligned}$$

Let $\hat{B}_n^- = B_{prior} - \hat{B}_n$ be the completely random measure consisting only of an ordinary component with jumps of size less than $1/n$. Let $\hat{I}_n^-$ be a counting process with parameters r and $\hat{B}_n^-$. Then, using the independence of $\hat{B}_n$ and $\hat{B}_n^-$, we have

$$\begin{aligned}
&\mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)B_{prior}\{\psi\} + f(\psi)I\{\psi\})\right\}\right] \\
&= \mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)\hat{B}_n\{\psi\} + f(\psi)\hat{I}_n\{\psi\})\right\}\right] \cdot \mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right].
\end{aligned}$$

So it is enough to show that

$$\mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right] \to 1, \quad n \to \infty. \tag{4.52}$$

In order to show Eq. (4.52) holds, we establish the following upper bounds:

$$\mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right] \leq 1, \tag{4.53}$$

and

$$\int_{\psi=0}^{1}(g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\}) \leq (\max_{\psi} g(\psi))\hat{B}_n^-[0,1] + (\max_{\psi} f(\psi))\hat{I}_n^-[0,1].$$

Henceforth we use the shorthand $c \triangleq (\max_{\psi} g(\psi))$ and $c' \triangleq (\max_{\psi} f(\psi))$. These quantities are finite by the assumptions on g and f. Choose $\epsilon > 0$. Further define the events

$$A_B \triangleq \{\hat{B}_n^-[0,1] > \epsilon\} \quad \text{and} \quad A_I \triangleq \{\hat{I}_n^-[0,1] > \epsilon\}.$$

By Chebyshev's inequality,

$$\mathbb{P}(A_{B,n}) < \mathbb{E}\left[\hat{B}_n^-[0,1]\right]/\epsilon \quad \text{and} \quad \mathbb{P}(A_{I,n}) < \mathbb{E}\left[\hat{I}_n^-[0,1]\right]/\epsilon.$$

Using these definitions, we can write

$$\begin{aligned}
&\mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1}(g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right] \\
&\geq \mathbb{E}_{prior}\left[\exp\left\{-c\hat{B}_n^-[0,1] - c'\hat{I}_n^-[0,1]\right\}\right] \\
&\geq \mathbb{E}_{prior}\left[\mathbb{1}(A_{B,n}^C \cap A_{I,n}^C)\exp\left\{-c\hat{B}_n^-[0,1] - c'\hat{I}_n^-[0,1]\right\}\right] \\
&\geq \mathbb{P}_{prior}(A_{B,n}^C \cap A_{I,n}^C) \cdot \exp\left\{-c\epsilon - c'\epsilon\right\}.
\end{aligned} \tag{4.54}$$

Now $\mathbb{P}_{prior}(A_{B,n}^C \cap A_{I,n}^C) = 1 - \mathbb{P}_{prior}(A_{B,n} \cup A_{I,n})$. And

$$\mathbb{P}_{prior}(A_{B,n} \cup A_{I,n}) \leq \mathbb{P}_{prior}(A_{B,n}) + \mathbb{P}_{prior}(A_{I,n})$$

$$\leq \epsilon^{-1} \left\{ \mathbb{E}\left[\hat{B}_n^-[0,1]\right] + \mathbb{E}\left[\hat{I}_n^-[0,1]\right] \right\} \to 0, \quad n \to \infty,$$

where the last line follows by noting

$$\mathbb{E}\left[\hat{B}_n^-[0,1]\right] = \int_{b=0}^{1/n} b\nu(db) \to 0, \quad n \to \infty,$$

since ν is continuous and $\int_{b=0}^{1} b\nu(db) < \infty$ by assumption, and

$$\begin{aligned}
\mathbb{E}\left[\hat{I}_n^-[0,1]\right] &= \sum_{m=1}^{\infty} \int_{b=0}^{1/n} Cb^{-1}(1-b)^{\theta-1} \binom{m+r-1}{m} (1-b)^r b^m \, db \\
&\text{where } C \text{ is a constant in } n \text{ (cf. Eq. (4.14))} \\
&= C \sum_m (2/n)^m \int_0^{1/2} (\tilde{b})^{m-1} (1-(2/n)\tilde{b})^{r+\theta-1} \binom{m+r-1}{m} \, d\tilde{b} \\
&\leq C2^{r+\theta-1} \sum_m (2/n)^m \int_0^{1/2} \tilde{b}^{m-1} (1-\tilde{b})^{r+\theta-1} \binom{m+r-1}{m} \, d\tilde{b} \\
&\leq 2^{r+\theta}(1/n) \sum_{m=1}^{\infty} C \int_0^1 \tilde{b}^{m-1}(1-\tilde{b})^{r+\theta-1} \binom{m+r-1}{m} \, d\tilde{b} \\
&= 2^{r+\theta}(1/n)C',
\end{aligned}$$

where C' is a constant in n (by Lemma 4.C.4). The final line goes to zero as $n \to \infty$.

So $\mathbb{P}_{prior}(A_{B,n}^C \cap A_{I,n}^C) \to 1$ as $n \to \infty$, and the bound in Eq. (4.54) yields:

$$\lim_{n\to\infty} \mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1} (g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right] \geq \exp\left\{-c\epsilon - c'\epsilon\right\}.$$

Since this result is true for every $\epsilon > 0$, we must have

$$\lim_{n\to\infty} \mathbb{E}_{prior}\left[\exp\left\{-\int_{\psi=0}^{1} (g(\psi)\hat{B}_n^-\{\psi\} + f(\psi)\hat{I}_n^-\{\psi\})\right\}\right] \geq 1.$$

Together with Eq. (4.53), this equation gives the desired result. □

Lemma 4.E.4. *For Φ_n and Φ defined in, respectively, Eqs. (4.48) and (4.49), we have the limit in Eq. (4.50):*

$$\int_{I\in\mathfrak{G}} \exp\left\{-\int_{\psi=0}^{1} f(\psi)I\{\psi\}\right\} (\Psi_n(I) - \Psi(I)) d\mathbb{P}_{marg,n}(I) \to 0. \tag{4.55}$$

Proof of Lemma 4.E.4: We start by choosing n large enough so that (1) the difference between the ordinary components in the truncated case and the non-truncated case are, in some sense, small enough and (2) the number of atoms in the truncated case is bounded with high probability. Under these two conditions, we will then show that $\Psi_n(I)$ and $\Psi(I)$ are sufficiently close in value by examining in turn each of the various types of atoms in the proposed posterior.

Therefore, choose $\epsilon > 0$. First note that by the assumption of finite integration of ν (Eq. (4.43)) we can choose n_0 such that for all $n > n_0$ we have

$$\int_{b=0}^{1/n} b\nu(db) < \epsilon. \tag{4.56}$$

This choice implies the existence of n_1 such that for all $n > n_1$ and all $i \geq 1$ we have Eq. (4.56) as well as

$$\int_{b=0}^{1/n} b^i (1-b)^r \nu(db) < \epsilon. \tag{4.57}$$

Second, since $I \sim \mathbb{P}_{marg,n}$ approaches $I \sim \mathbb{P}_{marg}$ in distribution by Lemma 4.E.3, there exist constants K' and n_2 such that the number of atoms K_n of I_n satisfies

$$\mathbb{P}_{marg,n}(K_n > K') < \epsilon \quad \text{for all} \quad n > n_2. \tag{4.58}$$

Moreover, conditional on $K_n \leq K'$, there exists a constant $\tilde{i}$ such that, under any $\mathbb{P}_{marg,n}$, all counts in I are bounded above by $\tilde{i}$ with probability at least $1 - \epsilon$.

It remains to use these conditions to bound

$$\int_{I \in \mathfrak{G}} \exp\left\{ - \int_{\psi=0}^{1} f(\psi) I\{\psi\} \right\} (\Psi_n(I) - \Psi(I)) d\mathbb{P}_{marg,n}(I).$$

For instance, since $\Psi_n(I)$ and $\Psi(I)$ are both bounded between zero and one, we have that

$$\begin{aligned} &\left| \int_{I \in \mathfrak{G}} \exp\left\{ - \int_{\psi=0}^{1} f(\psi) I\{\psi\} \right\} (\Psi_n(I) - \Psi(I)) d\mathbb{P}_{marg,n}(I) \right| \\ &\leq 2\epsilon + 2\epsilon + \int_{\substack{I \in \mathfrak{G} \\ K_n \leq K' \\ \text{atoms bounded by } \tilde{i}}} |\Psi_n(I) - \Psi(I)| \, d\mathbb{P}_{marg,n}(I). \end{aligned} \tag{4.59}$$

Next, we need to bound the second term on the righthand side of Eq. (4.59). To that end, we break Ψ_n and Ψ into their three constituent parts: the fixed atoms from the prior, the new fixed atoms in the proposed posterior, and the ordinary component in the proposed posterior. For Ψ_n, we have

$$\Psi_n(I) = \int_{B \in \mathfrak{M}_n} \exp\left\{ - \int_{\psi=0}^{1} g(\psi) B\{\psi\} \right\} \, d\mathbb{Q}(B : I)$$

$$= \int_{B \in \mathfrak{M}_n} \exp \left\{ - \sum_{\psi : I\{\psi\} \geq 1, \psi \notin \{u_1, \ldots, u_L\}} g(\psi) B\{\psi\} - \sum_{l=1}^{L} g(u_l) B\{u_l\} - \int_{\psi=0}^{1} g(\psi) B_{ord}\{\psi\} \right\} \, d\mathbb{Q}(B : I)$$

$$= \left[\prod_{\psi : I\{\psi\} \geq 1, \psi \notin \{u_1, \ldots, u_L\}} \int_{B \in \mathfrak{M}_n} \exp \{ -g(\psi) B\{\psi\} \} \, d\mathbb{Q}(B : I) \right]$$

$$\cdot \left[\prod_{l=1}^{L} \int_{B \in \mathfrak{M}_n} \exp \{ -g(u_l) B\{u_l\} \} \, d\mathbb{Q}(B : I) \right] \left[\int_{B \in \mathfrak{M}_n} \exp \left\{ - \int_{\psi=0}^{1} g(\psi) B_{ord}\{\psi\} \right\} \, d\mathbb{Q}(B : I) \right]$$

by the independence of these components under $\mathbb{Q}(B : I)$

$$= \left[\prod_{\psi : I\{\psi\} \geq 1, \psi \notin \{u_1, \ldots, u_L\}} c_{new,n}^{-1} \int_{b=0}^{1} \exp\{-g(\psi) b\} b^{I\{\psi\}} (1-b)^r \nu_n(db) \right]$$

$$\cdot \left[\prod_{l=1}^{L} \int_{B \in \mathfrak{M}_n} \exp \{ -g(u_l) B\{u_l\} \} \, d\mathbb{Q}(B : I) \right] \left[\exp \left\{ - \int_{b=0}^{1} \int_{\psi=0}^{1} \left(1 - e^{-g(\psi) b} \right) (1-b)^r \, d\psi \, \nu_n(db) \right\} \right]$$

The final factor results from Campbell's theorem. The analogous formula holds for Ψ by removing the n subscripts.

With the formulas for Ψ_n and Ψ in hand, we turn again to our desired bound. We follow Lemma 3 of Kim (1999a) in using the following fact: for $x_1, \ldots, x_M, y_1, \ldots, y_M \in \mathbb{R}$ and $|x_m|, |y_m| \leq 1$ for all m, we have

$$\left| \prod_{m=1}^{M} x_m - \prod_{m=1}^{M} y_m \right| \leq \sum_{m=1}^{M} |x_m - y_m|.$$

In particular, we apply this inequality to transform the difference in Ψ_n and Ψ into separate differences in each component, where we note that the prior fixed atom component is shared and therefore disappears. First, for notational convenience, define

$$C_n(I, \psi) \triangleq \int_{b=0}^{1} \exp\{-g(\psi) b\} b^{I\{\psi\}} (1-b)^r \nu_n(db)$$
$$C(I, \psi) \triangleq \int_{b=0}^{1} \exp\{-g(\psi) b\} b^{I\{\psi\}} (1-b)^r \nu(db)$$

Then

$$\int_{\substack{I \in \mathfrak{G} \\ K_n \leq K' \\ \text{atoms bounded by } \tilde{i}}} |\Psi_n(I) - \Psi(I)| \, d\mathbb{P}_{marg}(I)$$

$$\leq \int_{\substack{I \in \mathfrak{G} \\ K_n \leq K' \\ \text{atoms bounded by } \tilde{i}}} \left\{ \left[\sum_{\psi : I\{\psi\} \geq 1, \psi \notin \{u_1, \ldots, u_L\}} \left| [c_{new,n}(I\{\psi\})]^{-1} C_n(I, \psi) - [c_{new}(I\{\psi\})]^{-1} C(I, \psi) \right| \right] \right.$$

$$+ \left| \exp\left\{ - \int_{b=0}^{1} \int_{\psi=0}^{1} \left(1 - e^{-g(\psi)b}\right)(1-b)^r \, d\psi \, \nu_n(db) \right\} - \exp\left\{ - \int_{b=0}^{1} \int_{\psi=0}^{1} \left(1 - e^{-g(\psi)b}\right)(1-b)^r \, d\psi \, \nu(d \right. \tag{4.60}$$

From Eq. (4.57), we can conclude both that $|c_{new,n}(I\{\psi\}) - c_{new}(I\{\psi\})| \leq \epsilon$ and that $|C_n(I,\psi) - C(I,\psi)| \leq \epsilon$. Also $c_{new,n}(I\{\psi\}) \geq C_n(I,\psi)$ and likewise without the n subscript. So

$$\begin{aligned}
&\left| [c_{new,n}(I\{\psi\})]^{-1} C_n(I,\psi) - [c_{new}(I\{\psi\})]^{-1} C(I,\psi) \right| \\
&\leq \left| [c_{new,n}(I\{\psi\})]^{-1} C_n(I,\psi) - [c_{new,n}(I\{\psi\})]^{-1} C(I,\psi) \right| + \left| [c_{new,n}(I\{\psi\})]^{-1} C(I,\psi) - [c_{new}(I\{\psi\})]^{-1} C(I, \right. \\
&\leq [c_{new,n}(I\{\psi\})]^{-1} \epsilon + C(I,\psi) [c_{new,n}(I\{\psi\})]^{-1} [c_{new}(I\{\psi\})]^{-1} \epsilon \\
&\leq 2\epsilon [c_{new,n}(I\{\psi\})]^{-1} \leq 2\epsilon [c_{new}(I\{\psi\})]^{-1} + 2\epsilon^2.
\end{aligned}$$

The difference in the two exponential terms in Eq. (4.60) is similarly at most ϵ. So for large enough n and hence small enough ϵ we have

$$\begin{aligned}
&\int_{\substack{I \in \mathfrak{G} \\ K_n \leq K' \\ \text{atoms bounded by } \tilde{i}}} |\Psi_n(I) - \Psi(I)| \, d\mathbb{P}_{marg}(I) \\
&\leq \epsilon K' \left[2\epsilon \left[c_{new}(\tilde{i}) \right]^{-1} + 2\epsilon^2 \right] + \epsilon
\end{aligned}$$

Together with Eq. (4.59), this bound completes the proof. □

4.F Posterior inference details

Exact Gibbs slice sampler

We sample $b_{d,k}$ and ψ_k from their Gibbs conditionals as follows:

Sample ψ_k**.** The conditional posterior of ψ_k given $\mathbf{z}_{\cdot,\cdot}$ and $\mathbf{x}_{\cdot,\cdot}$ is proportional to

$$H(d\psi_k) \prod_{d=1}^{D} \prod_{n=1}^{N_d} F(dx_{d,n} \mid \psi_k)^{\mathbb{I}(z_{d,n}=k)}.$$

This has a closed form when H is conjugate to $F(\psi_k)$ and may otherwise be sampled using a generic univariate sampling procedure (e.g., random-walk Metropolis-Hastings or slice sampling).

Sample $b_{d,k}$**.** By beta-negative binomial conjugacy, the conditional posterior of $b_{d,k}$ given $z_{d,\cdot}$ and $b_{0,k}$ is a beta distribution:

$$b_{d,k} \sim \text{Beta}(\gamma_d \theta_d b_{0,k} + N_{d,k}, \theta_d(1 - \gamma_d b_{0,k}) + r_d),$$

where $N_{d,k} \triangleq \sum_n \mathbb{I}(z_{d,n} = k)$.

Sample $b_{0,k}$. To sample the shared beta process weights $b_{0,k}$, we turn to the size-biased construction of the beta process introduced by Thibaux and Jordan, 2007:

$$B_0 = \sum_{m=0}^{\infty} \sum_{i=1}^{C_m} b_{0,m,i} \delta_{\psi_{m,i,\cdot}},$$

where

$$C_m \overset{indep}{\sim} \text{Poisson}\left(\frac{\theta_0 \gamma_0}{\theta_0 + m}\right), \quad b_{0,m,i} \overset{indep}{\sim} \text{Beta}(1, \theta_0 + m), \quad \text{and} \quad \psi_{m,i,\cdot} \overset{iid}{\sim} H.$$

If we order the atoms by the rounds in which they were drawn, then the kth atom overall was drawn in round m_k, where

$$m_k \triangleq \min \left\{ m : \sum_{j=0}^{m} C_j \geq k \right\}.$$

Conditional on the round indices $(m_k)_{k=1}^{\infty}$, we have

$$B_0 = \sum_{k=1}^{\infty} b_{0,k} \delta_{\psi_k}$$

for

$$b_{0,k} \overset{indep}{\sim} \text{Beta}(1, \theta_0 + m_k) \quad \text{and} \quad \psi_k \overset{iid}{\sim} H.$$

The conditional density of $b_{0,k}$ given the remaining variables is therefore proportional to

$$(1 - b_{0,k})^{\theta_0 + m_k - 1} \prod_{d=1}^{D} \frac{1}{\Gamma(\gamma_d \theta_d b_{0,k}) \Gamma(\theta_d (1 - \gamma_d b_{0,k}))} \left(\frac{b_{d,k}}{1 - b_{d,k}} \right)^{\gamma_d \theta_d b_{0,k}} \tag{4.61}$$

and may be sampled using random-walk Metropolis-Hastings.

It remains then to sample the latent round indices m_k or, equivalently, their differences $h_k \triangleq m_k - m_{k-1}$, where $m_0 \triangleq 0$ for notational convenience. Let f_m and F_m denote the pmf and cdf of the Poisson$(\frac{\theta_0 \gamma_0}{\theta_0 + m})$ distribution respectively, and define $C_{m,j} \triangleq \sum_{k=1}^{j} \mathbb{I}(m_k = m)$. Since $C_m = \sum_{k=1}^{\infty} \mathbb{I}(m_k = m) \sim$ Poisson$(\frac{\theta_0 \gamma_0}{\theta_0 + m})$, it follows that

$$\begin{aligned} &\mathbb{P}(h_k < 0 \mid (h_j)_{j=1}^{k-1}) = 0, \\ &\mathbb{P}(h_k = 0 \mid (h_j)_{j=1}^{k-1}) = \frac{1 - F_{m_{k-1}}(C_{m_{k-1},k-1})}{1 - F_{m_{k-1}}(C_{m_{k-1},k-1} - 1)} \end{aligned}$$

for $m_{k-1} = \sum_{j=1}^{k-1} h_j$, and

$$\mathbb{P}(h_k = h \mid (h_j)_{j=1}^{k-1}) = \frac{f_{m_{k-1}}(C_{m_{k-1},k-1})}{1 - F_{m_{k-1}}(C_{m_{k-1},k-1} - 1)} (1 - f_{m_{k-1}+h}(0)) \prod_{g=1}^{h-1} f_{m_{k-1}+g}(0)$$

for all $h \in \mathbb{N}$. The conditional distribution of h_k given $(h_j)_{j=1}^{k-1}$ and $b_{0,k}$ is then

$$p(h_k \mid (h_j)_{j=1}^{k-1}, b_{0,k}) \propto (1-b_{0,k})^{h_k}(\theta_0 + h_k + m_{k-1})p(h_k \mid (h_j)_{j=1}^{k-1}),$$

which cannot be normalized in closed form due to the infinite summation. To permit posterior sampling of h_k, we introduce an auxiliary variable v_k with conditional distribution

$$v_k \sim \text{Unif}(0, \zeta_{0,h_k}(1-b_{0,k})^{h_k}),$$

where $(\zeta_{0,h})_{h=1}^{\infty}$ is a fixed positive sequence with $\lim_{h\to\infty} \zeta_{0,h} = 0$. Given v_k, we may slice sample h_k from the finite distribution

$$p(h_k \mid (h_j)_{j=1}^{k-1}, b_{0,k}) \propto \frac{\mathbb{I}(v_k \le \zeta_{0,h_k}(1-b_{0,k})^{h_k})}{\zeta_{0,h_k}}(\theta_0 + h_k + m_{k-1})p(h_k \mid (h_j)_{j=1}^{k-1}).$$

Collapsed sampling

In Eq. (4.61), we sampled $b_{0,k}$ conditional on $\mathbf{b}_{\cdot,k}$. A more efficient alternative is to integrate $\mathbf{b}_{\cdot,k}$ out of this conditional. We exploit the conjugacy of the beta and negative binomial distributions to derive the conditional distribution of $N_{d,k}$ given $b_{0,k}$, γ_d, θ_d, and r_d:

$$\begin{aligned} p(N_{d,k} \mid b_{0,k}, \gamma_d, \theta_d, r_d) &= \int p(N_{d,k} \mid b_{d,k}, r_d) p(b_{d,k} \mid b_{0,k}, \gamma_d, \theta_d) db_{d,k} \\ &= \int \frac{\Gamma(N_{d,k} + r_d)}{N_{d,k}!\ \Gamma(r_d)} \frac{\Gamma(\theta_d) b_{d,k}^{N_{d,k}+\gamma_d\theta_d b_{0,k}-1}(1-b_{d,k})^{r_d+\theta_d(1-\gamma_d b_{0,k})-1}}{\Gamma(\gamma_d\theta_d b_{0,k})\ \Gamma(\theta_d(1-\gamma_d b_{0,k}))} db_{d,k} \\ &= \frac{\Gamma(N_{d,k} + r_d)}{N_{d,k}!\ \Gamma(r_d)} \frac{\Gamma(\theta_d)\ \Gamma(N_{d,k}+\gamma_d\theta_d b_{0,k})\ \Gamma(r_d + \theta_d(1-\gamma_d b_{0,k}))}{\Gamma(N_{d,k}+r_d+\theta_d)\ \Gamma(\gamma_d\theta_d b_{0,k})\ \Gamma(\theta_d(1-\gamma_d b_{0,k}))}. \end{aligned}$$

The conditional density of $b_{0,k}$ with $\mathbf{b}_{\cdot,k}$ integrated out now takes the form

$$(1-b_{0,k})^{\theta_0+m_k-1} \prod_{d=1}^{D} \frac{\Gamma(N_{d,k}+\gamma_d\theta_d b_{0,k})\ \Gamma(r_d+\theta_d(1-\gamma_d b_{0,k}))}{\Gamma(\gamma_d\theta_d b_{0,k})\ \Gamma(\theta_d(1-\gamma_d b_{0,k}))}$$

and may be sampled using random-walk Metropolis-Hastings.

Finite approximation Gibbs sampler

The full conditional distribution of $b_{0,k}$ under the finite approximation of Eq. (4.10) is proportional to

$$b_{0,k}^{\theta_0\gamma_0/K-1}(1-b_{0,k})^{\theta_0(1-\gamma_0/K)-1} \prod_{d=1}^{D} \frac{1}{\Gamma(\gamma_d\theta_d b_{0,k})\Gamma(\theta_d(1-\gamma_d b_{0,k}))} \left(\frac{b_{d,k}}{1-b_{d,k}}\right)^{\gamma_d\theta_d b_{0,k}},$$

while the conditional density with $\mathbf{b}_{\cdot,k}$ integrated out is proportional to

$$b_{0,k}^{\theta_0\gamma_0/K-1}(1-b_{0,k})^{\theta_0(1-\gamma_0/K)-1}\prod_{d=1}^{D}\frac{\Gamma(N_{d,k}+\gamma_d\theta_d b_{0,k})\;\Gamma(r_d+\theta_d(1-\gamma_d b_{0,k}))}{\Gamma(\gamma_d\theta_d b_{0,k})\;\Gamma(\theta_d(1-\gamma_d b_{0,k}))}.$$

Random-walk Metropolis-Hastings may be used to sample $b_{0,k}$ from either distribution.

With this approximation in hand, we sample $\lambda_{d,k}$, $b_{d,k}$, and ψ_k precisely as described in Section 4.7. Since the number of components is finite, no auxiliary slice variables are needed to sample the component indices. Hence, we may sample $z_{d,n}$ from its discrete conditional distribution

$$\mathbb{P}(z_{d,n}=k)\propto F(dx_{d,n}\mid\psi_k)\lambda_{d,k}$$

given the remaining variables.

Chapter 5

Feature allocations, probability functions, and paintboxes

The problem of inferring a clustering of a data set has been the subject of much research in Bayesian analysis, and there currently exists a solid mathematical foundation for Bayesian approaches to clustering. In particular, the class of probability distributions over partitions of a data set has been characterized in a number of ways, including via exchangeable partition probability functions (EPPFs) and the Kingman paintbox. Here, we develop a generalization of the clustering problem, called feature allocation, where we allow each data point to belong to an arbitrary, non-negative integer number of groups, now called features or topics. We define and study an "exchangeable feature probability function" (EFPF)—analogous to the EPPF in the clustering setting—for certain types of feature models. Moreover, we introduce a "feature paintbox" characterization—analogous to the Kingman paintbox for clustering—of the class of exchangeable feature models. We provide a further characterization of the subclass of feature allocations that have EFPF representations.

5.1 Introduction

Exchangeability has played a key role in the development of Bayesian analysis in general and Bayesian nonparametric analysis in particular. Exchangeability can be viewed as asserting that the indices used to label the data points are irrelevant for inference, and as such is often a natural modeling assumption. Under such an assumption, one is licensed by de Finetti's theorem (De Finetti, 1931; Hewitt and Savage, 1955) to propose the existence of an underlying parameter that renders the data conditionally independent and identically distributed (iid) and to place a prior distribution on that parameter. Moreover, the theory of infinitely exchangeable sequences has advantages of simplicity over the theory of finite exchangeability, encouraging modelers to take a nonparametric stance in which the underlying "parameter" is infinite dimensional. Finally, the development of algorithms for posterior inference is often greatly simplified by the assumption of exchangeability, most notably in the case of Bayesian nonparametrics, where models based on the Dirichlet process and other combinatorial priors became useful tools in practice only when it was realized how to exploit exchangeability to develop inference procedures (Escobar, 1994).

The connection of exchangeability to Bayesian nonparametric modeling is well established in the case of models for clustering. The goal of a clustering procedure is to infer a partition of the data points. In the Bayesian setting, one works with random partitions, and, under an exchangeability assumption, the distribution on partitions should be invariant to a relabeling of the data points. The notion of an exchangeable random partition has been formalized by Kingman, Aldous, and others (Kingman, 1978; Aldous, 1985), and has led to the definition of an *exchangeable partition probability function* (EPPF) (Pitman, 1995). The EPPF is a mathematical function of the cardinalities of the groups in a partition. Exchangeability of the random partition is captured by the requirement that the EPPF be a symmetric function of these cardinalities. Furthermore, the exchangeability of a partition can be related to the exchangeability of a sequence of random variables representing the assignments of data points to clusters, for which a de Finetti mixing measure necessarily exists. This de Finetti measure is known as the *Kingman paintbox* (Kingman, 1978). The relationships among this circle of ideas are well understood: it is known that there is an equivalence among the class of exchangeable random partitions, the class of random partitions that possess an EPPF, and the class of random partitions generated by a Kingman paintbox; see Pitman (2006) for an overview of these relations. A specific example of these relationships is given by the Chinese restaurant process and the Dirichlet process, but several other examples are known and have proven useful in Bayesian nonparametrics.

Our focus in the current chapter is on an alternative to clustering models that we refer to as *feature allocation models*. While in a clustering model each data point is assigned to one and only one class, in a feature allocation model each data point can belong to multiple groups. It is often natural to view the groups as corresponding to traits or features, such that the notion that a data point belongs to multiple groups corresponds to the point exhibiting multiple traits or features. A Bayesian feature allocation model treats the feature assignments for a given data point as random and subject to posterior inference. A nonparametric Bayesian feature allocation model takes the number of features to also be random and subject to inference.

Research on nonparametric Bayesian feature allocation has been based around a single prior distribution, the Indian buffet process of Griffiths and Ghahramani (2006), which is known to have the beta process as its underlying de Finetti measure (Thibaux and Jordan, 2007). There does not yet exist a general definition of exchangeability for feature allocation models, nor counterparts of the EPPF or the Kingman paintbox.

In this chapter we supply these missing constructions. We provide a rigorous treatment of exchangeable feature allocations (in Section 8.3 and Section 5.3). In Section 5.4 we define a notion of *exchangeable feature probability function* (EFPF) that is the analogue for feature allocations of the EPPF for clustering. We then proceed to define a *feature paintbox* in Section 5.5. Finally, in Section 5.6 we discuss a class of models that we refer to as *feature frequency models* for which the construction of the feature paintbox is particularly

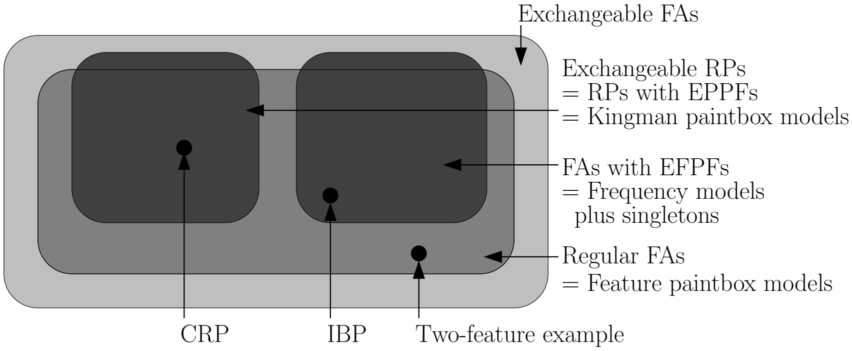


Figure 5.1: A summary of the relations described in this chapter. Rounded rectangles represent classes with the following abbreviations: RP for random partition, FA for random feature allocation, EPPF for exchangeable partition probability function, EFPF for exchangeable feature probability function. The large black dots represent particular models with the following abbreviations: CRP for Chinese restaurant process, IBP for Indian buffet process. The two-feature example refers to Example 5.4.4 with the choice $p_{11}p_{00} \neq p_{10}p_{01}$.

straightforward, and we discuss the important role that feature frequency models play in the general theory of feature allocations.

The Venn diagram shown in Figure 5.1 is a useful guide for understanding our results, and the reader may wish to consult this diagram in working through the chapter. As shown in the diagram, random partitions (RPs) are a special case of random feature allocations (FAs), and previous work on random partitions can be placed within our framework. Thus, in the diagram, we have depicted the equivalence already noted of exchangeable RPs, RPs that possess an EPPF, and Kingman paintboxes. We also see that random feature allocations have a somewhat richer structure: the class of FAs with EFPFs is not the same as those having an underlying feature paintbox. But the class of EFPFs is characterized in a different way; we will see that the class of feature allocations with EFPFs is equivalent to the class of FAs obtained from feature frequency models together with singletons of a certain distribution. Indeed, we will find that the class of clusterings with EPPFs is, in this way, analogous to the class of feature allocations with EFPFs when both are considered as subclasses of the general class of feature allocations. The diagram also shows several examples that we use to illustrate and develop our theory.

5.2 Feature allocations

We consider data sets with N points and let the points be indexed by the integers $[N] := \{1, 2, \ldots, N\}$. We also explicitly allow $N = \infty$, in which case the index set is $\mathbb{N} = \{1, 2, 3, \ldots\}$. For our discussion of feature allocations and partitioning it is sufficient to focus on the indices rather than the data points; thus, we will be discussing models for collections of subsets of $[N]$ and $\mathbb{N}$.

Our introduction to feature allocations follows Broderick, Jordan, and Pitman (2013). We define a *feature allocation* f_N of $[N]$ to be a multiset of non-empty subsets of $[N]$ called *features*, such that no index n belongs to infinitely many features. We write $f_N = \{A_1, \ldots, A_K\}$, where K is the number of features. An example feature allocation of $[6]$ is $f_6 = \{\{2,3\}, \{2,4,6\}, \{3\}, \{3\}, \{3\}\}$. Similarly, a feature allocation f_∞ of $\mathbb{N}$ is a multiset of non-empty subsets of $\mathbb{N}$ such that no index n belongs to infinitely many features. The total number of features in this case may be infinite, in which case we write $f_\infty = \{A_1, A_2, \ldots\}$. An example feature allocation of $\mathbb{N}$ is $f_\infty = \{\{n : n \text{ is prime}\}, \{n : n \text{ is not divisible by two}\}\}$. Finally, we may have $K = 0$, and $f_\infty = \emptyset$ is a valid feature allocation.

A *partition* is a special case of a feature allocation for which the features are restricted to be mutually exclusive and exhaustive. The features of a partition are often referred to as *blocks* or *clusters*. We note that a partition is always a feature allocation, but the converse statement does not hold in general; neither of the examples given above (f_6 and f_∞) are partitions.

We now turn to the problem of defining exchangeable feature allocations, extending previous work on exchangeable random partitions (Aldous, 1985). Let $\mathcal{F}_N$ be the space of all feature allocations of $[N]$. A *random feature allocation* F_N of $[N]$ is a random element of $\mathcal{F}_N$. Let $\sigma : \mathbb{N} \to \mathbb{N}$ be a finite permutation. That is, for some finite value N_σ, we have $\sigma(n) = n$ for all $n > N_\sigma$. Further, for any feature $A \subset \mathbb{N}$, denote the permutation applied to the feature as follows: $\sigma(A) := \{\sigma(n) : n \in A\}$. For any feature allocation F_N, denote the permutation applied to the feature allocation as follows: $\sigma(F_N) := \{\sigma(A) : A \in F_N\}$. Finally, let F_N be a random feature allocation of $[N]$. Then we say that a random feature allocation F_N is *exchangeable* if $F_N \overset{d}{=} \sigma(F_N)$ for every permutation of $[N]$.

In addition to exchangeability, we also require our distributions on feature allocations to exhibit a notion of coherence across different ranges of the index. Intuitively, we often imagine the indices as denoting time, and it is natural to suppose that the randomness at time n is coherent with the randomness at time $n + 1$. More formally, we say that a feature allocation f_M of $[M]$ is the *restriction* of a feature allocation f_N of $[N]$ for $M < N$ if

$$f_M = \{A \cap [M] : A \in f_N, A \cap [M] \neq \emptyset\}.$$

Let $\mathcal{R}_N(f_M)$ be the set of all feature allocations of $[N]$ whose restriction to $[M]$ is f_M.

Let $\mathbb{P}$ denote a probability measure on some probability space supporting (F_n). We say that the sequence of random feature allocations (F_n) is *consistent in distribution* if for all M and N such that $M < N$, we have

$$\mathbb{P}(F_M = f_M) = \sum_{f_N \in \mathcal{R}_N(f_M)} \mathbb{P}(F_N = f_N).$$

We say that the sequence (F_n) is *strongly consistent* if for all M and N such that $M < N$, we have

$$F_N \overset{a.s.}{\in} \mathcal{R}_N(F_M).$$

Given any (F_n) that is consistent in distribution, the Kolmogorov extension theorem implies that we can construct a sequence of random feature allocations that is strongly consistent and has the same finite dimensional distributions. So henceforth we simply use the term "consistency" to refer to strong consistency.

With this consistency condition, we can define a random feature allocation F_∞ of $\mathbb{N}$ as a consistent sequence of finite feature allocations. Thus F_∞ may be thought of as a random element of the space of such sequences: $F_\infty = (F_n)_{n=1}^{\infty}$. We say that F_N is a restriction of F_∞ to $[N]$ when it is the Nth element in this sequence. We let $\mathcal{F}_\infty$ denote the space of consistent feature allocation sequences, of which each random feature allocation is a random element. The sigma field associated with this space is generated by the finite-dimensional sigma fields of the restricted random feature allocations F_n.

We say that F_∞ is exchangeable if $F_\infty \overset{d}{=} \sigma(F_\infty)$ for every finite permutation σ. That is, for every permutation σ that changes no indices above N for some $N < \infty$, we require $F_N \overset{d}{=} \sigma(F_N)$, where F_N is the restriction of F_∞ to $[N]$.

5.3 Labeling features

Now that we have defined consistent, exchangeable random feature allocations, we want to characterize the class of all distributions on these allocations. We begin by considering some alternative representations of the feature allocation that are not merely useful, but indeed key to some of our later results.

A number of authors have made use of matrices as a way of representing feature allocations (Griffiths and Ghahramani, 2006; Thibaux and Jordan, 2007; Doshi et al., 2009). This representation, while a boon for intuition in some regards, requires care because a matrix presupposes an order on the features, which is not a part of the feature allocation a priori. We cover this distinction in some detail next.

We start by defining an *a priori labeled feature allocation*. Let $\hat{F}_{N,1}$ be the collection of indices in $[N]$ with feature 1, let $\hat{F}_{N,2}$ be the collection of indices in $[N]$ with feature 2, etc. Here, we think of a priori labels as being the ordered, positive natural numbers. This specification is different from (a priori unlabeled) feature allocations as defined above since there is nothing to distinguish the features in a feature allocation other than, potentially, the members of a feature. Consider the following analogy: an a priori labeled feature allocation is to a feature allocation as a classification is to a clustering. Indeed, when each index n belongs to exactly one feature in an a priori feature allocation, feature 1 is just class 1, feature 2 is class 2, and so on.

Another way to think of an a priori labeled feature allocation of $[N]$ is as a matrix of N rows filled with zeros and ones. Each column is associated with a feature. The (n, k) entry in the matrix is one if index n is in feature k and zero otherwise. However, just as—contrary to the classification case—we do not know the ordering of clusters in a clustering a priori, we do not a priori know the ordering of features in a feature allocation. To make use of a matrix representation for a feature allocation, we will need to introduce or find such an order.

The reasoning above suggests that introducing an order for features in a feature allocation would be useful. The next example illustrates that the probability $\mathbb{P}(F_N = f_N)$ in some sense undercounts features when they contain exactly the same indices: e.g., $A_j = A_k$ for some $j \neq k$. This fact will suggest to us that it is not merely useful, but indeed a key point of our theoretical development, to introduce an ordering on features.

Example 5.3.1 (A Bernoulli, two-feature allocation)**.** Given $q_A, q_B \in (0, 1)$, draw $Z_{n,A} \overset{iid}{\sim} \text{Bern}(q_A)$ and $Z_{n,B} \overset{iid}{\sim} \text{Bern}(q_B)$, independently, and construct the random feature allocation by collecting those indices with successful draws:

$$F_N := \{\{n : n \leq N, Z_{n,A} = 1\}, \{n : n \leq N, Z_{n,B} = 1\}\}.$$

One caveat here is that if either of the two sets in the multiset F_N is empty, we do not include it in the allocation. Note that calling the features A and B was merely for the purposes of construction, and in defining F_N, we have lost all feature labels. So F_N is a feature allocation, not an a priori labeled feature allocation.

Then the probability of the feature allocation $F_5 = f_5 := \{\{2, 3\}, \{2, 3\}\}$ is

$$q_A^2(1 - q_A)^3 q_B^2 (1 - q_B)^3,$$

but the probability of the feature allocation $F_5 = f_5' := \{\{2, 3\}, \{2, 5\}\}$ is

$$2q_A^2(1 - q_A)^3 q_B^2 (1 - q_B)^3.$$

The difference is that in the latter case the features can be distinguished, and so we must account for the two possible pairings of features to frequencies $\{q_A, q_B\}$.

Now, instead, let $\tilde{F}_N$ be F_N with the features ordered uniformly at random amongst all possible feature orderings. There is just a single possible ordering of f_5, so the probability of $\tilde{F}_5 = \tilde{f}_5 := (\{2, 3\}, \{2, 3\})$ is again

$$q_A^2(1 - q_A)^3 q_B^2 (1 - q_B)^3.$$

However, there are two orderings of f_5', each of which is equally likely. The probability of $\tilde{F}_N = \tilde{f}_5' := (\{2, 5\}, \{2, 3\})$ is

$$q_A^2(1 - q_A)^3 q_B^2 (1 - q_B)^3.$$

The same holds for the other ordering. ■

This example suggests that there are combinatorial factors that must be taken into account when working with the distribution of F_N directly. The example also suggests that we can avoid the need to specify such factors by instead working with a suitable randomized ordering of the random feature allocation F_N. We achieve this ordering in two steps.

The first step involves ordering the features via a procedure that we refer to as *order-of-appearance labeling.* The basic idea is that we consider data indices $n = 1, 2, 3$, and so on in order. Each time a new data point arrives, we examine the features associated with that data point. Each time we see a new feature, we label it with the lowest available feature label from $k = 1, 2, \ldots$.

In practice, the order-of-appearance scheme requires some auxiliary randomness since each index n may belong to zero, one, or many different features (though the number must be finite). When multiple features first appear for index n, we order them uniformly at random. That simple idea is explained in full detail as follows. Recursively suppose that there are K features among the indices $[N-1]$. Trivially there are zero features when no indices have been seen yet. Moreover, we suppose that we have features with labels 1 through K if $K \geq 1$, and if $K = 0$, we have no features. If features remain without labels, there exists some minimum index n in the data indices such that $n \notin \bigcup_{k=1}^{K} A_k$, where the union is $\emptyset$ if $K = 0$. It is possible that no features contain n. So we further note that there exists some minimum index m such that $m \notin \bigcup_{j=1}^{K} A_j$ but m is contained in some feature of the allocation. By construction, we must have $m \geq N$. Let K_m be the number of features containing m; K_m is finite by definition of a feature allocation. Let (U_k) denote a sequence of iid uniform random variables, independent of the random feature allocation. Assign $U_{K+1}, \ldots, U_{K+K_m}$ to these new features and determine their order of appearance by the order of these random variables. While features remain to be labeled, continue the recursion with N now equal to m and K now equal to $K + K_m$.

Example 5.3.2 (Feature labeling schemes)**.** Consider the feature allocation

$$f_6 = \{\{2,5,4\},\{3,4\},\{6,4\},\{3\},\{3\}\}. \tag{5.1}$$

And consider the random variables

$$U_1, U_2, U_3, U_4, U_5 \overset{iid}{\sim} \text{Unif}[0,1].$$

We see from f_6 that index 1 has no features. Index 2 has exactly one feature, so we assign this feature, $\{2,5,4\}$, to have order-of-appearance label 1. While U_1 is associated with this feature, we do not need to break any ties at this point, so it has no effect.

Index 3 is associated with three features. We associate each feature with exactly one of U_2, U_3, and U_4 (the next three available U_k). For instance, pair $\{3,4\}$ with U_2, $\{3\}$ with U_3, and the other $\{3\}$ with U_4. Suppose it happens that $U_3 < U_2 < U_4$. Then the feature $\{3\}$ paired with U_3 receives label 2 (the next available order-of-appearance label). The feature $\{3,4\}$ receives label 3. And the feature $\{3\}$ paired with U_4 receives label 4.

Index 4 has three features, but $\{2,5,4\}$ and $\{3,4\}$ are already labeled. So the only remaining feature, $\{6,4\}$, receives the next available order-of-appearance label: 5. U_5 is associated with this feature, but since we do not need to break ties here, it has no effect. Indices 5 and 6 belong to already-labeled features.

So the features can be listed with order-of-appearance indices as

$$A_1 = \{2,5,4\}, A_2 = \{3\}, A_3 = \{3,4\}, A_4 = \{3\}, A_5 = \{6,4\}. \tag{5.2}$$

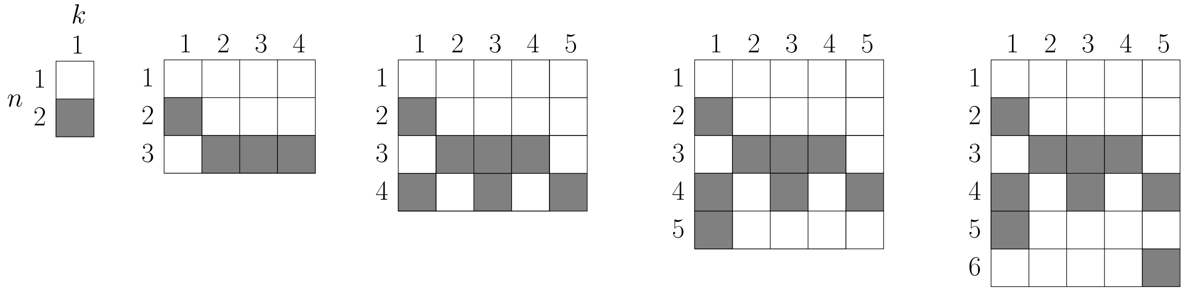


Figure 5.2: Order-of-appearance binary matrix representations of the sequence of feature allocations on $[2], [3], [4], [5]$, and $[6]$ found by restricting f_6 in Example 5.3.2. Rows correspond to indices n, and columns correspond to order-of-appearance feature labels k. A gray square indicates a 1 entry, and a white square indicates a 0 entry. Y_n°, the set of order-of-appearance feature assignments of index n, is easily read off from the matrix as the set of columns with entry in row n equal to 1.

Let Y_n° indicate the set of order-of-appearance feature labels for the features to which index n belongs; i.e., if the features are labeled according to order of appearance as in Eq. (5.2), then $Y_n^\circ = \{k : n \in A_k\}$. By definition of a feature allocation, Y_n° must have finite cardinality. The order-of-appearance labeling gives $Y_1^\circ = \emptyset, Y_2^\circ = \{1\}, Y_3^\circ = \{2,3,4\}, Y_4^\circ = \{1,3,5\}, Y_5^\circ = \{1\}, Y_6^\circ = \{5\}$.

Order-of-appearance labeling is well-suited for matrix representations of feature allocations. The rows of the matrix correspond to indices n and the columns correspond to features with order-of-appearance labels k. The matrix representation of the order-of-appearance labeling and resulting feature assignments (Y_n°) for $n \in [6]$ is depicted in Figure 5.2. ■

Note that when the feature allocation is a partition, there is exactly one feature containing any m, so this scheme reduces to the order-of-appearance scheme for cluster labeling.

Consider an exchangeable feature allocation F_∞. Give order-of-appearance labels to the features of this allocation, and let Y_n° be the set of feature labels for features containing n. So Y_n° is a random finite subset of $\mathbb{N}$. It can be thought of as a simple point process on $\mathbb{N}$; a discussion of measurability of such processes may be found in Kallenberg (2002, p. 178). Our process is even simpler than a simple point process as it is globally finite rather than merely locally finite.

Note that $(Y_n^\circ)_{n=1}^\infty$ is not necessarily exchangeable. For instance, consider again Example 5.3.1. If Y_1° is non-empty, $1 \in Y_1^\circ$ with probability one. If Y_2° is non-empty, with positive probability it may not contain 1. To restore exchangeability we extend an idea due to Aldous (1985) in the setting of random partitions; in our feature allocation extension, we associate to each feature a draw from a uniform random variable on $[0,1]$. Drawing these random variables independently we maintain consistency across different values of N. We refer to these random variables as *uniform random feature labels.*

Note that the use of a uniform distribution is for convenience; we simply require that

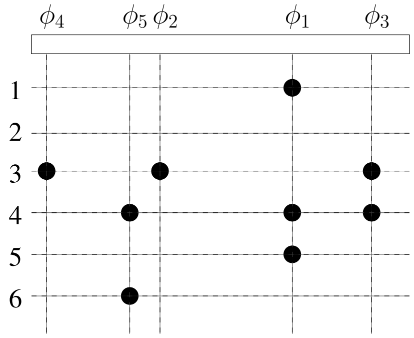


Figure 5.3: An illustration of the uniform random feature labeling in Example 5.3.3. The top rectangle is the unit interval. The uniform random labels are depicted along the interval with vertical dotted lines at their locations. The indices [6] are shown to the left. A black circle shows appears when an index occurs in the feature with a given label. The matrix representations of this feature allocation in Figure 5.4 can be recovered from this plot.

features receive distinct labels with probability one, so any other continuous distribution would suffice. We also note that in a full-fledged model based on random feature allocations these labels often play the role of parameters and are used in defining the likelihood. For further discussion of such constructions, see Broderick, Jordan, and Pitman (2013).

Thus, let (ϕ_k) be a sequence of iid uniform random variables, independent of both (U_k) and F_∞. Construct a new feature labeling by taking the feature labeled k in the order-of-appearance labeling and now label it ϕ_k. In this case, let $Y_n^\dagger$ denote the set of feature labels for features to which n belongs. Call this a *uniform random labeling.* $Y_n^\dagger$ can be thought of as a (globally finite) simple point process on $[0, 1]$. Again, we refer the reader to Kallenberg (2002, p. 178) for a discussion of measurability.

Example 5.3.3 (Feature labeling schemes (continued))**.** Again consider the feature allocation

$$f_6 = \{\{2, 5, 4\}, \{3, 4\}, \{6, 4\}, \{3\}, \{3\}\}.$$

Now consider the random variables

$$U_1, U_2, U_3, U_4, U_5, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5 \overset{iid}{\sim} \text{Unif}[0, 1].$$

Recall from Example 5.3.2 that $U_1, \ldots, U_5$ gave us the order-of-appearance labeling of the features. This labeling allowed us to index the features as in Eq. (5.2), copied here:

$$A_1 = \{2, 5, 4\}, A_2 = \{3\}, A_3 = \{3, 4\}, A_4 = \{3\}, A_5 = \{6, 4\}.$$

With this order-of-appearance labeling in hand, we can assign a uniform random label to each feature. In particular, we assign the uniform random label ϕ_k to the feature with order-of-appearance label k: $A_1 = \{2, 5, 4\}$ gets label ϕ_1, $A_2 = \{3\}$ gets label ϕ_2, $A_3 = \{3, 4\}$

gets label ϕ_3, $A_4 = \{3\}$ gets label ϕ_4, and $A_5 = \{6, 4\}$ gets label ϕ_5. Let $Y_n^\dagger$ indicate the set of uniform random feature labels for the features to which index n belongs. The uniform random labeling gives

$$Y_1^\dagger = \emptyset, Y_2^\dagger = \{\phi_1\}, Y_3^\dagger = \{\phi_2, \phi_3, \phi_4\}, Y_4^\dagger = \{\phi_1, \phi_3, \phi_5\}, Y_5^\dagger = \{\phi_1\}, Y_6^\dagger = \{\phi_5\}. \tag{5.3}$$

■

Lemma 5.3.4. *Give the features of an exchangeable feature allocation F_∞ uniform random labels, and let $Y_n^\dagger$ be the set of feature labels for features containing n. So $Y_n^\dagger$ is a random finite subset of $[0, 1]$. Then the sequence $(Y_n^\dagger)_{n=1}^\infty$ is exchangeable.*

Proof. Note that $(Y_n^\dagger)_{n=1}^\infty = g((\phi_k)_k, (U_k)_k, F_\infty)$ for some measurable function g. Consider any finite permutation σ that does not change any index n with $n > N$ for some fixed, finite N. Let K represent the (potentially random but finite) number of features in F_N. If we construct order-of-appearance labels using the same $(U_k)_k$ as above and now $\sigma(F_\infty)$ instead of F_∞, the labels will not differ from the original order-of-appearance labels after the first K features. Therefore, there exists some finite permutation τ—which may be a function of $(U_k)_{k=1}^K$, σ, and F_N and hence random—such that $(Y_{\sigma(n)}^\dagger)_n = g((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty))$.

Now

$$((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} ((\phi_k)_k, (U_k)_k, \sigma(F_\infty))$$

since the iid sequence $(\phi_k)_k$, the iid sequence $(U_k)_k$, and F_∞ are independent by construction and

$$((\phi_k)_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} ((\phi_k)_k, (U_k)_k, F_\infty)$$

since the feature allocation is exchangeable and the independence used above still holds. So

$$g((\phi_{\tau(k)})_k, (U_k)_k, \sigma(F_\infty)) \stackrel{d}{=} g((\phi_k)_k, (U_k)_k, F_\infty).$$

It follows that the sequence $(Y_n^\dagger)_n$ is exchangeable. □

We can recover the full feature allocation F_∞ from the sequence $Y_1^\dagger, Y_2^\dagger, \ldots$. In particular, if $\{x_1, x_2, \ldots\}$ are the unique values in $\{Y_1^\dagger, Y_2^\dagger, \ldots\}$, then the features are $\{\{n : x_k \in Y_n^\dagger\} : k = 1, 2, \ldots\}$. The feature allocation can similarly be recovered from the order-of-appearance label collections (Y_n°).

We can also recover a new *random ordered feature allocation* $\tilde{F}_N$ from the sequence $(Y_n^\dagger)$. In particular, $\tilde{F}_N$ is the sequence—rather than the collection—of features $\{n : x_k \in Y_n^\dagger\}$ such that the feature with smallest label ϕ_k occurs first, and so on. This construction achieves our goal of avoiding the combinatorial factors needed to work with the distribution of F_N, while retaining exchangeability and consistency.

Example 5.3.5 (Feature labeling schemes (continued))**.** Once more, consider the feature allocation

$$f_6 = \{\{2, 5, 4\}, \{3, 4\}, \{6, 4\}, \{3\}, \{3\}\}.$$

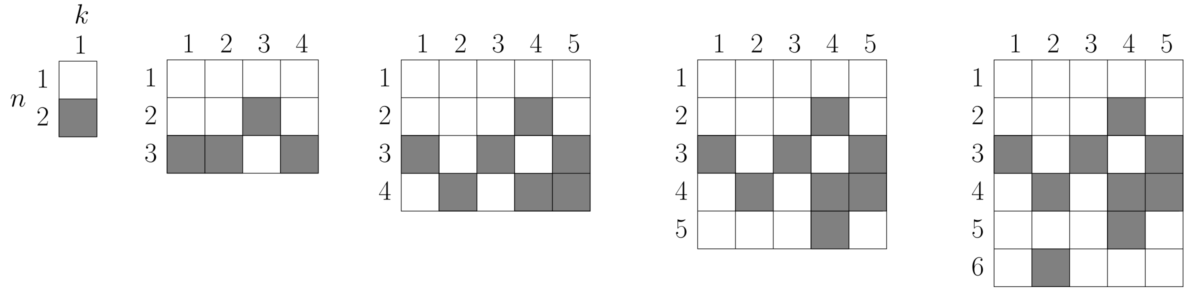


Figure 5.4: The same consistent sequence of feature allocations in Figure 5.2 but now with the uniform random order of Example 5.3.5 instead of the order of appearance illustrated in Figure 5.2.

and the uniform random labeling in Eq. (5.3). If it happens that $\phi_4 < \phi_5 < \phi_2 < \phi_1 < \phi_3$, then the random ordered feature allocation is

$$\tilde{f}_6 = (\{3\}, \{6,4\}, \{3\}, \{2,5,4\}, \{3,4\}).$$

■

Recall that we were motivated by Example 5.3.1 to produce such a random ordering scheme to avoid obfuscating combinatorial factors in the probability of a feature allocation. From another perspective, these factors arise because the random labeling is in some sense more natural than alternative labelings; again, consider random labels as iid parameters for each feature. While order-of-appearance labeling is common due to its pleasant aesthetic representation in matrix form (compare Figures 5.2 and 5.4), one must be careful to remember that the order-of-appearance label sets (Y_n°) are not exchangeable. We will use random labeling extensively below since, among other nice properties, it preserves exchangeability of the sets of feature labels associated with the indices.

5.4 Exchangeable feature probability function

In general, given a probability of a random feature allocation, $\mathbb{P}(F_N = f_N)$, we can find the probability of a random ordered feature allocation $\mathbb{P}(\tilde{F}_N = \tilde{f}_N)$ as follows. Let H be the number of distinct features of F_N, and let $(\tilde{K}_1, \ldots, \tilde{K}_H)$ be the multiplicities of these distinct features in decreasing order. Then

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = \binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H}^{-1} \mathbb{P}(F_N = f_N), \tag{5.4}$$

where

$$\binom{K}{\tilde{K}_1, \ldots, \tilde{K}_H} := \frac{K!}{\tilde{K}_1! \cdots \tilde{K}_H!}.$$

For partitions, the effect of this multiplicative factor is the same across all partitions with the same number of clusters; for some number of clusters K, it is just $1/K!$. In the general feature case, the multiplicative factor may be different for different feature configurations with the same number of features.

Example 5.4.1 (A Bernoulli, two-feature allocation (continued))**.** Consider F_N constructed as in Example 5.3.1. Denote the sizes of the two features by $M_{N,1}$ and $M_{N,2}$. Then

$$\begin{aligned}\mathbb{P}(\tilde{F}_N = \tilde{f}_N) &= \frac{1}{2} q_A^{M_{N,1}} (1-q_A)^{N-M_{N,1}} q_B^{M_{N,2}} (1-q_B)^{N-M_{N,2}} \\ &\quad + \frac{1}{2} q_A^{M_{N,2}} (1-q_A)^{N-M_{N,2}} q_B^{M_{N,1}} (1-q_B)^{N-M_{N,1}} \\ &= p(N, M_{N,1}, M_{N,2}). \end{aligned} \tag{5.5}$$

Here, p is some function of the number of indices N and the feature sizes $(M_{N,1}, M_{N,2})$ that we note is symmetric in $(M_{N,1}, M_{N,2})$; i.e., $p(N, M_{N,1}, M_{N,2}) = p(N, M_{N,2}, M_{N,1})$. ■

When the feature allocation probability admits the representation

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N) = p(N, |A_1|, \ldots, |A_K|) \tag{5.6}$$

for every ordered feature allocation $\tilde{f}_N = (A_1, \ldots, A_K)$ and some function p that is symmetric in all arguments after the first, we call p the *exchangeable feature probability function* (EFPF). We take care to note that the exchangeable partition probability function (EPPF), which always exists for partitions, is not a special case of the EFPF. Indeed, the EPPF assigns zero probability to any multiset in which an index occurs in more than one feature of the multiset; e.g., $\{\{1\}, \{2\}\}$ is a valid partition and a valid feature allocation of $[2]$, but $\{\{1\}, \{1\}\}$ is a valid feature allocation but not a valid partition of $[2]$. Thus, the EPPF must examine the feature indices of a feature allocation to judge their exclusivity and thereby assign a probability. By contrast, the indices in the multiset provide no such information to the EFPF; only the sizes of the multiset features are relevant in the EFPF case.

Proposition 5.4.2. *The class of exchangeable feature allocations with EFPFs is a strict but non-empty subclass of the class of exchangeable feature allocations.*

Proof. Example 5.4.3 below shows that the class of feature allocations with EFPFs is non-empty, and Example 5.4.4 below establishes that there exist simple exchangeable feature allocations without EFPFs. □

Example 5.4.3 (Three-parameter Indian buffet process)**.** The Indian buffet process (IBP) (Griffiths and Ghahramani, 2006) is a generative model for a random feature allocation that is specified recursively in a manner akin to the Chinese restaurant process (Aldous, 1985) in the case of partitions. The metaphor involves a set of "customers" that enter a restaurant and sample a set of "dishes." Order the customers by placing them in one-to-one

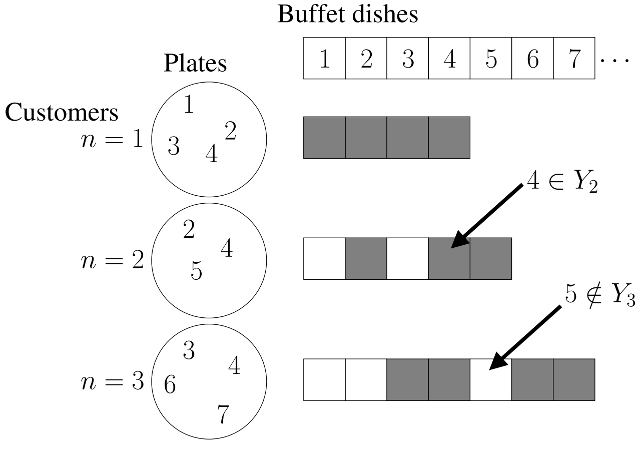


Figure 5.5: Illustration of an Indian buffet process in the order-of-appearance representation of Figure 5.2. The buffet (*top*) consists of a vector of dishes, corresponding to features. Each customer—corresponding to a data point—who enters the restaurant first decides whether or not to choose dishes that the other customers have already sampled. The customer then selects a random number of new dishes, not previously sampled by any customer. A gray box in position (n, k) indicates customer n has sampled dish k, and a white box indicates the customer has not sampled the dish. In the example, the second customer has sampled exactly those dishes indexed by 2, 4, and 5: $Y_2^\circ = \{2, 4, 5\}$.

correspondence with the indices $n \in \mathbb{N}$. The dishes in the restaurant correspond to feature labels. Customers in the Indian buffet can sample any non-negative integer number of dishes. The set of dishes chosen by a customer n is just Y_n°, the collection of feature labels for the features to which n belongs, and the procedure described below provides a way to construct Y_n° recursively.

We describe an extended version (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012) of the Indian buffet that includes two extra parameters beyond the single *mass parameter* γ ($\gamma > 0$) originally specified by Griffiths and Ghahramani (2006); in particular, we include a *concentration parameter* θ ($\theta > 0$) and a *discount parameter* α ($\alpha \in [0, 1)$). We abbreviate this three-parameter IBP as "3IBP." The single-parameter IBP may be recovered by setting $\theta = 1$ and $\alpha = 0$.

We start with a single customer, who enters the buffet and chooses $K_1^+ \sim \text{Poisson}(\gamma)$ dishes. None of the dishes have been sampled by any other customers since no other customers have yet entered the restaurant. An order-of-appearance labeling gives the dishes labels $1, \ldots, K_1^+$ if $K_1^+ > 0$.

Recursively, the nth customer chooses which dishes to sample in two phases. First, for each dish k that has previously been sampled by any customer in $1, \ldots, n-1$, customer n samples dish k with probability

$$\frac{M_{n-1,k} - \alpha}{\theta + n - 1},$$

for $M_{n,k}$ equal to the number of customers indexed $1, \ldots, n$ who have tried dish k. As each dish represents a feature, sampling a dish represents that the customer index n belongs to that feature. And $M_{n,k}$ is the size of the feature labeled k in the feature allocation of $[n]$.

Next, customer n chooses

$$K_n^+ \sim \text{Poisson}\left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right)$$

new dishes to try. If $K_n^+ > 0$, then the dishes receive unique order-of-appearance labels $K_{n-1}+1, \ldots, K_n$. Here, K_n represents the number of sampled dishes after n customers: $K_n = K_{n-1} + K_n^+$ (with base case $K_0 = 0$).

With this generative model in hand, we can find the probability of a particular feature allocation. We discover its form by enumeration. At each round n, we have a Poisson number of new features, K_n^+, represented. The probability factor associated with these choices is a product of Poisson densities:

$$\prod_{n=1}^{N} \frac{1}{K_n^+!} [C(n, \gamma, \theta, \alpha)]^{K_n^+} \exp\left(-C(n, \gamma, \theta, \alpha)\right),$$

where

$$C(n, \gamma, \theta, \alpha) := \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}.$$

Let R_k be the round on which the kth dish, in order of appearance, is first chosen. Then the denominators for future dish choice probabilities are the factors in the product $(\theta+R_k) \cdot (\theta+R_k+1) \cdots (\theta+N-1)$. The numerators for the times when the dish is chosen are the factors in the product $(1-\alpha) \cdot (2-\alpha) \cdots (M_{N,k}-1-\alpha)$. The numerators for the times when the dish is not chosen yield $(\theta+R_k-1+\alpha) \cdots (\theta+N-1-M_{N,k}+\alpha)$. Let $A_{n,k}$ represent the collection of indices in the feature with label k after n customers have entered the restaurant. Then $M_{n,k} = |A_{n,k}|$.

Finally, let $\tilde{K}_1, \ldots, \tilde{K}_H$ be the multiplicities of distinct features formed by this model. We note that there are

$$\left[\prod_{n=1}^{N} K_n^+!\right] / \left[\prod_{h=1}^{H} \tilde{K}_h!\right]$$

rearrangements of the features generated by this process that all yield the same feature allocation. Since they all have the same generating probability, we simply multiply by this factor to find the feature allocation probability.

Multiplying all factors together[1] and taking $f_n = \{A_{N,1}, \ldots, A_{N,K_N}\}$ yields

$$\mathbb{P}(F_N = f_N)$$

[1] Readers curious about how the R_k terms disappear may observe that

$$\prod_{k=1}^{K_N} \frac{\Gamma(\theta+R_k)}{\Gamma(\theta+R_k+\alpha-1)} = \prod_{n=1}^{N} \left(\frac{\Gamma(\theta+n)}{\Gamma(\theta+n+\alpha-1)}\right)^{K_N^+}.$$

$$= \left(\prod_{h=1}^{H} \tilde{K}_h! \right)^{-1} \left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \right)^{K_N} \exp\left(- \sum_{n=1}^{N} \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)} \right)$$
$$\cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(M_{N,k}-\alpha)}{\Gamma(1-\alpha)} \cdot \frac{\Gamma(\theta+N-M_{N,k}+\alpha)}{\Gamma(\theta+N)} \right].$$

It follows from Eq. (5.4) that the probability of a uniform random ordering of the feature allocation is

$$\mathbb{P}(\tilde{F}_N = \tilde{f}_N)$$
$$= \frac{1}{K_N!} \left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \right)^{K_N} \exp\left(- \sum_{n=1}^{N} \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)} \right)$$
$$\cdot \left[\prod_{k=1}^{K_N} \frac{\Gamma(M_{N,k}-\alpha)}{\Gamma(1-\alpha)} \cdot \frac{\Gamma(\theta+N-M_{N,k}+\alpha)}{\Gamma(\theta+N)} \right]. \tag{5.7}$$

The distribution of $\tilde{F}_N$ has no dependence on the ordering of the indices in $[N]$. Hence, the distribution of F_N depends only on the same quantities—the number of indices and the feature sizes—and the feature multiplicities. So we see that the 3IBP construction yields an exchangeable random feature allocation. Consistency follows from the recursive construction and exchangeability. Therefore, Eq. (5.7) is seen to be in EFPF form given by Eq. (5.6). ■

The three-parameter Indian buffet process has an EFPF representation, but the following simple model does not.

Example 5.4.4 (A general two-feature allocation). We here describe an exchangeable, consistent random feature allocation whose (randomly ordered) distribution does not depend only on the number of indices N and the sizes of the features of the allocation.

Let $p_{10}, p_{01}, p_{11}, p_{00}$ be fixed frequencies that sum to one. Let Y_n represent the collection of features to which index n belongs. For $n \in \{1, 2\}$, choose Y_n independently and identically according to:

$$Y_n = \begin{cases} \{1\} & \text{with probability } p_{10} \\ \{2\} & \text{with probability } p_{01} \\ \{1,2\} & \text{with probability } p_{11} \\ \emptyset & \text{with probability } p_{00}. \end{cases}$$

We form a feature allocation from these labels as follows. For each label (1 or 2), collect those indices n with the given label appearing in Y_n to form a feature.

Now consider two possible outcome feature allocations: $f_2 = \{\{2\},\{2\}\}$, and $f_2' = \{\{1\},\{2\}\}$. The probability of any ordering $\tilde{f}_2$ of f_2 under this model is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) = p_{10}^0 \, p_{01}^0 \, p_{11}^1 \, p_{00}^1.$$

To see this result, note the distinction between indices $\{1, 2\}$ and the feature labels $\{1, 2\}$ used in an intermediate step above. Likewise, the probability of any ordering $\tilde{f}'_2$ of f'_2 is

$$\mathbb{P}(\tilde{F}_2 = \tilde{f}'_2) = p_{10}^1 \, p_{01}^1 \, p_{11}^0 \, p_{00}^0.$$

It follows from these two probabilities that we can choose values of $p_{10}, p_{01}, p_{11}, p_{00}$ such that $\mathbb{P}(\tilde{F}_2 = \tilde{f}_2) \neq \mathbb{P}(\tilde{F}_2 = \tilde{f}'_2)$. But $\tilde{f}_2$ and $\tilde{f}'_2$ have the same feature counts and N value ($N = 2$). So there can be no such symmetric function p, as in Eq. (5.5), for this model. ■

5.5 The Kingman paintbox and feature paintbox

Since the class of exchangeable feature models with EFPFs is a strict subclass of the class of exchangeable feature models, it remains to find a characterization of the latter class. Noting that the sequence of feature collections $Y_n^\dagger$ is an exchangeable sequence when the uniform random labeling of features is used, we might turn to the de Finetti mixing measure of this exchangeable sequence for such a characterization.

Indeed, in the partition case, the Kingman paintbox (Kingman, 1978; Aldous, 1985) provides just such a characterization.

Theorem 5.5.1 (Kingman paintbox). *Let $\Pi_\infty := (\Pi_n)_{n=1}^\infty$ be an exchangeable random partition of $\mathbb{N}$, and let $(M_{n,k}^\downarrow, k \geq 1)$ be the decreasing rearrangement of cluster sizes of Π_n with $M_{n,k}^\downarrow = 0$ if Π_n has fewer than k clusters. Then $M_{n,k}^\downarrow/n$ has an almost sure limit $\rho_k^\downarrow$ as $n \to \infty$ for each k. Moreover, the conditional distribution of Π_∞ given $(\rho_k^\downarrow, k \geq 1)$ is as if Π_∞ were generated by random sampling from a random distribution with ranked atoms $(\rho_k^\downarrow, k \geq 1)$.*

When the partition clusters are labeled with uniform random labels rather than by the ranking in the statement of the theorem above, Kingman's paintbox provides the de Finetti mixing measure for the sequence of partition labels of each index n. Two representations of an example Kingman paintbox are illustrated in Figure 5.6. The Kingman paintbox is so named since we imagine each subinterval of the unit interval as containing paint of a certain color; the colors have a one-to-one mapping with the uniform random cluster labels. A random draw from the unit interval is painted with the color of the Kingman paintbox subinterval into which it falls. While Figure 5.6 depicts just four subintervals and hence at most four clusters, the Kingman paintbox may in general have a countable number of subintervals and hence clusters. Moreover, these subintervals may themselves be random.

Note that the ranked atoms need not sum to one; in general, $\sum_k \rho_k^\downarrow \leq 1$. When random sampling from the Kingman paintbox does not select some atom k with $\rho_k^\downarrow > 0$, a new cluster is formed but it is necessarily never selected again for another index. In particular, then, a corollary of the Kingman paintbox theorem is that there are two types of clusters: those with unbounded size as the number of indices N grows to infinity and those with exactly one member as N grows to infinity; the latter are sometimes referred to as *singletons* or

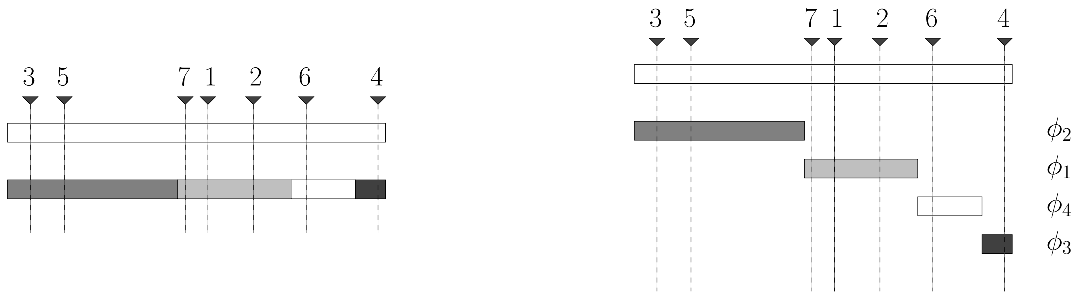


Figure 5.6: *Left*: An example Kingman paintbox. The upper rectangle represents the unit interval. The lower rectangles represent a partition of the unit interval into four subintervals corresponding to four clusters. The horizontal locations of the seven vertical lines represent seven uniform random draws from the unit interval. The resulting partition of $[7]$ is $\{\{3,5\},\{7,1,2\},\{6\},\{4\}\}$. *Right*: An alternate representation of the same Kingman paintbox, now with each subinterval separated out into its own vertical level. To the right of each cluster subinterval is a uniform random label (with index determined by order of appearance) for the cluster.

collectively as *Kingman dust*. In the feature case, we impose one further regularity condition that essentially rules out dust. Consider any feature allocation F_∞. Recall that we use the notation $Y_n^\dagger$ to indicate the set of features to which index n belongs. We assume that, for each n, with probability one there exists some m with $m \neq n$ such that $Y_m^\dagger = Y_n^\dagger$. Equivalently, with probability one there is no index with a unique feature collection. We call a random feature allocation that obeys this condition a *regular feature allocation*.

We can prove the following theorem for the feature case, analogous to the Kingman paintbox construction for partitions.

Theorem 5.5.2 (Feature paintbox)**.** *Let $F_\infty := (F_n)$ be an exchangeable, consistent, regular random feature allocation of $\mathbb{N}$. There exists a random sequence $(C_k)_{k=1}^\infty$ such that C_k is a countable union of subintervals of $[0,1]$ (and may be empty) and such that F_∞ has the same distribution as F'_∞ where F'_∞ is generated as follows. Randomly sample $(U'_n)_n$ iid uniform in $[0,1]$. Let $Y_n := \{k : U'_n \in C_k\}$ represent a collection of feature labels for index n, and let F'_∞ be the induced feature allocation from these label collections.*

Proof. Given F_∞ as in the theorem statement, we can construct $(Y_n^\dagger)_{n=1}^\infty$ as in Lemma 5.3.4. Then, according to Lemma 5.3.4, $(Y_n^\dagger)_{n=1}^\infty$ is an exchangeable sequence. Note that $Y_n^\dagger$ defines a partition: $n \sim m$ (i.e., n and m belong to the same cluster of the partition) if and only if $Y_n^\dagger = Y_m^\dagger$. This partition is exchangeable since the feature allocation is. Moreover, since we assume there are no singletons in the induced partition (by regularity), the Kingman paintbox theorem implies that the Kingman paintbox atoms sum to one.

By de Finetti's theorem (Aldous, 1985), there exists α such that α is the directing random measure for $(Y_n^\dagger)$. Condition on $\alpha = \mu$. Write $\mu = \sum_{j=1}^\infty q_j \delta_{x_j}$, where the q_j satisfy $q_j \in (0,1]$ and are written in monotone decreasing order: $q_1 \geq q_2 \geq \cdots$. The condition that the atoms

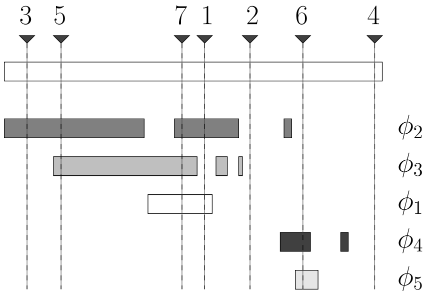


Figure 5.7: An example feature paintbox. The top rectangle represents the unit interval. Each vertical level below the top rectangle represents a subset of the unit interval corresponding to a feature. To the right of each subset is a uniform random label for the feature. For example, using the notation of Theorem 5.5.2, the topmost subset is C_2 corresponding to feature label ϕ_2. The vertical dashed lines represent uniform random draws; i.e., U'_n for index n. The resulting feature allocation of [7] for this realization of the construction is $\{\{3,5,7,1\},\{5,7\},\{7,1\},\{6\},\{6\}\}$. The collection of feature labels for index 7 is $Y_7 = \{\phi_2, \phi_3, \phi_1\}$. The collection of feature labels for index 4 is $Y_4 = \emptyset$.

of the paintbox sum to one translates to $\sum_{j=1}^{\infty} q_j = 1$. The (x_j) are the (countable) unique values of $Y_n^{\dagger}$, ordered to agree with the q_j. The strong law of large numbers yields

$$N^{-1}\#\{n : n \leq N, Y_n^{\dagger} = x_j\} \to q_j, \quad N \to \infty.$$

Since $\sum_{j=1}^{\infty} q_j = 1$, we can partition the unit interval into subintervals of length q_j. The jth such subinterval starts at $s_j := \sum_{l=1}^{j-1} q_l$ and ends at $e_j := s_{j+1}$. For $k = 1, 2, \ldots$, define $C_k := \bigcup_{j:\phi_k \in x_j} [s_j, e_j)$. We call the $(C_k)_{k=1}^{\infty}$ the *feature paintbox*.

Then F_∞ has the same distribution as the following construction. Let $(U'_1, U'_2, \ldots)$ be an iid sequence of uniform random variables. For each n, define $Y_n = \{k : U'_n \in C_k\}$ to be the collection of features, now labeled by positive integers, to which n belongs. Let F'_∞ be the feature allocation induced by the (Y_n). □

A point to note about this feature paintbox construction is that the ordering of the feature paintbox subsets C_k in the proof is given by the order of appearance of features in the original feature allocation F_∞. This ordering stands in contrast to the ordering of atoms by size in the Kingman paintbox. Making use of such a size-ordering would be more difficult in the feature case due to the non-trivial intersections of feature subsets. A particularly important implication is that the conditional distribution of F_∞ given $(C_k)_k$ is not the same as that of F'_∞ given $(C_k)_k$ (cf. Pitman (1995) for similar ordering issues in the partition case).

An example feature paintbox is illustrated in Figure 5.7. Again, we may think of each feature paintbox subset as containing paint of a certain color (where these colors have a one-to-one mapping with the uniform random labels). Draws from the unit interval to determine the feature allocation may now be painted with some subset of these colors rather than just a single color.

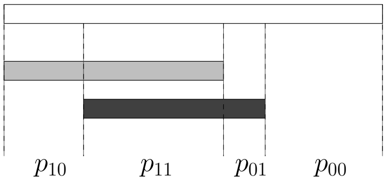


Figure 5.8: A feature paintbox for the two-feature allocation in Example 5.4.4. The top rectangle is the unit interval. The middle rectangle is the feature paintbox subset for feature 1. The lower rectangle is the feature paintbox subset for feature 2.

Next, we revisit earlier examples to find their feature paintbox representations.

Example 5.5.3 (A general two-feature allocation (continued))**.** The feature paintbox for the random feature allocation in Example 5.4.4 consists of two features. The total measure of the paintbox subset for feature 1 is $p_{10} + p_{11}$. The total measure of the paintbox subset for feature 2 is $p_{01} + p_{11}$. The total measure of the intersection of these two subsets is p_{11}. A depiction of this paintbox appears in Figure 5.8. ■

Example 5.5.4 (Three-parameter Indian buffet process (continued))**.** The 3IBP turns out to be an instance of a general class of exchangeable feature models that we refer to as *feature frequency models*. This class of models not only provides a straightforward way to construct feature paintbox representations in general, but also plays a key role in our general theory, providing a link between feature paintboxes and EFPFs. In the following section, we define feature frequency models, develop the general construction of paintboxes from feature frequency models, and then return to the construction of the feature paintbox for the 3IBP as an example. We subsequently turn to the general theoretical characterization of feature frequency models. ■

5.6 Feature frequency models

We now discuss a general class of exchangeable feature models for which it is straightforward to describe the feature paintbox. Let (V_k) be a sequence of (not necessarily independent) random variables with values in $[0, 1]$ such that $\sum_{k=1}^{\infty} V_k < \infty$ almost surely. Let $\phi_k \overset{iid}{\sim}$ Unif$[0, 1]$ and independent of the (V_k). A *feature frequency model* is built around a random measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$. We may draw a feature allocation given B as follows. For each data point n, independently draw its features like so: for each feature indexed by k, independently make a Bernoulli draw with success probability V_k. If the draw is a success, n belongs to the feature indexed by k (i.e., the feature with label ϕ_k). If the draw is a failure, n does not belong to the feature indexed by k. The feature allocation is induced in the usual way from these labels.

The condition that the frequencies have an almost surely finite sum guarantees, by the Borel-Cantelli lemma, that the number of features exhibited by any index n is almost surely

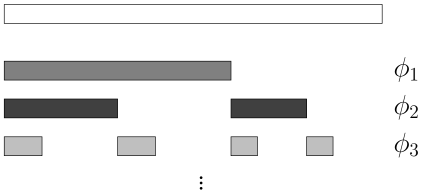


Figure 5.9: An example feature paintbox for a feature frequency model (Section 5.6). One such model is the 3IBP (Example 5.6.1).

finite, as required in the definition of a feature allocation. We obtain exchangeable feature allocations simply by virtue of the fact that the feature allocations are independently and identically distributed given B. The Bernoulli draws from the feature frequencies guarantee that the feature allocation is regular.

Before constructing the feature paintbox for such a model, we note that V_k is the total length of the paintbox subset for the feature indexed by k. In this sense, it is the frequency of this feature (hence the name "feature frequency model"). And ϕ_k is the uniform random feature label for the feature with frequency V_k. Finally, to achieve the independent Bernoulli draws across k required by the feature allocation specification, we need for the intersection of any two paintbox subsets to have length equal to the product of the two paintbox subset lengths. This desideratum can be achieved with a recursive construction.

First, divide the unit interval into one subset (call it I_1) of length V_1 and another subset (call it I_0) of length $1 - V_1$. Then I_1 is the paintbox subset for the feature indexed by 1. Recursively, suppose we have paintbox subsets for features indexed 1 to $K - 1$. Let e be a binary string of length $K - 1$. Suppose that I_e is the intersection of (a) all paintbox subsets for features indexed by k ($k < K$) where the kth digit of e is 1 and (b) all paintbox subset complements for features indexed by k ($k < K$) where the kth digit of e is 0. For every e, we construct $I_{(e,1)}$ to be a subset of I_e with total length equal to V_K times the length of I_e. We construct $I_{(e,0)}$ to be $I_e \backslash I_{(e,1)}$.

Finally, the paintbox subset for the feature indexed by K is the union of all $I_{e'}$ with e' a binary string of length K such that the final digit of e' is 1. An example of such a paintbox is illustrated in Figure 5.9.

Example 5.6.1 (Three-parameter Indian buffet process (continued))**.** We show that the three-parameter Indian buffet process is an example of a feature frequency model, and thus its feature paintbox can be constructed according to the general recipe that we have just presented.

The underlying random measure for the three-parameter Indian buffet process is known as the *three-parameter beta process* (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012). This random measure, denoted B, can be constructed explicitly via the following recursion (with $K_0 = 0$ and $n = 1, 2, \ldots$), which extends the results of Thibaux and Jordan

(2007):

$$
\begin{aligned}
K_n^+ &\sim \text{Poisson}\left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+n)} \cdot \frac{\Gamma(\theta+\alpha-1+n)}{\Gamma(\theta+\alpha)}\right), \\
K_n &= K_{n-1} + K_n^+ \\
V_k &\sim \text{Beta}(1-\alpha, \theta+n-1+\alpha), \quad k = K_{n-1}+1, \ldots, K_n \\
\phi_k &\sim \text{Unif}[0,1] \\
B &= \sum_{k=1}^{\infty} V_k \delta_{\phi_k},
\end{aligned}
$$

where we recall that the ϕ_k are assumed to be drawn from the uniform distribution for simplicity in this chapter, but in general they may be drawn from a continuous distribution that serves as a prior for the parameters defining a likelihood.

Given $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$, the feature allocation is drawn according to the procedure outlined for feature frequency models conditioned on the underlying random measure. Teh and Görür (2009) demonstrate that the distribution of the resulting feature allocation is the same as if it were generated according to a three-parameter Indian buffet process. An alternative proof proceeds as in the two-parameter case covered by Broderick, Jordan, and Pitman (2013). ■

We have seen that the 3IBP can be represented as a feature frequency model. It is straightforward to observe that the two-feature model in Examples 5.4.4 and 5.5.3 cannot be represented as a feature frequency model unless the intersection of the feature subsets has length p_{11} equal to the product of the feature subset lengths ($p_{10} + p_{11}$ and $p_{01} + p_{11}$); i.e., unless $(p_{10} + p_{11})(p_{01} + p_{11}) = p_{11}$ (cf. Figure 5.8). Therefore, we have the following result similar to Proposition 5.4.2.

Proposition 5.6.2. *The class of feature frequency models is a strict but non-empty subclass of the class of exchangeable feature allocations.*

In proving Propositions 5.6.2 and 5.4.2, we used the 3IBP as an example that belongs to both the class of feature models with EFPFs and the class of feature frequency models. Moreover, in both cases we used two-feature models as an example of exchangeable feature models that do not belong to these subclasses; in particular, we used two-feature models in which the feature combination probabilities $p_{10}, p_{01}, p_{11}, p_{00}$ are not in the necessary proportions. These observations suggest that feature frequency models and EFPFs may be linked. We flesh out the relationship between the two representations in the next few results.

We start with *a priori labeled* features. Recall from Section 5.3 that an a priori labeled feature allocation is to a feature allocation what a classification is to a clustering; that is, the feature labels are known in advance. The case where we know the feature order in advance is somewhat easier and gives intuition for the type of result we would like in the true feature allocation case. In particular, we prove the results for the case of two a priori labeled features in Theorem 5.6.3 and then the case of an unbounded number of a priori labeled features in Theorem 5.6.4.

From there, we move on to the (a priori) unlabeled case that is the focus of the chapter and prove the equivalence of EFPFs and a slight extension of feature frequency models in Theorem 5.6.5.

Theorem 5.6.3. *Consider a model with two a priori labeled features: feature 1 and feature 2. If the two features are generated from labeled feature frequencies, the probability of an a priori labeled feature allocation of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2 takes the form $\check{p}(N; M_{N,1}, M_{N,2})$, where we make no symmetry assumptions about $\check{p}$ here and also allow any of $M_{N,1}$ and $M_{N,2}$ to be zero. Conversely, if the probability of any a priori labeled feature allocation can be written as $\check{p}(N; M_{N,1}, M_{N,2})$, then the feature allocation has the same distribution as if it were generated from labeled feature frequencies.*

Proof. Note that throughout this proof we consider the probability of *a particular* labeled feature allocation of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2, as distinct from the probability of all labeled feature allocations of $[N]$ with $M_{N,1}$ occurrences of feature 1 and $M_{N,2}$ occurrences of feature 2. The latter, which is not addressed here, would be the sum over instances of the former. In particular, recalling the matrix representation from Section 5.3, there are

$$\binom{N}{M_{N,1}}\binom{N}{M_{N,2}}$$

possible $N \times 2$ matrices with $M_{N,1}$ ones in the first column and $M_{N,2}$ ones in the second column.

The reader may feel there is some similarity in this setup to the two-feature allocation of Examples 5.4.4 and 5.5.3. We note that the quantities $p_{10}, p_{01}, p_{11}, p_{00}$—which retain essentially the same meaning as in Figure 5.8—may now be random and that their order is pre-specified and non-random.

First, we calculate the probability of a certain labeled feature configuration under this model. Let $M'_{n,10}$ be the number of indices in $[n]$ with feature 1 but not feature 2. Let $M'_{n,01}$ be the number of indices in $[n]$ with feature 2 but not feature 1. Let $M'_{n,00}$ count the indices with neither feature, and let $M'_{n,11}$ count the indices with both features. Then

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \mathbb{E}(p_{10}^{M'_{N,10}} p_{01}^{M'_{N,01}} p_{11}^{M'_{N,11}} p_{00}^{M'_{N,00}}). \tag{5.8}$$

Denote the total probabilities of features 1 and 2 as, respectively, $q_1 = p_{10} + p_{11}$ and $q_2 = p_{01} + p_{11}$. Suppose that we have a feature frequency model. This assumption implies that

$$p_{10} \overset{a.s.}{=} q_1(1-q_2), \quad p_{01} \overset{a.s.}{=} (1-q_1)q_2, \quad p_{11} \overset{a.s.}{=} q_1 q_2, \quad p_{00} \overset{a.s.}{=} (1-q_1)(1-q_2), \tag{5.9}$$

where any one of the equalities in Eq. (5.9) implies the others. It follows that

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \mathbb{E}[q_1^{M_{N,1}}(1-q_1)^{N-M_{N,1}} q_2^{M_{N,2}}(1-q_2)^{N-M_{N,2}}], \tag{5.10}$$

where $M_{n,1} = M'_{n,10} + M'_{n,11}$ is the total number of indices with feature 1, and likewise $M_{n,2} = M'_{n,01} + M'_{n,11}$ is the total number of indices with feature 2.

So we see that making a feature frequency model assumption yields a feature allocation probability in Eq. (5.10) that depends only on $N, M_{N,1}, M_{N,2}$. Since we retain the known labeling in this example, the probability is not symmetric in $M_{N,1}$ and $M_{N,2}$.

In the other direction, suppose we know that

$$\mathbb{P}(\hat{F}_{N,1} = \hat{f}_{N,1}, \hat{F}_{N,2} = \hat{f}_{N,2}) = \check{p}(N, M_{N,1}, M_{N,2}) \tag{5.11}$$

for some function $\check{p}$. Again, we make no symmetry assumptions about $\check{p}$ here, and any of $M_{N,1}$ and $M_{N,2}$ may be zero. Then frequencies $p_{10}, p_{01}, p_{11}, p_{00}$ must exist by the law of large numbers; we note they may be random.

The assumption in Eq. (5.11) implies that the configurations

$$\begin{aligned}(M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) &= (2, 2, 0, 0)\\ (M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) &= (0, 0, 2, 2)\\ (M'_{4,10}, M'_{4,01}, M'_{4,00}, M'_{4,11}) &= (1, 1, 1, 1)\end{aligned}$$

have the same probability. That is, by Eq. (5.8),

$$\mathbb{E}[p_{10}^2 p_{01}^2] = \mathbb{E}[p_{11}^2 p_{00}^2] = \mathbb{E}[p_{10} p_{01} p_{11} p_{00}].$$

It follows that

$$\mathbb{E}[(p_{10}p_{01} - p_{11}p_{00})^2] = \mathbb{E}[p_{10}^2 p_{01}^2 + p_{11}^2 p_{00}^2 - 2 p_{10} p_{01} p_{11} p_{00}] = 0.$$

So it must be that $p_{10}p_{01} \overset{a.s.}{=} p_{11}p_{00}$. Recall that this condition is familiar from Example 5.4.4.

Adding $p_{10}p_{11}$ to both sides of the almost sure equality and then further adding $p_{11}(p_{01} + p_{11})$ to both sides yields

$$(p_{10} + p_{11})(p_{01} + p_{11}) \overset{a.s.}{=} p_{11}(p_{10} + p_{01} + p_{11} + p_{00}),$$

which reduces to

$$q_1 q_2 \overset{a.s.}{=} p_{11}$$

from the definitions of q_1 and q_2 and from the fact that $p_{10} + p_{01} + p_{11} + p_{00} = 1$.

By Eq. (5.9) and surrounding text, we see that Eq. (5.11) implies our model is a feature frequency model. Thus, the equivalence between models with a priori labeled EFPFs and a priori labeled feature frequency models in the case of two features results from simple algebraic manipulations. □

Extending the argument above becomes more tedious when more than two features are involved. In the case of multiple, or even countably many, labeled features, a more elegant proof exists.

Theorem 5.6.4. *Consider a model with features a priori labeled* $1, 2, 3, \ldots$. *If the features are generated from labeled feature frequencies, the probability of an a priori labeled feature allocation of* $[N]$ *with* K *or fewer features and* $M_{N,k}$ *occurrences of feature* k *for* $k \in \{1, \ldots, K\}$ *takes the form* $\breve{p}(N; M_{N,1}, \ldots, M_{N,K})$, *where we make no symmetry assumptions about* $\breve{p}$ *here and note that any of* $M_{N,1}, \ldots, M_{N,K}$ *may be zero. Call* $\breve{p}$ *a* labeled EFPF. *Conversely, if the probability of any a priori labeled feature allocation can be written as* $\breve{p}(N; M_{N,1}, \ldots, M_{N,K})$, *then the feature allocation has the same distribution as if it were generated from labeled feature frequencies.*

Proof. First, consider the claim that every labeled feature frequency model has a labeled EFPF. This claim is intuitively clear since the independent Bernoulli draws at each atom of the (potentially random) measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$ result in a probability that depends only on the number of occurrences of the corresponding feature and not any interactions between features.

To show this direction formally, we consider a fixed, labeled feature allocation $\hat{f}_N = (A_{N,1}, A_{N,2}, \ldots, A_{N,K})$ with $M_{N,k} := |A_{N,k}|$ and note that

$$\begin{aligned}
&\mathbb{P}(\hat{F}_N = \hat{f}_N) \\
&= \mathbb{E}\left[\mathbb{P}(\hat{F}_N = \hat{f}_N | B)\right] \\
&= \mathbb{E}\left[\left(\prod_{k=1}^{K} V_k^{M_{N,k}} (1 - V_k)^{N - M_{N,k}}\right) \cdot \left(\prod_{k=K+1}^{\infty} (1 - V_k)^N\right)\right].
\end{aligned}$$

It follows that $\mathbb{P}(\hat{F}_N = \hat{f}_N)$ has $\breve{p}$ form.

Now consider the other direction. We start with a labeled feature allocation F_∞. In this case, we know that for every labeled feature allocation of $[N]$,

$$\hat{f}_N = (A_{N,1}, \ldots, A_{N,K}),$$

we have that a function $\breve{p}$ exists in the form

$$\mathbb{P}(\hat{F}_N = \hat{f}_N) = \breve{p}(N, |A_{N,1}|, \ldots, |A_{N,K}|), \tag{5.12}$$

with no additional symmetry assumptions for $\breve{p}$ and where the block sizes $M_{N,k} = |A_{N,k}|$ may be zero.

Let $Z_{n,k}$ be one if n belongs to the kth feature (i.e., $n \in A_{N,k}$) or zero otherwise. Let $b_1, \ldots, b_k$ be values in $\{0, 1\}$. Our goal is to show that conditional on some (as yet unknown) labeled feature frequencies, the probability of feature presence factorizes as independent Bernoulli draws:

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | V_1, \ldots, V_K) = \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1 - b_k}. \tag{5.13}$$

By the assumption on $\check{p}$, the labeled feature sizes $M_{N,1}, \ldots, M_{N,K}$ are sufficient for the distribution of the labeled feature allocation. Let ξ_N be the sigma-field of events invariant under permutations of the first N indices. We note that $M_{N,1}, \ldots, M_{N,K}$ are measurable with respect to ξ_N and start by considering

$$\begin{aligned}&\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \\ &= \prod_{k=1}^{K} \mathbb{P}(Z_{1,k} = b_k | Z_{1,1} = b_1, \ldots, Z_{1,k-1} = b_{k-1}, \xi_N).\end{aligned} \tag{5.14}$$

Then since the feature sizes are sufficient for the feature allocation distribution, we have

$$\begin{aligned}&\mathbb{P}(Z_{1,k} = b_k | Z_{1,1} = b_1, \ldots, Z_{1,k-1} = b_{k-1}, \xi_N) \\ &= \mathbb{P}(Z_{1,k} = b_k | \xi_N) \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{P}(Z_{n,k} = b_k | \xi_N) \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Z_{n,k} = b_k\} | \xi_N \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{Z_{n,k} = b_k\}.\end{aligned}$$

The last line follows since the sum is measurable in ξ_N. By the strong law of large numbers, the final sum converges almost surely as $N \to \infty$ to some potentially random value in $[0, 1]$; call it V_k if $b_k = 1$. By Eq. (5.14), then, we have

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \xrightarrow{\text{a.s.}} \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1-b_k}. \tag{5.15}$$

We next observe that the lefthand side of Eq. (5.15) is a reverse martingale. (ξ_N) is a reversed filtration since $\xi_N \supseteq \xi_{N+1}$ for all N. Moreover, (1) $\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N)$ is measurable with respect to ξ_N; (2) the same quantity is integrable; and (3) by the tower law,

$$\mathbb{E}\left[\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) | \xi_{N+1}\right] = \mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_{N+1}).$$

Since $\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N)$ is a reverse martingale, we have that

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_N) \xrightarrow{\text{a.s.}} \mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_\infty)$$

for $\xi_\infty = \bigcap_{n=1}^{\infty} \xi_n$ by reverse martingale convergence. Together with Eq. (5.15), this convergence implies that

$$\mathbb{P}(Z_{1,1} = b_1, \ldots, Z_{1,K} = b_K | \xi_\infty) = \prod_{k=1}^{K} V_k^{b_k} (1 - V_k)^{1-b_k},$$

and since the V_k are measurable with respect to ξ_∞, the tower law yields Eq. (5.13), as was to be shown. □

While illustrative, the two previous results do not directly deal with feature allocations as defined earlier in this chapter; namely, they do not show any equivalence between EFPFs and feature frequency models in the case where the features are unlabeled (which is exactly the case where EFPFs are defined). We will show in the unlabeled case that every feature frequency model has an EFPF and that every regular feature allocation with an EFPF is an feature frequency model. In fact, we can consider a general—i.e., not necessarily regular—feature allocation and characterize the EFPF representation in this case.

Theorem 5.6.5. *Let λ be a non-negative random variable (which may have some arbitrary joint law with the feature frequencies in a feature frequency model). We can obtain an exchangeable feature allocation by generating a feature allocation from a feature frequency model and then, for each index n, including an independent* Poisson(λ)*-distributed number of features of the form $\{n\}$ in addition to those features previously generated (which may also include index n). A feature allocation of this type has an EFPF. Conversely, every feature allocation with an EFPF has the same distribution as one generated by this construction for some joint distribution of λ and the feature frequencies.*

Proof. Suppose a feature allocation $\tilde{f}$ is generated as described by the construction in Theorem 5.6.5 with (potentially random) measure $B = \sum_{k=1}^{\infty} V_k \delta_{\phi_k}$ giving the frequencies in the feature frequency model component. We wish to show that the feature allocation has an EFPF. We will make use of the fact that an equivalent way to generate the Poisson component of the feature allocation is to draw Poisson$\,(N\lambda)$ singletons and then assign each uniformly at random to an index in $[N]$.

Consider $\tilde{f}_N = (A_1, A_2, \ldots, A_K)$. Let $S = \{k : |A_k| = 1\}$ represent the feature indices of the singletons of the feature allocation. These features may have been generated either from the feature frequency model or from the Poisson component. To find the probability of the feature allocation, we consider each possible association of singletons to one of these components. For any such association, let $\tilde{S}$ represent those singletons assigned to the Poisson component; that is, $\tilde{S} \subseteq S$. Let $\tilde{K} = K - |\tilde{S}|$ represent the number of remaining features, which we denote by

$$(\tilde{A}_1, \ldots, \tilde{A}_{\tilde{K}}).$$

Then the probability of this feature allocation satisfies

$$\begin{aligned}
&\mathbb{P}(\tilde{F}_N = \tilde{f}_N) \\
&= \mathbb{E}\left[\mathbb{P}(\tilde{F}_N = \tilde{f}_N | B, \lambda)\right] \\
&= \mathbb{E}\left[\sum_{\tilde{S}:\tilde{S}\subseteq S} N^{-\tilde{S}} \text{Poisson}\left(\tilde{S} | N\lambda\right) \sum_{\substack{(i_1,\ldots,i_{\tilde{K}}) \\ \text{distinct}}} \right.
\end{aligned}$$

$$\frac{1}{K!}\left(V_{i_1}^{|\tilde{A}_1|}(1-V_{i_1})^{N-|\tilde{A}_1|}\cdots V_{i_{\tilde{K}}}^{|\tilde{A}_{\tilde{K}}|}(1-V_{i_{\tilde{K}}})^{N-|\tilde{A}_{\tilde{K}}|} \prod_{\substack{l\in\mathbb{N}\\ l\notin\{i_1,\ldots,i_{\tilde{K}}\}}} (1-V_l)^N \right)\Bigg] .$$

The final expression depends only on the number of data points N and feature sizes and is symmetric in the feature sizes. So it has EFPF form.

In the other direction, we sidestep the issue of feature ordering by looking at the number of features to which each data index belongs. The advantage of this approach is that this number does not depend on the feature order. The following result is the key to making use of this observation.

Lemma 5.6.6. *Let K_n be a sequence of positive integers. For each n, suppose we have (constants)*

$$1 \geq p_{n,1} \geq p_{n,2} \geq \ldots \geq p_{n,K_n} > 0.$$

And, for completeness, suppose $p_{n,k} = 0$ for $k > K_n$. Let $X_{n,k} \sim \mathrm{Bern}(p_{n,k})$, independently across n and k and with $k = 1 : K_n$. Define $\#_n := \sum_{k=1}^{K_n} X_{n,k}$. Then the following are equivalent.

1. *$\#_n \xrightarrow{d} \#$ for some finite-valued random variable $\#$ on $\{0, 1, 2, \ldots\}$.*
2. *There exist (constants) $\{p_k\}_{k=1}^{\infty}$ and λ such that $p_k \in [0, 1]$ and $\lambda > 0$ and further such that, $\forall k = 1, 2, \ldots,$*

$$p_{n,k} \to p_k, \quad n \to \infty \tag{5.16}$$

and

$$\sum_{k=1}^{K_n} p_{n,k} \to \sum_{k=1}^{\infty} p_k + \lambda, \quad n \to \infty. \tag{5.17}$$

In this case, we further have

$$1 \geq p_1 \geq p_2 \geq \cdots , \tag{5.18}$$

and

$$\# \overset{d}{=} Y + \sum_{k=1}^{\infty} X_k, \tag{5.19}$$

where $X_k \sim \mathrm{Bern}(p_k)$, independently across k, and $Y \sim \mathrm{Poisson}(\lambda)$.

The proof of Lemma 5.6.6 appears in Appendix 5.B; this lemma is essentially a special case of a more general result in Appendix 5.A.

In this direction of the proof of Theorem 5.6.5, we want to show that if we assume that the probability of a feature allocation takes EFPF form, then the allocation has the same distribution as if it were generated according to a feature frequency model with a Poisson-distributed number of singleton features for each n. To see how Lemma 5.6.6 may be useful, we let $\hat{\#}$ be the number of features in which index 1 occurs. Recall that in order to use the EFPF, we apply a uniform random ordering to the features of our feature allocation. Examining $\hat{\#}$ is advantageous since it is invariant to the ordering of the features, and we can thereby avoid complicated considerations that may arise related to the feature ordering and consistency of ordering across feature allocations of increasing index sets.

Indeed, recall that once we have chosen a uniform random ordering for the features, the EFPF assumption tells us that any feature allocation with the requisite feature sizes and number of indices has the same probability. Let K_N be the number of features containing indices $[N]$. If $M_{N,k}$ is the size of the kth feature (under the uniform random ordering) after N indices, then there are

$$\binom{N}{M_{N,1}} \cdots \binom{N}{M_{N,K_N}}$$

such configurations. $M_{N,1}/N$ have index 1 in the first feature. For each such allocation, there are equally many configurations of the remaining features. So, for each such allocation, $M_{N,2}/N$ have index 1 in the second feature. And so on. That is, we have that, conditionally on the feature sizes, the number of features with index 1 has the same distribution as a sum of Bernoulli random variables:

$$\sum_{k=1}^{K_N} \tilde{X}_{N,k}, \quad \tilde{X}_{N,k} \overset{indep}{\sim} \text{Bern}(M_{N,k}/N). \tag{5.20}$$

First, we note that the feature sizes are sufficient for the distribution by the EFPF assumption. So we may, in fact, condition on ξ_N, which we define to be the sigma-field of events invariant under permutations of the indices $n = 1, \ldots, N$. That is, $\hat{\#}|\xi_N$ has the same distribution as the sum in Eq. (5.20).

Second, we note that the sum in Eq. (5.20) has no dependence on the ordering of the features. In particular, then, let $1 \geq p_{N,1} \geq p_{N,2} \geq \cdots \geq p_{N,K_N}$ be the sizes of the features divided by N and ordered so as to be monotonically decreasing. Again, note that we are only considering those features including some data index in $[N]$. It follows that

$$\hat{\#}|\xi_N \overset{d}{=} \sum_{k=1}^{K_N} \tilde{X}_{N,k}, \quad \tilde{X}_{N,k} \overset{indep}{\sim} \text{Bern}(p_{N,k}). \tag{5.21}$$

So we see that we have circumvented ordering concerns and can simply use a size ordering in what follows.

At this point, it seems natural to apply Lemma 5.6.6 to $\hat{\#}|\xi_N$. To do so, we need to show that $\hat{\#}|\xi_N$ converges in distribution to some random variable with non-negative integer values as $N \to \infty$. To that end, we note that (ξ_N) is a reversed filtration: $\xi_N \supseteq \xi_{N+1}$ for all N. And further $\mathbb{P}(\hat{\#} = j|\xi_N)$ is a reversed martingale since (1) $\mathbb{P}(\hat{\#} = j|\xi_N)$ is measurable with respect to ξ_N; (2) $\mathbb{P}(\hat{\#} = j|\xi_N)$ is integrable; and (3) by the tower law, $\mathbb{E}[\mathbb{P}(\hat{\#} = j|\xi_N)|\xi_{N+1}] = \mathbb{P}(\hat{\#} = j|\xi_{N+1})$. It follows that

$$\mathbb{P}(\hat{\#} = j|\xi_N) \xrightarrow{\text{a.s.}} \mathbb{P}(\hat{\#} = j|\xi_\infty)$$

and hence

$$\hat{\#}|\xi_N \xrightarrow{d} \hat{\#}|\xi_\infty \quad \text{a.s.}$$

for $\xi_\infty = \bigcap_{n=1}^\infty \xi_n$ by reverse martingale convergence.

So we may apply Lemma 5.6.6 conditional on ξ_∞. By the lemma, we have that, conditional on ξ_∞,

$$\begin{aligned} \hat{\#} &\overset{d}{=} Y + \sum_{k=1}^\infty X_k \\ Y &\sim \text{Poisson}(\lambda) \\ X_k &\overset{indep}{\sim} \text{Bern}(p_k) \end{aligned}$$

for some $\lambda \geq 0$ and some $1 \geq p_1 \geq p_2 \geq \cdots$. The conditioning on ξ_∞ means that, in general, λ and the frequencies $1 \geq p_1 \geq p_2 \geq \cdots$ may be positive random variables, as was to be shown. □

5.7 Discussion

It has been known for some time that the class of exchangeable partitions is the same as the class of partitions generated by the Kingman paintbox, which is in turn the same as the class of partitions with exchangeable partition probability functions (EPPFs). In this chapter, we have developed an analogous set of concepts for the feature allocation problem. We defined a feature allocation as an extension of partitions in which indices may belong to multiple groups, now called features. We have developed analogues of the EPPF and the Kingman paintbox, which we refer to as the exchangeable feature partition function (EFPF) and the feature paintbox, respectively. The feature paintbox allows us to construct a feature allocation via iid draws from an underlying collection of sets in the unit interval. In the special cases of partitions and feature frequency models the construction of these sets is particularly straightforward.

The Venn diagram presented earlier in Figure 5.1 summarizes our results and also suggests a number of open areas for further investigation. In particular it would be useful to develop a fuller understanding of the regularity condition on feature allocations that allows the connection to the feature paintbox. It would also be of interest to carry the program further by exploring generalizations of the partition and feature allocation framework to other combinatorial representations, such as the setting in which we allow multiplicity within, as well as across, features (Broderick, Mackey, et al., 2014; Zhou et al., 2012).

5.A Intermediate lemmas leading to Lemma 5.6.6

To prove Lemma 5.6.6, we will make use of a few definitions and lemmas. We start with two definitions. First, suppose we have constants $p_1, p_2, p_3, \ldots$ such that

$$1 \geq p_1 \geq p_2 \geq p_3 \geq \ldots \geq 0$$

and a constant λ such that $0 \leq \lambda < \infty$. Then we say that the random variable # has the *extended Poisson-binomial distribution* with parameters $(\lambda, p_1, p_2, \ldots)$ if there exist independent random variables $X_0, X_1, X_2, \ldots$ with

$$\begin{aligned} X_0 &\sim \text{Poisson}(\lambda) \\ X_k &\sim \text{Bern}(p_k), \quad k = 1, 2, \ldots \end{aligned}$$

such that

$$\# = X_0 + \sum_{k=1}^{\infty} X_k.$$

The terminology "extended Poisson-binomial distribution" is motivated by the familiar *Poisson-binomial distribution* (Y. H. Wang, 1993; S. X. Chen and Liu, 1997; O. Johnson, Kontoyiannis, and Madiman, 2011), which describes the special case of the above where $\lambda = 0$ and $p_k = 0$ for all $k > K$ for some finite K.

Second, we say that μ is the *spike size-location measure* with parameters $(\lambda, p_1, p_2, \ldots)$ if μ puts mass λ at 0 and mass p_k at p_k for $k = 1, 2, \ldots$. With these definitions in hand, we can state the following lemmas.

Lemma 5.A.1. *Let # have the extended Poisson-binomial distribution with parameters* $(\lambda, p_1, p_2, \ldots)$.

Then

1. *# is a.s. finite if and only if* $\sum_{k=1}^{\infty} p_k < \infty$.
2. *If # is a.s. finite, then the parameters* $(\lambda, p_1, p_2, \ldots)$ *are uniquely determined by the distribution of #.*

In particular, since the parameters $(\lambda, p_1, p_2, \ldots)$ uniquely determine the distribution of #, Lemma 5.A.1 tells us that there is a bijection between the distribution of # and the parameters $(\lambda, p_1, p_2, \ldots)$ when # is a.s. finite. See Appendix 5.C for the proof of Lemma 5.A.1.

The next lemma tells us that this correspondence between distributions and parameters is also continuous in a sense.

Lemma 5.A.2. *For* $n = 1, 2, \ldots$, *let* $\#_n$ *have the extended Poisson-binomial distribution with parameters* $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$. *Let* μ_n *be the spike size-location measure with parameters* $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$.

Then the following two statements are equivalent:

1. *$\#_n$ converges in distribution to a finite-valued limit random variable.*

2. *μ_n converges weakly to some finite measure on $[0, 1]$.*

If the convergence holds, the limiting random variable (call it $\#$) has an extended Poisson-binomial distribution, and the limiting measure (call it μ) is a spike size-location measure. In this case, $\#$ and μ have the same parameters; call the parameters $(\lambda, p_1, p_2, \ldots)$.

This lemma is suggested by, and provides an extension to, previous results on triangular arrays of random variables with row sums converging in distribution; cf. Kallenberg (2002). See Appendix 5.D for the proof of Lemma 5.A.2.

Lemma 5.6.6 highlights a special case of Lemmas 5.A.1 and 5.A.2 that we use to prove the equivalence in Theorem 5.6.5.

5.B Proof of Lemma 5.6.6

We can rephrase the statement of Lemma 5.6.6 in terms of the terminology introduced in Appendix 5.A. In particular, we are given a sequence of random variables $\#_n$, where $\#_n$ has an extended Poisson-binomial distribution with parameters

$$(0, p_{n,1}, p_{n,2}, \ldots, p_{n,K_n}, 0, 0, \ldots).$$

Then we see that Lemma 5.6.6 is essentially a special case of Lemma 5.A.2 where λ_n and all but finitely many of the $p_{n,k}$ are equal to zero; this special case is exactly the usual Poisson-binomial distribution.

(1) ⇒ (2). We assume that $\#_n$ converges in distribution to some finite-valued random variable $\#$, and we wish to show that the $p_{n,k}$ converge to some limiting p_k as $n \to \infty$ for each k, and likewise that $\sum_{k=1}^{K_n} p_{n,k}$ converges to $\sum_{k=1}^{\infty} p_k + \lambda$ for some non-negative constant λ. The $p_{n,k}$ are just the ordered atom sizes of the spike size-location measures μ_n in Lemma 5.A.2. By Lemma 5.A.2, the μ_n converge weakly to some spike size-location measure μ.

Denote the parameters of μ by $(\lambda, p_1, p_2, \ldots)$. The convergence of μ_n to μ yields both the desired convergence of the atom sizes (Eq. (5.16), repeated here)

$$p_{n,k} \to p_k, \quad n \to \infty$$

and the desired convergence of the total mass of μ_n (Eq. (5.17), repeated here)

$$\sum_{k=1}^{K_n} p_{n,k} \to \sum_{k=1}^{\infty} p_k + \lambda, \quad n \to \infty.$$

(2) ⇒ (1). Now we assume that the $p_{n,k}$ converge to some limiting p_k as $n \to \infty$ for each k, and likewise that $\sum_{k=1}^{K_n} p_{n,k}$ converges to $\sum_{k=1}^{\infty} p_k + \lambda$ for some appropriate positive constants $\{p_k\}, \lambda$. We wish to show that $\#_n$ converges in distribution to some finite-valued random variable $\#$.

The assumed convergences guarantee the weak convergence of the spike size-location measures μ_n to some finite measure on $[0,1]$. Lemma 5.A.2 then guarantees that $\#_n$ converges in distribution to some finite-valued random variable $\#$.

Assume (1) and (2). We wish to show that $1 \ge p_1 \ge p_2 \ge \ldots$ (Eq. (5.18)), but this result follows from the monotonicity of the $p_{n,k}$.

Eq. (5.19) in the original lemma statement can be rephrased as wanting to show that $\#$ has the extended Poisson-binomial distribution with parameters $(\lambda, p_1, p_2, \ldots)$. This follows directly from the final statement in Lemma 5.A.2 and our identification of the limiting spike size-location measure μ as having parameters $(\lambda, p_1, p_2, \ldots)$ in a previous part of this proof ("(1) ⇒ (2)"). □

5.C Proof of Lemma 5.A.1

Throughout we assume that $\#$ has the extended Poisson-binomial distribution with parameters $(\lambda, p_1, p_2, \ldots)$.

(1). We want to show that $\#$ is a.s. finite if and only if $\sum_{k=1}^{\infty} p_k < \infty$. Since $\#$ is extended Poisson-binomially distributed, we can write $\# = X_0 + \sum_{k=1}^{\infty} X_k$ for independent $X_0 \sim \text{Poisson}(\lambda)$ and $X_k \sim \text{Bern}(p_k)$ for $k = 1, 2, \ldots$. First suppose $\sum_{k=1}^{\infty} p_k < \infty$. Then $\sum_{k=1}^{\infty} X_k$ is a.s. finite by the Borel-Cantelli lemma. Second, suppose $\sum_{k=1}^{\infty} p_k = \infty$. Then $\sum_{k=1}^{\infty} X_k$ is a.s. infinite by the second Borel-Cantelli lemma. Since X_0 is a.s. finite by construction, the result follows.

(2). We want to show that if $\#$ is a.s. finite, then the parameters $(\lambda, p_1, p_2, \ldots)$ are uniquely determined by the distribution of $\#$. To that end, let μ be the spike size-location measure with parameters $(\lambda, p_1, p_2, \ldots)$. Note that μ need not be a probability measure but is finite by the assumption that $\#$ is a.s. finite together with part (1) of the lemma.

To better understand the distribution of $\#$, we write the probability generating function of $\#$. For s with $0 \le s \le 1$, we have

$$\mathbb{E}s^{\#} = e^{-\lambda(1-s)} \prod_{k=1}^{\infty} \left[1 - (1-s)p_k\right],$$

which implies that for s with $0 < s \le 1$ we have

$$-\log \mathbb{E}s^{\#} = \lambda(1-s) - \sum_{k=1}^{\infty} \log\left[1 - (1-s)p_k\right] \tag{5.22}$$

$$= \lambda(1-s) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{1}{j} (1-s)^j p_k^j$$

from the Taylor series expansion of the logarithm

$$= \lambda(1-s) + \sum_{j=1}^{\infty} \frac{1}{j} (1-s)^j \sum_{k=1}^{\infty} p_k^j$$

interchanging the order of summation since the summands are non-negative

$$= (1-s)\mu\{0\} + \sum_{j=1}^{\infty} \frac{1}{j} (1-s)^j \int_{(0,1]} x^{j-1} \mu(dx) \tag{5.23}$$

$$= \sum_{j=1}^{\infty} \frac{1}{j} (1-s)^j m_{j-1}, \tag{5.24}$$

where

$$m_j := \int_{[0,1]} x^j \mu(dx)$$

is the jth moment of the measure μ.

Now the distribution of # uniquely determines the probability generating function of #, which by Eq. (5.24) uniquely determines the sequence of moments of the measure μ. In turn, μ is a bounded measure on $[0,1]$ and hence uniquely determined by its moments. And the parameters $(\lambda, p_1, p_2, \ldots)$ are uniquely determined by μ. □

5.D Proof of Lemma 5.A.2

For $n = 1, 2, \ldots$, we assume $\#_n$ has the extended Poisson-binomial distribution with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$. We further assume μ_n has the spike size-location measure with parameters $(\lambda_n, p_{n,1}, p_{n,2}, \ldots)$.

(2) ⇒ (1). Suppose the μ_n converge weakly to some finite measure μ on $[0,1]$. We want to show that $\#_n$ converges in distribution to a finite-valued limit random variable.

In Appendix 5.C, we noted that we can express the probability generating function of an extended Poisson-binomial distribution in terms of a spike size-location measure with the same parameters. In particular, by Eq. (5.23), we can write the negative log of the probability generating function of $\#_n$ as

$$-\log \mathbb{E} s^{\#_n} = \int_{[0,1]} f_s(x) \, \mu_n(dx),$$

where

$$f_s(x) := \sum_{j=1}^{\infty} \frac{1}{j} (1-s)^j x^{j-1} = \begin{cases} -x^{-1} \log\left[1-(1-s)x\right] & x > 0 \\ 1-s & x = 0 \end{cases}. \tag{5.25}$$

Since $f_s(x)$ is bounded in x for each fixed s with $0 < s \leq 1$, we have by the assumption of weak convergence of μ_n that

$$\lim_{n \to \infty} - \log \mathbb{E} s^{\#_n} = \int_{[0,1]} f_s(x) \; \mu(dx).$$

Moreover, since μ is finite by assumption, we have that the result is finite for each s with $0 < s \leq 1$. It follows that $\#_n$ converges in distribution to a finite random variable $\#$, with probability generating function given by

$$\mathbb{E} s^{\#} = \exp \left\{ - \int_{[0,1]} f_s(x) \; \mu(dx) \right\}. \tag{5.26}$$

Assume (1). Now suppose the $\#_n$ converge in distribution to a finite random variable $\#$. The next two parts of the proof will rely on an intermediate step: showing that μ_n has bounded total mass in this case.

To show that μ_n has bounded total mass, first note that $\mathbb{E}\#_n$ is exactly the total mass of μ_n:

$$\begin{aligned} \mathbb{E}\#_n &= \lambda_n + \sum_{k=1}^{\infty} p_{n,k} =: \Sigma_n, \\ \text{and } \mathrm{Var}\#_n &= \lambda_n + \sum_{k=1}^{\infty} p_{n,k}(1 - p_{n,k}). \end{aligned}$$

Noting that $\mathrm{Var}\#_n \leq \Sigma_n$ allows us to apply Chebyshev's inequality to find

$$\begin{aligned} 1/4 &\geq \mathbb{P}(|\#_n - \mathbb{E}\#_n| \geq 2\sqrt{\mathrm{Var}\#_n}) \\ 3/4 &\leq \mathbb{P}(|\#_n - \Sigma_n| \leq 2\sqrt{\mathrm{Var}\#_n}) \\ &\leq \mathbb{P}(|\#_n - \Sigma_n| \leq 2\sqrt{\Sigma_n}) \\ &\leq \mathbb{P}(\#_n \geq \Sigma_n - 2\sqrt{\Sigma_n}). \end{aligned}$$

Since $\#_n$ converges in distribution by assumption, the sequence $\#_n$ is tight. Choose ϵ such that $1/2 > \epsilon > 0$. Then there exists some N_ϵ such that, for all $n \geq 1$, we have $\mathbb{P}(\#_n \leq N_\epsilon) > 1 - \epsilon > 1/2$. It follows that, for all $n \geq 1$,

$$1/4 \leq \mathbb{P}(N_\epsilon \geq \Sigma_n - 2\sqrt{\Sigma_n}).$$

Since Σ_n is non-random, it must be that $\mathbb{P}(N_\epsilon \geq \Sigma_n - 2\sqrt{\Sigma_n}) = 1$. That is, the total mass of μ_n is bounded.

Assume (1) and (2). Suppose $\#_n$ converges in distribution to some finite-valued limit random variable $\#$ and that μ_n converges weakly to some finite measure μ. We want to show that $\#$ has an extended Poisson-binomial distribution, that μ is a spike size-location measure, and that $\#$ and μ have the same parameters.

We start by showing that μ is discrete. Choose any $\epsilon > 0$. Since the mass of μ_n is bounded across n by the previous part of the proof ("Assume (1)"), the number of atoms of μ_n greater than ϵ is bounded across n. It follows that the number of atoms of μ has the same bound. So μ is discrete. Since μ_n converges weakly to μ, we see that μ must have atoms with sizes and locations $p_1, p_2, \ldots$ such that

$$1 \geq p_1 \geq p_2 \geq \ldots$$

as well as a potential atom, with size we denote by λ, at zero. That is, μ is a spike size-location measure with parameters $(\lambda, p_1, p_2, \ldots)$.

In a previous part of the proof ("(2) $\Rightarrow$ (1)"), we expressed the probability generating function of $\#$ as a function of μ (Eq. (5.26)). With this relation in hand, we can reverse the series of equations presented in Appendix 5.C and ending in Eq. (5.23) to find the form of the probability generating function for $\#$ (Eq. (5.22)). In particular, Eq. (5.22) tells us that $\#$ is an extended Poisson-binomial random variable with parameters $(\lambda, p_1, p_2, \ldots)$. In particular, we emphasize that $\#$ has the same parameters as μ, which we have already shown above is a spike size-location measure.

(1) $\Rightarrow$ (2) Now step back and assume that $\#_n$ converges in distribution to a finite-valued limit random variable; call it $\#$. We wish to show that μ_n converges weakly to some finite measure on $[0, 1]$.

By a previous part of this proof ("Assume (1)"), the mass of μ_n is bounded across n. Moreover, by construction, all of the mass for each μ_n is concentrated on $[0, 1]$. So it must be that the sequence μ_n is tight. It follows that if every weakly convergent subsequence μ_{n_j} has the same limit μ, then μ_n converges weakly to μ.

Consider a subsequence $(n_j)_j$ of $\mathbb{N}$. We know $\#_{n_j}$ converges in distribution to $\#$ by the assumption that $\#_n$ converges in distribution to $\#$. The previous part of this proof ("Assume (1) and (2)") gives that the form of the limit of μ_{n_j} is determined by $\#$; namely, the limit is a spike size-location measure with parameters shared by $\#$. In particular, then, the limit μ must be the same for every subsequence, and the desired result is shown. □

Chapter 6

Posteriors, conjugacy, and exponential families for completely random measures

We demonstrate how to calculate posteriors for general CRM-based priors and likelihoods for Bayesian nonparametric models. We further show how to represent Bayesian nonparametric priors as a sequence of finite draws using a size-biasing approach—and how to represent full Bayesian nonparametric models via finite marginals. Motivated by conjugate priors based on exponential family representations of likelihoods, we introduce a notion of exponential families for CRMs, which we call exponential CRMs. This construction allows us to specify automatic Bayesian nonparametric conjugate priors for exponential CRM likelihoods. We demonstrate that our exponential CRMs allow particularly straightforward recipes for size-biased and marginal representations of Bayesian nonparametric models. Along the way, we prove that the gamma process is a conjugate prior for the Poisson likelihood process and the beta prime process is a conjugate prior for a process we call the odds Bernoulli process. We deliver a size-biased representation of the gamma process and a marginal representation of the gamma process coupled with a Poisson likelihood process.

6.1 Introduction

An important milestone in Bayesian analysis was the development of a general strategy for obtaining conjugate priors based on exponential family representations of likelihoods (DeGroot, 1970). While slavish adherence to exponential-family conjugacy can be criticized, conjugacy continues to occupy an important place in Bayesian analysis, for its computational tractability in high-dimensional problems and for its role in inspiring investigations into broader classes of priors (e.g., via mixtures, limits, or augmentations). The exponential family is, however, a parametric class of models, and it is of interest to consider whether similar general notions of conjugacy can be developed for Bayesian nonparametric models.

Indeed, the nonparametric literature is replete with nomenclature that suggests the exponential family, including familiar names such as "Dirichlet," "beta," "gamma," and "Poisson." These names refer to aspects of the random measures underlying Bayesian nonparametrics, either the Lévy measure used in constructing certain classes of random measures or properties of marginals obtained from random measures. In some cases, conjugacy results have been established that parallel results from classical exponential families; in particular, the Dirichlet process is known to be conjugate to a multinomial process likelihood (Ferguson, 1973), the beta process is conjugate to a Bernoulli process (Kim, 1999a; Thibaux and Jordan, 2007) and to a negative binomial process (Broderick, Mackey, et al., 2014). Moreover, various useful representations for marginal distributions, including stick-breaking and size-biased representations, have been obtained by making use of properties that derive from exponential families. It is striking, however, that these results have been obtained separately, and with significant effort; a general formalism has not yet emerged. In this chapter, we provide the single, holistic framework so strongly suggested by the nomenclature. Within this single framework, we show that it is straightforward to calculate posteriors and establish conjugacy. Our framework includes the specification of a Bayesian nonparametric analog of the finite exponential family, which allows us to provide automatic and constructive nonparametric conjugate priors given a likelihood specification as well as general recipes for marginal and size-biased representations.

A broad class of Bayesian nonparametric priors—including those built on the Dirichlet process (Ferguson, 1973), the beta process (Hjort, 1990), the gamma process (Titsias, 2008), and the negative binomial process (Zhou et al., 2012; Broderick, Mackey, et al., 2014)—can be viewed as models for the allocation of data points to traits. These processes give us pairs of traits together with rates or frequencies with which the traits occur in some population. Corresponding likelihoods assign each data point in the population to some finite subset of traits conditioned on the trait frequencies. What makes these models nonparametric is that the number of traits in the prior is countably infinite. Then the (typically random) number of traits to which any individual data point is allocated is unbounded, but also there are always new traits to which as-yet-unseen data points may be allocated. That is, such a model allows the number of traits in any data set to grow with the size of that data set.

A principal challenge of working with such models arises in posterior inference. There is a countable infinity of trait frequencies in the prior which we must integrate over to calculate the posterior of trait frequencies given allocations of data points to traits. Bayesian nonparametric models sidestep the full infinite-dimensional integration in three principal ways: conjugacy, size-biased representations, and marginalization.

In its most general form, conjugacy simply asserts that the prior is in the same family of distributions as the posterior. When the prior and likelihood are in finite-dimensional conjugate exponential families, conjugacy can turn posterior calculation into, effectively, vector addition. As a simple example, consider a model with beta-distributed prior, $\theta \sim \text{Beta}(\theta|\alpha, \beta)$, for some fixed hyperparameters α and β. For the likelihood, let each observation x_n with $n \in \{1, \ldots, N\}$ be iid Bernoulli-distributed conditional on parameter θ: $x_n \overset{iid}{\sim} \text{Bern}(x|\theta)$.

Then the posterior is simply another beta distribution, $\mathrm{Beta}(\theta|\alpha_{post}, \beta_{post})$, with parameters updated via addition: $\alpha_{post} := \alpha + \sum_{n=1}^{N} x_n$ and $\beta_{post} := \beta + N - \sum_{n=1}^{N} x_n$. While conjugacy is certainly useful and popular in the case of finite parameter cardinality, there is arguably a stronger computational imperative for its use in the infinite-parameter case. Indeed, the core prior-likelihood pairs of Bayesian nonparametrics are generally proven (Hjort, 1990; Kim, 1999a; Thibaux and Jordan, 2007; Broderick, Mackey, et al., 2014), or assumed to be (Titsias, 2008; Thibaux, 2008), conjugate. When such proofs exist, though, thus far they have been specialized to specific pairs of processes. In what follows, we demonstrate a general way to calculate posteriors for a class of distributions that includes all of these classical Bayesian nonparametric models. We also define a notion of exponential family representation for the infinite-dimensional case and show that, given a Bayesian nonparametric exponential family likelihood, we can readily construct a Bayesian nonparametric conjugate prior.

Size-biased sampling provides a finite-dimensional distribution for each of the individual prior trait frequencies (Thibaux and Jordan, 2007; Paisley, Zaas, et al., 2010). Such a representation has played an important role in Bayesian nonparametrics in recent years, allowing for either exact inference via slice sampling (Damien, Wakefield, and Walker, 1999; Neal, 2003)—as demonstrated by Teh, Görür, and Ghahramani (2007); Broderick, Mackey, et al. (2014)—or approximate inference via truncation (Doshi et al., 2009; Paisley, Carin, and Blei, 2011). This representation is particularly useful for building hierarchical models (Thibaux and Jordan, 2007). We show that our framework yields such representations in general, and we show that our construction is especially straightforward to use in the exponential family framework that we develop.

Marginal processes avoid directly representing the infinite-dimensional prior and posterior altogether by integrating out the trait frequencies. Since the trait allocations are finite for each data point, the marginal processes are finite for any finite set of data points. Again, thus far, such processes have been shown to exist separately in special cases; for example, the Indian buffet process (Griffiths and Ghahramani, 2006) is the marginal process for the beta process prior paired with a Bernoulli process likelihood (Thibaux and Jordan, 2007). We show that the integration that generates the marginal process from the full Bayesian model can be generally applied in Bayesian nonparametrics and takes a particularly straightforward form when using conjugate exponential family priors and likelihoods. We further demonstrate that, in this case, a basic, constructive recipe exists for the general marginal process in terms of only finite-dimensional distributions.

Our results are built on the general class of stochastic processes known as *completely random measures* (CRMs) (Kingman, 1967). We review CRMs in Section 6.2 and we discuss what assumptions are needed to form a full Bayesian nonparametric model from CRMs in Section 6.2. Given a general Bayesian nonparametric prior and likelihood (Section 6.2), we demonstrate in Section 6.3 how to calculate the posterior. Although the development up to this point is more general, we next introduce a concept of exponential families for CRMs (Section 6.4) and call such models *exponential CRMs*. We show that we can generate automatic conjugate priors given exponential CRM likelihoods in Section 6.4. Finally, we show how we can generate recipes for size-biased representations (Section 6.5) and marginal processes (Section 6.6), which are particularly straightforward in the exponential CRM case (Corollary 6.5.2 in Section 6.5 and Corollary 6.6.2 in Section 6.6). We illustrate our results on a number of examples and derive new conjugacy results, size-biased representations, and marginal processes along the way.

6.2 Bayesian models based on completely random measures

As we have discussed, we view Bayesian nonparametric models as being composed of two parts: (1) a collection of pairs of traits together with their frequencies or rates and (2) for each data point, an allocation to different traits. Both parts can be expressed as *random measures*. Recall that a random measure is a random element whose values are measures.

We represent each trait by a point ψ in some space Ψ of traits. Further, let θ_k be the frequency, or rate, of the trait represented by ψ_k, where k indexes the countably many traits. In particular, $\theta_k \in \mathbb{R}_+$. Then (θ_k, ψ_k) is a tuple consisting of the frequency of the kth trait together with its trait descriptor. We can represent the full collection of pairs of traits with their frequencies by the discrete measure on Ψ that places weight θ_k at location ψ_k:

$$\Theta = \sum_{k=1}^{K} \theta_k \delta_{\psi_k}, \tag{6.1}$$

where the cardinality K may be finite or infinity.

Next, we form data point X_n for the nth individual. The data point X_n is viewed as a discrete measure. Each atom of X_n represents a pair consisting of (1) a trait to which the nth individual is allocated and (2) a degree to which the nth individual is allocated to this particular trait. That is,

$$X_n = \sum_{k=1}^{K_n} x_{n,k} \delta_{\psi_{n,k}}, \tag{6.2}$$

where again $\psi_{n,k} \in \Psi$ represents a trait and now $x_{n,k} \in \mathbb{R}_+$ represents the degree to which the nth data point belongs to trait $\psi_{n,k}$. K_n is the total number of traits to which the nth data point belongs.

Here and in what follows, we treat $X_{1:N} = \{X_n : n \in [N]\}$ as our observed data points for $[N] := \{1, 2, 3, \ldots, N\}$. In practice $X_{1:N}$ is often incorporated into a more complex Bayesian hierarchical model. For instance, in topic modeling, ψ_k represents a topic; that is, ψ_k is a distribution over words in a vocabulary (Blei, Ng, and Jordan, 2003; Teh, Jordan, et al., 2006). θ_k might represent the frequency with which the topic ψ_k occurs in a corpus of documents. $x_{n,k}$ might be a positive integer and represent the number of words in topic $\psi_{n,k}$ that occur in the nth document. So the nth document has a total length of $\sum_{k=1}^{K_n} x_{n,k}$ words. In this case, the actual observation consists of the words in each document, and the topics are latent. Not only are the results concerning posteriors, conjugacy, and exponential family representations that we develop below useful for inference in such models, but in fact our results are especially useful in such models—where the traits and any ordering on the traits are not known in advance.

Next, we want to specify a full Bayesian model for our data points $X_{1:N}$. To do so, we must first define a prior distribution for the random measure Θ as well as a likelihood for each random measure X_n conditioned on Θ. We let Σ_Ψ be a σ-algebra of subsets of Ψ, where we assume all singletons are in Σ_Ψ. Then we consider random measures Θ and X_n whose values are measures on Ψ. Note that for any random measure Θ and any measurable set $A \in \Sigma_\Psi$, $\Theta(A)$ is a random variable.

Completely random measures

We can see from Eqs. (6.1) and (6.2) that we desire a distribution on random measures that yields discrete measures almost surely. A particularly simple form of random measure called a *completely random measure* has been shown to have this property (Kingman, 1967).

A completely random measure Θ is defined as a random measure that satisfies one additional property; for any disjoint, measurable sets $A_1, A_2, \ldots, A_K \in \Sigma_\Psi$, we require that $\Theta(A_1), \Theta(A_2), \ldots, \Theta(A_K)$ be independent random variables. Kingman (1967) showed that a completely random measure can always be decomposed into a sum of three independent parts:

$$\Theta = \Theta_{det} + \Theta_{fix} + \Theta_{ord}. \tag{6.3}$$

Here, Θ_{det} is the deterministic component, Θ_{fix} is the *fixed-location* component, and Θ_{ord} is the *ordinary* component. In particular, Θ_{det} is any deterministic measure. It is straightforward to include a deterministic measure in a statistical model, so—without loss of generality in our treatment and according to the prevailing norm in using models based on CRMs—in what follows we will set $\Theta_{det} \equiv 0$. We define the remaining two parts next.

The fixed-location component is called the "fixed component" by Kingman (1967), but we change the name slightly here to emphasize that Θ_{fix} is defined to be constructed from a set of random weights at fixed (i.e., deterministic) locations. That is,

$$\Theta_{fix} = \sum_{k=1}^{K_{fix}} \theta_{fix,k} \delta_{\psi_{fix,k}}, \tag{6.4}$$

where the number of fixed-location atoms, K_{fix}, may be either finite or infinity; $\psi_{fix,k}$ is deterministic, and $\theta_{fix,k}$ is a non-negative, real-valued random variable (since Φ is a measure). Without loss of generality, we assume that the locations $\psi_{fix,k}$ are all distinct. Then, by the independence assumption of CRMs, we must have that $\theta_{fix,k}$ are independent random variables across k. Although the fixed-location atoms are often ignored in the Bayesian nonparametrics literature, we will see that the fixed-location component has a key role to play in establishing Bayesian nonparametric conjugacy and in the CRM representations we present.

The third and final component is the ordinary component. Let $\#(A)$ denote the cardinality of some countable set A. Let μ be any σ-finite, deterministic measure on $\mathbb{R}_+ \times \Psi$, where $\mathbb{R}_+$ is equipped with the Borel σ-algebra and $\Sigma_{\mathbb{R}_+ \times \Psi}$ is the resulting product σ-algebra given Σ_Ψ. Recall that a *Poisson point process* with rate measure μ on $\mathbb{R}_+ \times \Psi$ is a random countable subset Π of $\mathbb{R}_+ \times \Psi$ such that two properties hold (Kingman, 1993):

1. For any $A \in \Sigma_{\mathbb{R}_+ \times \Psi}$, $\#(\Pi \cap A) \sim \text{Poisson}(\mu(A))$.
2. For any disjoint $A_1, A_2, \ldots, A_K \in \Sigma_{\mathbb{R}_+ \times \Psi}$, $\#(\Pi \cap A_1), \#(\Pi \cap A_2), \cdots, \#(\Pi \cap A_K)$ are independent random variables.

To generate an ordinary component, start with a Poisson point process on $\mathbb{R}_+ \times \Psi$, characterized by its rate measure $\mu(d\theta \times d\psi)$. This process yields Π, a random and countable set of points: $\Pi = \{(\theta_{ord,k}, \psi_{ord,k})\}_{k=1}^{K_{ord}}$, where K_{ord} may be finite or infinity. Form the ordinary component measure by letting $\theta_{ord,k}$ be the weight of the atom located at $\psi_{ord,k}$:

$$\Theta_{ord} = \sum_{k=1}^{K_{ord}} \theta_{ord,k} \delta_{\psi_{ord,k}}. \tag{6.5}$$

Recall that we stated at the start of Section 6.2 that CRMs yield a.s. discrete random measures. To check this assertion, note that Θ_{fix} is a.s. discrete by construction (Eq. (6.4)) and Θ_{ord} is a.s. discrete by construction (Eq. (6.5)). When we set $\Theta_{det} \equiv 0$ as above, we are left with $\Theta = \Theta_{fix} + \Theta_{ord}$ by Eq. (6.3). So Θ is also discrete, as desired.

Prior and likelihood

The prior that we place on Θ will be a fully general CRM (minus any deterministic component) with one additional assumption on the rate measure of the ordinary component. That is, before incorporating the additional assumption, we say that Θ has a fixed-location component with K_{fix} atoms, where the kth atom has arbitrary distribution $F_{fix,k}$: $\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$. K_{fix} may be finite or infinity, and Θ has an ordinary component characterized by rate measure $\mu(d\theta \times d\psi)$. The additional assumption we make is that the distribution on the weights in the ordinary component is assumed to be decoupled from the distribution on the locations. The locations are typically more interesting in other parts of a full Bayesian model hierarchy and have been discussed extensively elsewhere (Neal, 2000; C. Wang and Blei, 2013). Moreover, it is the weights that affect the allocation of data points to traits. So henceforth we assume that the rate measure decomposes as

$$\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi), \tag{6.6}$$

where ν is any σ-finite, deterministic measure on $\mathbb{R}_+$ and G is any proper distribution on Ψ.

Given the factorization of μ in Eq. (6.6), an the ordinary component of Θ can be generated by letting $\{\theta_{fix,k}\}_{k=1}^{K_{ord}}$ be the points of a Poisson point process generated on $\mathbb{R}_+$ with rate ν.[1]

[1] Recall that K_{ord} may be finite or infinity depending on ν and is random when taking finite values.

We then draw the locations $\{\psi_{fix,k}\}_{k=1}^{K_{ord}}$ iid according to $G(d\psi)$: $\psi_{fix,k} \overset{iid}{\sim} G(d\psi)$. Finally, for each k, $\theta_{fix,k}\delta_{\psi_{fix,k}}$ is an atom in Θ_{ord}. This factorization will allow us to focus our attention on the trait frequencies, and not the trait locations, in what follows. Moreover, going forward, we will assume G is diffuse (i.e., G has no atoms) so that the ordinary component atoms are all at a.s. distinct locations, which are further a.s. distinct from the fixed locations.

Since we have seen that Θ is an a.s. discrete random measure, we can write it as

$$\Theta = \sum_{k=1}^{K} \theta_k \delta_{\psi_k}, \tag{6.7}$$

where $K := K_{fix} + K_{ord}$ may be finite or infinity and every ψ_k is a.s. unique. That is, we will sometimes find it helpful notationally to use Eq. (6.7) instead of separating the fixed and ordinary components. At this point, we have specified the prior for Θ in our general model.

Next, we specify the likelihood; i.e., we specify how to generate the data points X_n given Θ. We will assume each X_n is generated iid given Θ across the data indices n. We will let X_n be a CRM with only a fixed-location component given Θ. In particular, the atoms of X_n will be located at the atom locations of Θ, which are fixed when we condition on Θ:

$$X_n := \sum_{k=1}^{K} x_{n,k} \delta_{\psi_k}.$$

Here, $x_{n,k}$ is drawn according to some distribution H that may take θ_k, the weight of Θ at location ψ_k, as a parameter; i.e.,

$$x_{n,k} \overset{indep}{\sim} H(dx|\theta_k) \quad \text{independently across } n \text{ and } k. \tag{6.8}$$

Note that while every atom of X_n is located at an atom of Θ, it is not necessarily the case that every atom of Θ has a corresponding atom in X_n. In particular, if $x_{n,k}$ is zero for any k, there is no atom in X_n at ψ_k.

Bayesian nonparametrics

So far we have described a prior and likelihood that may be used to form a Bayesian model. We have already stated above that forming a *Bayesian nonparametric* model imposes some restrictions on the prior and likelihood. We formalize these restrictions in Assumptions A0, A1, and A2 below.

Recall that the premise of Bayesian nonparametrics is that the number of traits represented in a collection of data can grow with the number of data points. More explicitly, we achieve the desideratum that the number of traits is unbounded, and may always grow as new data points are collected, by modeling a countable infinity of traits. This assumption requires that the prior have a countable infinity of atoms. These must either be fixed-location atoms or ordinary component atoms. Fixed-location atoms represent known traits in some sense since we must know the fixed locations of the atoms in advance. Conversely, ordinary component atoms represent unknown traits, as yet to be discovered, since both their locations and associated rates are unknown a priori. Since we cannot know (or represent) a countable infinity of traits a priori, we cannot start with a countable infinity of fixed-location atoms.

A0. The number of fixed-location atoms in Θ is finite.

Since we require a countable infinity of traits in total and they cannot come from the fixed-location atoms by Assumption A0, the ordinary component must contain a countable infinity of atoms. This assumption will be true if and only if the rate measure on the trait frequencies has infinite mass.

A1. $\nu(\mathbb{R}_+) = \infty$.

Finally, an implicit part of the starting premise is that each data point be allocated to only a finite number of traits; we do not expect to glean an infinite amount of information from finitely represented data. Thus, we require that the number of atoms in every X_n be finite. By Assumption A0, the number of atoms in X_n that correspond to fixed-location atoms in Θ is finite. But by Assumption A1, the number of atoms in Θ from the ordinary component is infinite. So there must be some restriction on the distribution of values of X at the atoms of Θ (that is, some restriction on H in Eq. (6.8)) such that only finitely many of these values are nonzero.

In particular, note that if $H(dx|\theta)$ does not contain an atom at zero for any θ, then a.s. every one of the countable infinity of atoms of X will be nonzero. One consequence of this observation is that $H(dx|\theta)$ cannot be purely continuous for all θ. Though this line of reasoning does not necessarily preclude a mixed continuous and discrete H, we henceforth assume that $H(dx|\theta)$ is discrete, with support $\mathbb{Z}_* = \{0, 1, 2, \ldots\}$, for all θ.

In what follows, we write $h(x|\theta)$ for the probability mass function of x given θ. So our requirement that each data point be allocated to only a finite number of traits translates into a requirement that the number of atoms of X_n with values in $\mathbb{Z}_+ = \{1, 2, \ldots\}$ be finite. Note that, by construction, the pairs $\{(\theta_{ord,k}, x_{ord,k})\}_{k=1}^{K_{ord}}$ form a marked Poisson point process with rate measure $\mu_{mark}(d\theta \times dx) := \nu(d\theta)h(x|\theta)$. And the pairs with $x_{ord,k}$ equal to any particular value $x \in \mathbb{Z}_+$ further form a thinned Poisson point process with rate measure $\nu_x(d\theta) := \nu(d\theta)h(x|\theta)$. In particular, the number of atoms of X with weight x is Poisson-distributed with mean $\nu_x(\mathbb{R}_+)$. So the number of atoms of X is finite if and only if the following assumption holds.

A2. $\sum_{x=1}^{\infty} \nu_x(\mathbb{R}_+) < \infty$ for $\nu_x := \nu(d\theta)h(x|\theta)$.

Thus Assumptions A0, A1, and A2 capture our Bayesian nonparametric desiderata. We illustrate the development so far with an example.

Example 6.2.1. The *beta process* (Hjort, 1990) provides an example distribution for Θ. In its most general form, sometimes called the *three-parameter beta process* (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012), the beta process has an ordinary component whose weight rate measure has a beta distribution kernel,

$$\nu(d\theta) = \gamma\theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}d\theta, \tag{6.9}$$

with support on $(0,1]$. Here, the three fixed hyperparameters are γ, the *mass parameter*; c, the *concentration parameter*; and α, the *discount parameter*.[2] Moreover, each of its K_{fix} fixed-location atoms, $\theta_k\delta_{\psi_k}$, has a beta-distributed weight (Broderick, Mackey, et al., 2014):

$$\theta_{fix,k} \sim \text{Beta}(\theta|\rho_{fix,k}, \sigma_{fix,k}), \tag{6.10}$$

where $\rho_{fix,k}, \sigma_{fix,k} > 0$ are fixed hyperparameters of the model.

By Assumption A0, K_{fix} is finite. By Assumption A1, $\nu(\mathbb{R}_+) = \infty$. To achieve this infinite-mass restriction, the beta kernel in Eq. (6.9) must be improper; i.e., either $-\alpha \leq 0$ or $c + \alpha \leq 0$. Also, note that we must have $\gamma > 0$ since ν is a measure (and the case $\gamma = 0$ would be trivial).

Often the beta process is used as a prior paired with a *Bernoulli process* likelihood (Thibaux and Jordan, 2007). The Bernoulli process specifies that, given $\Theta = \sum_{k=1}^{\infty}\theta_k\delta_{\psi_k}$, we draw

$$x_{n,k} \overset{indep}{\sim} \text{Bern}(x|\theta_k),$$

which is well-defined since every atom weight θ_k of Θ is in $(0,1]$ by the beta process construction. Thus,

$$X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}.$$

Finally, then, we may apply Assumption A2, which specifies that the number of atoms in each observation X_n is finite; in this case, the assumption means

$$\begin{aligned}
&\sum_{x=1}^{\infty}\int_{\theta\in\mathbb{R}_+}\nu(d\theta)\cdot h(x|\theta)\\
&= \int_{\theta\in(0,1]}\nu(d\theta)\cdot h(1|\theta)\\
&\text{since } \theta \text{ is supported on } (0,1] \text{ and } x \text{ is supported on } \{0,1\}\\
&= \int_{\theta\in(0,1]}\gamma\theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}d\theta\cdot\theta
\end{aligned}$$

[2] In (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012), the ordinary component features the beta distribution kernel in Eq. (6.9) multiplied not only by γ but also by a more complex, positive, real-valued expression in c and α. Since all of γ, c, and α are fixed hyperparameters, and γ is an arbitrary positive real value, any other constant factors containing the hyperparameters can be absorbed into γ, as in the main text here.

$$
\begin{aligned}
&= \gamma \int_{\theta \in (0,1]} \theta^{1-\alpha-1} (1-\theta)^{c+\alpha-1} d\theta \\
&< \infty.
\end{aligned}
$$

The integral here is finite if and only if $1-\alpha$ and $c+\alpha$ are the parameters of a proper beta distribution: i.e., if and only if $\alpha < 1$ and $c > -\alpha$. Together with the restrictions above, these restrictions imply the following allowable parameter ranges for the beta process fixed hyperparameters:

$$
\begin{aligned}
\gamma &> 0 \\
\alpha &\in [0,1) \\
c &> -\alpha \\
\rho_{fix,k}, \sigma_{fix,k} &> 0 \quad \text{for all } k \in [K_{fix}].
\end{aligned}
\tag{6.11}
$$

These correspond to the hyperparameter ranges previously found in (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2012). ■

6.3 Posteriors

In Section 6.2, we defined a full Bayesian model consisting of a CRM prior for Θ and a CRM likelihood for an observation X conditional on Θ. Now we would like to calculate the posterior distribution of $\Theta | X$.

Theorem 6.3.1 (Bayesian nonparametric posteriors). *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$ with*

$$\theta_{fix,k} \overset{indep}{\sim} F_{fix,k}(d\theta)$$

for proper distribution $F_{fix,k}$ and deterministic $\psi_{fix,k}$. The ordinary component of Θ has rate measure

$$\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi),$$

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X be generated conditional on Θ according to $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ with $x_k \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X and Θ jointly satisfy Assumption A2 so that

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) < \infty.$$

Then let Θ_{post} be a random measure with the distribution of $\Theta | X$. Θ_{post} is a completely random measure with three parts.

1. *For each $k \in [K_{fix}]$, Θ_{post} has a fixed-location atom at $\psi_{fix,k}$ with weight $\theta_{post,fix,k}$ distributed according to the finite-dimensional posterior $F_{post,fix,k}(d\theta)$ that comes from prior $F_{fix,k}$, likelihood h, and observation $X(\{\psi_{fix,k}\})$.*

2. *Let $\{x_{new,k}\delta_{\psi_{new,k}} : k \in [K_{new}]\}$ be the atoms of X that are not at fixed-locations in the prior of Θ. K_{new} is finite by Assumption A2. Then Θ_{post} has a fixed-location atom at $x_{new,k}$ with random weight $\theta_{post,new,k}$, whose distribution $F_{post,new,k}(d\theta)$ is proportional to*
$$\nu(d\theta)h(x_{new,k}|\theta).$$

3. *The ordinary component of Θ_{post} has rate measure*
$$\nu_{post}(d\theta) := \nu(d\theta)h(0|\theta).$$

Proof. To prove the theorem, we consider in turn each of the two parts of the prior: the fixed-location component and the ordinary component. First, consider any fixed-location atom, $\theta_{fix,k}\delta_{\psi_{fix,k}}$, in the prior. All of the other fixed-location atoms in the prior, as well as the prior ordinary component, are drawn independently from the weight $\theta_{fix,k}$. So it follows that all of X except $x_{fix,k} := X(\{\psi_{fix,k}\})$ is independent of $\theta_{fix,k}$. Thus the posterior has a fixed atom located at $\psi_{fix,k}$ whose weight, which we denote $\theta_{post,fix,k}$, has distribution

$$F_{post,fix,k}(d\theta) \propto F_{fix,k}(d\theta)h(x_{fix,k}|\theta),$$

which follows from the usual finite Bayes Theorem.

Next, consider the ordinary component in the prior. Let

$$\Psi_{fix} = \{\psi_{fix,1}, \ldots, \psi_{fix,K_{fix}}\}$$

be the set of fixed-location atoms in the prior. Recall that Ψ_{fix} is deterministic, and since G is continuous, all of the fixed-location atoms and ordinary component atoms of Θ are at a.s. distinct locations. So the measure X_{fix} defined by

$$X_{fix}(A) := X(A \cap \Psi_{fix})$$

can be derived purely from X, without knowledge of Θ. It follows that the measure X_{ord} defined by

$$X_{ord}(A) := X(A \cap (\Psi \backslash \Psi_{fix}))$$

can be derived purely from X without knowledge of Θ. X_{ord} is the same as the observed data point X but with atoms only at atoms of the ordinary component of Θ and not at the fixed-location atoms of Θ.

Now for any value $x \in \mathbb{Z}_+$, let

$$\{\psi_{new,x,1}, \ldots, \psi_{new,x,K_{new,x}}\}$$

be all of the locations of atoms of size x in X_{ord}. By Assumption A2, the number of such atoms, $K_{new,x}$, is finite. Further let $\theta_{new,x,k} := \Theta(\{\psi_{new,x,k}\})$. Then the values $\{\theta_{new,x,k}\}_{k=1}^{K_{new,x}}$ are generated from a thinned Poisson point process with rate measure

$$\nu_x(d\theta) := \nu(d\theta)h(x|\theta). \tag{6.12}$$

And since $\nu_x(\mathbb{R}_+) < \infty$ by assumption, each $\theta_{new,x,k}$ has distribution equal to the normalized rate measure in Eq. (6.12). Note that $\theta_{new,x,k}\delta_{\psi_{new,x,k}}$ is a fixed-location atom in the posterior now that its location is known from the observed X_{ord}.

By contrast, if a likelihood draw at an ordinary component atom in the prior returned a zero, that atom is not observed in X_{ord}. Such atom weights in Θ_{post} thus formed a marked Poisson point process with rate measure

$$\nu(d\theta)h(0|\theta),$$

as was to be shown. □

In Theorem 6.3.1, we consider generating Θ and then a single data point X conditional on Θ. Now suppose we generate Θ and then N data points, $X_1, \ldots, X_N$, iid conditional on Θ. In this case, Theorem 6.3.1 may be iterated to find the posterior $\Theta|X_{1:N}$. We now illustrate the results of the theorem with an example.

Example 6.3.2. Suppose we again start with a beta process prior for Θ as in Example 6.2.1. This time we consider a *negative binomial process likelihood* (Zhou et al., 2012; Broderick, Mackey, et al., 2014). The negative binomial process specifies that, given $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, we draw $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ with

$$x_k \overset{indep}{\sim} \mathrm{NegBin}(x|r, \theta_k),$$

for some fixed hyperparameter $r > 0$. So

$$X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}.$$

In this case, Assumption A2 translates into the following restriction.

$$\begin{aligned}
&\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot h(x|\theta) \\
&= \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot [1 - h(0|\theta)] \\
&= \int_{\theta \in (0,1]} \gamma \theta^{-\alpha-1}(1-\theta)^{c+\alpha-1} d\theta \cdot [1 - (1-\theta)^r] \\
&\text{since the support of } \nu(d\theta) \text{ is } (0,1]
\end{aligned}$$

$$< \infty.$$

Since $1-(1-\theta)^r$ is asymptotically equivalent to $r\theta$ as $\theta \to 0$, we require

$$\int_{\theta \in (0,1]} \theta^{1-\alpha-1}(1-\theta)^{c+\alpha-1} d\theta < \infty,$$

which is satisfied if and only if $1-\alpha$ and $c+\alpha$ are the parameters of a proper beta distribution. Thus, we have the same parameter restrictions as in Eq. (6.11).

Now we calculate the posterior given the beta process prior on Θ and the negative binomial process likelihood for X conditional on Θ. In particular, the posterior has the distribution of Θ_{post}, a CRM with three parts given by Theorem 6.3.1.

First, at each fixed atom $\psi_{fix,k}$ of the prior with weight $\theta_{fix,k}$ given by Eq. (6.10), there is a fixed atom in the posterior with weight $\theta_{post,fix,k}$. Let $x_{post,fix,k} := X(\{\psi_{fix,k}\})$. Then $\theta_{post,fix,k}$ has distribution

$$\begin{aligned} F_{post,fix,k}(d\theta) &\propto F_{fix}(d\theta) \cdot h(x_{post,fix,k}|\theta) \\ &= \mathrm{Beta}(\theta|\rho_{fix,k}, \sigma_{fix,k})\ d\theta \cdot \mathrm{NegBin}(x_{post,fix,k}|r,\theta) \\ &\propto \theta^{\rho_{fix,k}-1}(1-\theta)^{\sigma_{fix,k}-1}\ d\theta \cdot \theta^{x_{post,fix,k}}(1-\theta)^r \\ &\propto \mathrm{Beta}\left(\theta\,|\rho_{fix,k} + x_{post,fix,k}, \sigma_{fix,k} + r\right)\ d\theta. \end{aligned} \tag{6.13}$$

Second, for any atom $x_{new,k}\delta_{\psi_{new,k}}$ in X that is not at a fixed location in the prior, Θ_{post} has a fixed atom at $\psi_{new,k}$ whose weight $\theta_{post,new,k}$ has distribution

$$\begin{aligned} F_{post,new,k}(d\theta) &\propto \nu(d\theta) \cdot h(x_{new,k}|\theta) \\ &= \nu(d\theta) \cdot \mathrm{NegBin}(x_{new,k}|r,\theta) \\ &\propto \theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}\ d\theta \cdot \theta^{x_{new,k}}(1-\theta)^r \\ &\propto \mathrm{Beta}\left(\theta\,|-\alpha + x_{new,k}, c+\alpha+r\right)\ d\theta, \end{aligned} \tag{6.14}$$

which is a proper distribution since we have the following restrictions on its parameters. For one, by assumption, $x_{new,k} \geq 1$. And further, by Eq. (6.11), we have $\alpha \in [0,1)$ as well as $c+\alpha > 0$ and $r > 0$.

Third, the ordinary component of Θ_{post} has rate measure

$$\begin{aligned} \nu(d\theta)h(0|\theta) &= \gamma\theta^{-\alpha-1}(1-\theta)^{c+\alpha-1}\ d\theta \cdot (1-\theta)^r \\ &= \gamma\theta^{-\alpha-1}(1-\theta)^{c+r+\alpha-1}\ d\theta. \end{aligned}$$

Not only have we found the posterior distribution Θ_{post} above, but now we can note that the posterior is in the same form as the prior with updated ordinary component hyperparameters:

$$\begin{aligned} \gamma_{post} &= \gamma \\ \alpha_{post} &= \alpha \end{aligned}$$

$$c_{post} = c + r.$$

The posterior also has old and new beta-distributed fixed atoms with beta distribution hyperparameters given in Eq. (6.13) and Eq. (6.14), respectively. Thus, we have proven that the beta process is, in fact, conjugate to the negative binomial process. An alternative proof was first given by Broderick, Mackey, et al. (2014). ■

As in Example 6.3.2, we can use Theorem 6.3.1 not only to calculate posteriors but also, once those posteriors are calculated, to check for conjugacy. This approach unifies existing disparate approaches to Bayesian nonparametric conjugacy. However, it still requires the practitioner to guess the right conjugate prior for a given likelihood. In the next section, we define a notion of exponential families for CRMs, and we show how to automatically construct a conjugate prior for any exponential family likelihood.

6.4 Exponential families

Exponential families are what typically make conjugacy so powerful in the finite case. For one, when a finite likelihood belongs to an exponential family, then existing results give an automatic conjugate, exponential family prior for that likelihood. In this section, we review finite exponential families, define *exponential CRMs*, and show that analogous automatic conjugacy results can be obtained for exponential CRMs. Our development of exponential CRMs will also allow particularly straightforward results for size-biased representations (Corollary 6.5.2 in Section 6.5) and marginal processes (Corollary 6.6.2 in Section 6.6).

In the finite-dimensional case, suppose we have some (random) parameter θ and some (random) observation x whose distribution is conditioned on θ. We say the distribution $H_{exp,like}$ of x conditional on θ is in an exponential family if

$$\begin{aligned} H_{exp,like}(dx|\theta) &= h_{exp,like}(x|\theta)\ dx \\ &= \kappa(x) \exp\left\{\langle \eta(\theta), \phi(x)\rangle - A(\theta)\right\}\ \mu(dx), \end{aligned} \tag{6.15}$$

where $\eta(\theta)$ is the *natural parameter*, $\phi(x)$ is the *sufficient statistic*, $\kappa(x)$ is the *base density*, and $A(\theta)$ is the *log partition function*. We denote the density of $H_{exp,like}$ here, which exists by definition, by $h_{exp,like}$. The measure μ—with respect to which the density $h_{exp,like}$ exists—is typically Lebesgue measure when $H_{exp,like}$ is diffuse or counting measure when $H_{exp,like}$ is atomic. $A(\theta)$ is determined by the condition that $H_{exp,like}(dx|\theta)$ have unit total mass on its support.

It is a classic result that the following distribution for $\theta \in \mathbb{R}^D$ constitutes a conjugate prior:

$$\begin{aligned} F_{exp,prior}(d\theta) &= f_{exp,prior}(\theta)\ d\theta \\ &= \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right] - B(\xi, \lambda)\right\}\ d\theta. \end{aligned} \tag{6.16}$$

$F_{exp,prior}$ is another exponential family distribution, now with natural parameter $(\xi', \lambda)'$, sufficient statistic $(\eta(\theta)', -A(\theta))'$, and log partition function $B(\xi, \lambda)$. Note that the logarithms of the densities in both Eq. (6.15) and Eq. (6.16) are linear in $\eta(\theta)$ and $-A(\theta)$. So, by Bayes Theorem, the posterior $F_{exp,post}$ also has these quantities as sufficient statistics in θ, and we can see $F_{exp,post}$ must have the following form.

$$\begin{aligned} &F_{exp,post}(d\theta|x) \\ &= f_{exp,post}(\theta|x)\; d\theta \\ &= \exp\left\{\langle \xi + \phi(x), \eta(\theta)\rangle + (\lambda+1)\,[-A(\theta)] - B(\xi + \phi(x), \lambda + 1)\right\}\; d\theta. \end{aligned} \tag{6.17}$$

Thus we see that $F_{exp,post}$ belongs to the same exponential family as $F_{exp,prior}$ in Eq. (6.16), and hence $F_{exp,prior}$ is a conjugate prior for $H_{exp,like}$ in Eq. (6.15).

Exponential families for completely random measures

In the finite-dimensional case, we saw that for any exponential family likelihood, as in Eq. (6.15), we can always construct a conjugate exponential family prior, given by Eq. (6.16).

In order to prove a similar result for CRMs, we start by defining a notion of exponential families for CRMs.

Definition 6.4.1. We say that a CRM Θ is an *exponential CRM* if it has the following two parts. First, let Θ have K_{fix} fixed-location atoms, where K_{fix} may be finite or infinite. The kth fixed-location atom is located at any $\psi_{fix,k}$, unique from the other fixed locations, and has random weight $\theta_{fix,k}$, whose distribution has density $f_{fix,k}$:

$$f_{fix,k}(\theta) = \kappa(\theta) \exp\left\{\langle \eta(\zeta_k), \phi(\theta)\rangle - A(\zeta_k)\right\},$$

for some base density κ, natural parameter function η, sufficient statistic ϕ, and log partition function A shared across atoms. Here, ζ_k is an atom-specific parameter.

Second, let Θ have an ordinary component with rate measure $\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi)$ for some proper distribution G and weight rate measure ν of the form

$$\nu(d\theta) = \gamma \exp\left\{\langle \eta(\zeta), \phi(\theta)\rangle\right\}.$$

In particular, η and ϕ are shared with the fixed-location atoms, and fixed hyperparameters γ and ζ are unique to the ordinary component.

Automatic conjugacy for completely random measures

With Definition 6.4.1 in hand, we can specify an automatic Bayesian nonparametric conjugate prior for an exponential CRM likelihood.

Theorem 6.4.2 (Automatic conjugacy)**.** *Let $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, in accordance with Assumption A1. Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. In particular, the distribution of the weight x_k at ψ_k of X has the following density conditional on the weight θ_k at ψ_k of Θ:*

$$h(x|\theta_k) = \kappa(x) \exp \left\{ \langle \eta(\theta_k), \phi(x) \rangle - A(\theta_k) \right\}.$$

Then a conjugate prior for Θ is the following exponential CRM distribution. First, let Θ have $K_{prior,fix}$ fixed-location atoms, in accordance with Assumption A0. The kth such atom has random weight $\theta_{fix,k}$ with proper density

$$f_{prior,fix,k}(\theta) = \exp \left\{ \langle \xi_{fix,k}, \eta(\theta) \rangle + \lambda_{fix,k} \left[-A(\theta) \right] - B(\xi_{fix,k}, \lambda_{fix,k}) \right\},$$

where $(\eta', -A)'$ here is the sufficient statistic and B is the log partition function. $\xi_{fix,k}$ and $\lambda_{fix,k}$ are fixed hyperparameters for this atom weight.

Second, let Θ have ordinary component characterized by any proper distribution G and weight rate measure

$$\nu(d\theta) = \gamma \exp \left\{ \langle \xi, \eta(\theta) \rangle + \lambda \left[-A(\theta) \right] \right\},$$

where γ, ξ, and λ are fixed hyperparameters of the weight rate measure chosen to satisfy Assumptions A1 and A2.

Proof. To prove the conjugacy of the prior for Θ with the likelihood for X, we calculate the posterior distribution of $\Theta|X$ using Theorem 6.3.1. Let Θ_{post} be a CRM with the distribution of $\Theta|X$. Then, by Theorem 6.3.1, Θ_{post} has the following three parts.

First, at any fixed location $\psi_{fix,k}$ in the prior, let $x_{fix,k}$ be the value of X at that location. Then Θ_{post} has a fixed-location atom at $\psi_{fix,k}$, and its weight $\theta_{post,fix,k}$ has distribution

$$\begin{aligned}
&F_{post,fix,k}(d\theta) \\
&\propto f_{prior,fix,k}(\theta)\ d\theta \cdot h(x_{fix,k}|\theta) \\
&\propto \exp \left\{ \langle \xi_{fix,k}, \eta(\theta) \rangle + \lambda_{fix,k} \left[-A(\theta) \right] \right\}\ d\theta \cdot \exp \left\{ \langle \eta(\theta), \phi(x_{fix,k}) \rangle - A(\theta) \right\}\ d\theta \\
&= \exp \left\{ \langle \xi_{fix,k} + \phi(x_{fix,k}), \eta(\theta) \rangle + (\lambda_{fix,k} + 1) \left[-A(\theta) \right] \right\}\ d\theta.
\end{aligned}$$

It follows, from putting in the normalizing constant, that the distribution of $\theta_{post,fix,k}$ has density

$$\begin{aligned}
f_{post,fix,k}(\theta) = \exp \big\{ &\langle \xi_{fix,k} + \phi(x_{fix,k}), \eta(\theta) \rangle + (\lambda_{fix,k} + 1) \left[-A(\theta) \right] \\
&- B(\xi_{fix,k} + \phi(x_{fix,k}), \lambda_{fix,k} + 1) \big\}.
\end{aligned}$$

Second, for any atom $x_{new,k} \delta_{\psi_{new,k}}$ in X that is not at a fixed location in the prior, Θ_{post} has a fixed atom at $\psi_{new,k}$ whose weight $\theta_{post,new,k}$ has distribution

$$\begin{aligned}
F_{post,new,k}(\theta) &\propto \nu(d\theta) \cdot h(x_{new,k}|\theta) \\
&\propto \exp \left\{ \langle \xi, \eta(\theta) \rangle + \lambda \left[-A(\theta) \right] \right\} \cdot \exp \left\{ \langle \eta(\theta), \phi(x_{new,k}) \rangle - A(\theta) \right\}\ d\theta
\end{aligned}$$

$$= \exp\left\{\langle \xi + \phi(x_{new,k}), \eta(\theta)\rangle + (\lambda + 1)\left[-A(\theta)\right]\right\}\ d\theta$$

and hence density

$$\begin{aligned} f_{post,new,k}(\theta) = \exp\{&\langle \xi + \phi(x_{new,k}), \eta(\theta)\rangle + (\lambda + 1)\left[-A(\theta)\right] \\ &- B(\xi + \phi(x_{new,k}), \lambda + 1)\}. \end{aligned}$$

Third, the ordinary component of Θ_{post} has weight rate measure

$$\begin{aligned} &\nu(d\theta) \cdot h(0|\theta) \\ &= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\} \cdot \kappa(0) \exp\left\{\langle \eta(\theta), \phi(0)\rangle - A(\theta)\right\} \\ &= \gamma\kappa(0) \cdot \exp\left\{\langle \xi + \phi(0), \eta(\theta)\rangle + (\lambda + 1)\left[-A(\theta)\right]\right\}. \end{aligned}$$

Thus, the posterior rate measure is in the same exponential CRM form as the prior rate measure with updated hyperparameters:

$$\begin{aligned} \gamma_{post} &= \gamma\kappa(0) \\ \xi_{post} &= \xi + \phi(0) \\ \lambda_{post} &= \lambda + 1. \end{aligned}$$

Since we see that the posterior fixed-location atoms are likewise in the same exponential CRM form as the prior, we have shown that conjugacy holds, as desired. □

We next use Theorem 6.4.2 to give proofs of conjugacy in cases where conjugacy has not previously been established in the Bayesian nonparametrics literature.

Example 6.4.3. Let X be generated according to a *Poisson likelihood process*[3] conditional on Θ. That is, $X = \sum_{k=1}^{\infty} x_k \delta_{\psi_k}$ conditional on $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$ has an exponential CRM distribution with only a fixed-location component. The weight x_k at location ψ_k has support on $\mathbb{Z}_*$ and has a Poisson density with parameter $\theta_k \in \mathbb{R}_+$:

$$\begin{aligned} h(x|\theta_k) &= \frac{1}{x!}\theta_k^x e^{-\theta_k} \\ &= \frac{1}{x!}\exp\left\{x \log(\theta_k) - \theta_k\right\}. \end{aligned} \tag{6.18}$$

The final line is rewritten to emphasize the exponential family form of this density, with

$$\begin{aligned} \kappa(x) &= \frac{1}{x!} \\ \phi(x) &= x \\ \eta(\theta) &= \log(\theta) \end{aligned}$$

[3] We use the term "Poisson likelihood process" to distinguish this specific Bayesian nonparametric likelihood from the Poisson point process.

$$A(\theta) = \theta.$$

By Theorem 6.4.2, this Poisson likelihood process has a Bayesian nonparametric conjugate prior for Θ with two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms, where $K_{prior,fix} < \infty$ by Assumption A0. The kth such atom has random weight $\theta_{fix,k}$ with density

$$\begin{aligned} f_{prior,fix,k}(\theta) &= \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \theta^{\xi_{fix,k}} e^{-\lambda_{fix,k}\theta} \exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \mathrm{Gamma}(\theta \,|\xi_{fix,k} + 1, \lambda_{fix,k}), \end{aligned} \tag{6.19}$$

where $\mathrm{Gamma}(\theta|a, b)$ denotes the gamma density with shape parameter $a > 0$ and rate parameter $b > 0$. So we must have fixed hyperparameters $\xi_{fix,k} > -1$ and $\lambda_{fix,k} > 0$. Further,

$$\exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} = \lambda_{fix,k}^{\xi_{fix,k}+1} / \Gamma(\xi_{fix,k} + 1)$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\begin{aligned} \nu(d\theta) &= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\}\ d\theta \\ &= \gamma \theta^{\xi} e^{-\lambda\theta}\ d\theta. \end{aligned} \tag{6.20}$$

Note that Theorem 6.4.2 guarantees that the weight rate measure will have the same distributional kernel in θ as the fixed-location atoms.

Finally, we need to choose the allowable hyperparameter ranges for γ, ξ, and λ. $\gamma > 0$ to ensure ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper gamma distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda \leq 0$. By Assumption A2, we must have

$$\begin{aligned} &\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot h(x|\theta) \\ &= \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot [1 - h(0|\theta)] \\ &= \int_{\theta \in \mathbb{R}_+} \gamma \theta^{\xi} e^{-\lambda\theta}\ d\theta \cdot \left[1 - e^{-\theta}\right] \\ &< \infty. \end{aligned}$$

To ensure the integral over $[1, \infty)$ is finite, we must have $\lambda > 0$. To ensure the integral over $(0, 1)$ is finite, we note that $1 - e^{-\theta}$ is asymptotically equivalent to θ as $\theta \to 0$. So we require

$$\int_{\theta \in (0,1)} \gamma \theta^{\xi+1} e^{-\lambda\theta}\ d\theta < \infty,$$

which is satisfied if and only if $\xi + 2 > 0$.

Finally, then the hyperparameter restrictions can be summarized as:

$$\begin{aligned}
\gamma &> 0 \\
\xi &\in (-2, -1] \\
\lambda &> 0 \\
\xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} &> 0 \quad \text{for all } k \in [K_{prior,fix}].
\end{aligned}$$

The ordinary component of the conjugate prior for Θ discovered in this example is typically called a *gamma process*. Here, we have for the first time specified the distribution of the fixed-location atoms of the gamma process and, also for the first time, proved that the gamma process is conjugate to the Poisson likelihood process. We highlight this result as a corollary to Theorem 6.4.2.

Corollary 6.4.4. *Let the Poisson likelihood process be a CRM with fixed-location atom weight distributions as in Eq. (6.18). Let the gamma process be a CRM with fixed-location atom weight distributions as in Eq. (6.19) and ordinary component weight measure as in Eq. (6.20). Then the gamma process is a conjugate Bayesian nonparametric prior for the Poisson likelihood process.*

■

Example 6.4.5. Next, let X be generated according to a new process we call an *odds Bernoulli process*. We have previously seen a typical Bernoulli process likelihood in Example 6.2.1. In the odds Bernoulli process, we say that X, conditional on Θ, has an exponential CRM distribution. In this case, the weight of the kth atom, x_k, conditional on θ_k has support on $\{0, 1\}$ and has a Bernoulli density with odds parameter $\theta_k \in \mathbb{R}_+$:

$$\begin{aligned}
h(x|\theta_k) &= \theta_k^x (1 + \theta_k)^{-1} \\
&= \exp\left\{ x \log(\theta_k) - \log(1 + \theta_k) \right\}.
\end{aligned} \tag{6.21}$$

That is, if ρ is the probability of a successful Bernoulli draw, then $\theta = \rho/(1 - \rho)$ represents the odds ratio of the probability of success over the probability of failure.

The final line of Eq. (6.21) is written to emphasize the exponential family form of this density, with

$$\begin{aligned}
\kappa(x) &= 1 \\
\phi(x) &= x \\
\eta(\theta) &= \log(\theta) \\
A(\theta) &= \log(1 + \theta).
\end{aligned}$$

By Theorem 6.4.2, the likelihood for X has a Bayesian nonparametric conjugate prior for Θ. This conjugate prior has two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms. The kth such atom has random weight $\theta_{fix,k}$ with density

$$\begin{aligned} f_{prior,fix,k}(\theta) &= \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \theta^{\xi_{fix,k}}(1+\theta)^{-\lambda_{fix,k}} \exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} \\ &= \operatorname{BetaPrime}\left(\theta\left|\xi_{fix,k}+1, \lambda_{fix,k} - \xi_{fix,k} - 1\right.\right), \end{aligned} \tag{6.22}$$

where BetaPrime$(\theta|a, b)$ denotes the beta prime density with shape parameters $a > 0$ and $b > 0$. Further,

$$\exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} = \frac{\Gamma(\lambda_{fix,k})}{\Gamma(\xi_{fix,k}+1)\Gamma(\lambda_{fix,k} - \xi_{fix,k} - 1)}$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\begin{aligned} \nu(d\theta) &= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\}\, d\theta \\ &= \gamma\theta^{\xi}(1+\theta)^{-\lambda}\, d\theta. \end{aligned} \tag{6.23}$$

We need to choose the allowable hyperparameter ranges for γ, ξ, and λ. $\gamma > 0$ to ensure ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper beta prime distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda - \xi - 1 \leq 0$. By Assumption A2, we must have

$$\begin{aligned} &\sum_{x=1}^{\infty} \int_{\theta\in\mathbb{R}_+} \nu(d\theta)\cdot h(x|\theta) \\ &= \int_{\theta\in\mathbb{R}_+} \nu(d\theta)\cdot h(1|\theta) \\ &\text{since the support of } x \text{ is } \{0,1\} \\ &= \int_{\theta\in\mathbb{R}_+} \gamma\theta^{\xi}(1+\theta)^{-\lambda}\, d\theta \cdot \theta^{1}(1+\theta)^{-1} \\ &= \gamma \int_{\theta\in\mathbb{R}_+} \theta^{\xi+1}(1+\theta)^{-\lambda-1}\, d\theta \\ &< \infty. \end{aligned}$$

Since the integrand is the kernel of a beta prime distribution, we simply require that this distribution be proper; i.e., $\xi + 2 > 0$ and $\lambda - \xi - 1 > 0$.

The hyperparameter restrictions can be summarized as:

$$\gamma > 0$$

$$\begin{aligned}
&\xi \in (-2, -1] \\
&\lambda > \xi + 1 \\
\xi_{fix,k} > -1 \text{ and } &\lambda_{fix,k} > \xi_{fix,k} + 1 \quad \text{for all } k \in [K_{prior,fix}].
\end{aligned}$$

We call the distribution for Θ described in this example the *beta prime process.* Its ordinary component has previously been defined by Broderick, Mackey, et al. (2014). But this result represents the first time the beta prime process is described in full, including parameter restrictions and fixed-location atoms, as well as the first proof of its conjugacy with the odds Bernoulli process. We highlight the latter result as a corollary to Theorem 6.4.2 below.

Corollary 6.4.6. *Let the odds Bernoulli process be a CRM with fixed-location atom weight distributions as in Eq. (6.21). Let the beta process be a CRM with fixed-location atom weight distributions as in Eq. (6.22) and ordinary component weight measure as in Eq. (6.23). Then the beta process is a conjugate Bayesian nonparametric prior for the odds Bernoulli process.*

■

6.5 Size-biased representations

We have shown in Section 6.4 that our exponential CRM (Definition 6.4.1) is useful in that we can find an automatic Bayesian nonparametric conjugate prior given an exponential CRM likelihood. We will see in this section and the next that exponential CRMs allow us to build representations that allow tractable inference despite the infinite-dimensional nature of the models we are using.

The best-known size-biased representation of a random measure in Bayesian nonparametrics is the *stick-breaking* representation of the Dirichlet process Θ_{DP} (Sethuraman, 1994):

$$\begin{aligned}
\Theta_{DP} &= \sum_{k=1}^{\infty} \theta_{DP,k} \delta_{\psi_k} \\
\theta_{DP,k} &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \quad \text{for } k \in \mathbb{Z}_* \\
\beta_k &\overset{iid}{\sim} \text{Beta}(1, c) \\
\psi_k &\overset{iid}{\sim} G,
\end{aligned} \tag{6.24}$$

where c is a fixed hyperparameter satisfying $c > 0$.

The name stick-breaking originates from thinking of the unit interval as a stick of length one. At each round k, only some of the stick remains; β_k describes the proportion of the remaining stick that is broken off in round k, and $\theta_{DP,k}$ describes the total amount of remaining stick that is broken off in round k. By construction, not only is each $\theta_{DP,k} \in (0, 1)$ but in fact the $\theta_{DP,k}$ add to one (the total stick length) and thus describe a distribution.

Eq. (6.24) is called a *size-biased* representation for the following reason. Since the weights $\{\theta_{DP,k}\}_{k=1}^{\infty}$ describe a distribution, we can make draws from this distribution; each such draw is sometimes thought of as a multinomial draw with a single trial. In that vein, typically we imagine that our data points $X_{mult,n}$ are described as iid draws conditioned on Θ_{DP}, where $X_{mult,n}$ is a random measure with just a single atom:

$$\begin{aligned} X_{mult,n} &= \delta_{\psi_{mult,n}} \\ \psi_{mult,n} &= \psi_k \text{ with probability } \theta_{DP,k}. \end{aligned} \tag{6.25}$$

Then the limiting proportion of data points $X_{mult,n}$ with atom at $\psi_{mult,1}$ (the first atom location chosen) is $\theta_{DP,1}$. The limiting proportion of data points with atom at the next unique atom location chosen will have size $\theta_{DP,2}$, and so on (Broderick, Jordan, and Pitman, 2013).

The representation in Eq. (6.24) is so useful because there is a familiar, finite-dimensional distribution for each of the atom weights $\theta_{DP,k}$ of the random measure Θ_{DP}. This representation allows approximate inference via truncation (Ishwaran and James, 2001) or exact inference via slice sampling (Walker, 2007; Kalli, Griffin, and Walker, 2011).

Since the weights $\{\theta_{DP,k}\}_{k=1}^{\infty}$ are constrained to sum to one, the Dirichlet process is not a CRM.[4] However, size-biased representations have been explored in the past for particular CRM examples, notably the beta process (Paisley, Zaas, et al., 2010; Broderick, Jordan, and Pitman, 2012). And even though there is no interpretation of these representations in terms of a single stick representing all probability mass, they are sometimes referred to as stick-breaking representations as a nod to the popularity of Dirichlet process stick-breaking.

In the beta process case, such size-biased representations have already been shown to allow approximate inference via truncation (Doshi et al., 2009; Paisley, Carin, and Blei, 2011) or exact inference via slice sampling (Teh, Görür, and Ghahramani, 2007; Broderick, Mackey, et al., 2014). Here we provide general recipes for the creation of these representations and illustrate our recipes by discovering previously unknown size-biased representations.

We have seen that a general CRM Θ takes the form of an a.s. discrete random measure:

$$\sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}. \tag{6.26}$$

The fixed-location atoms are straightforward to simulate; there are finitely many by Assumption A0, their locations are fixed, and their weights are assumed to come from finite-dimensional distributions. The infinite-dimensionality of the Bayesian nonparametric CRM comes from the ordinary component (cf. Section 6.2 and Assumption A1). So far the only description we have of the ordinary component is its generation from the countable infinity of points in a Poisson point process. The next result constructively demonstrates that we can represent the distributions of the CRM weights $\{\theta_k\}_{k=1}^{\infty}$ in Eq. (6.26) as a sequence of finite-dimensional distributions, much as in the familiar Dirichlet process case.

[4] In fact, the Dirichlet process is a normalized gamma process (cf. Example 6.4.3) (Ferguson, 1973).

Theorem 6.5.1 (Size-biased representations)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$. The ordinary component of Θ has rate measure*

$$\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi),$$

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X_n be generated iid given Θ according to $X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X_n and Θ jointly satisfy Assumption A2 so that

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) < \infty.$$

Then we can write

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}} \\
\psi_{m,x,k} &\overset{iid}{\sim} G \textit{ iid across } m, x, j \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho \,\middle|\, \int_{\theta} \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta)\right) \textit{ across } m, x \\
\theta_{m,x,j} &\overset{indep}{\sim} F_{size,m,x}(d\theta) \propto \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \\
&\qquad \textit{iid across } j \textit{ and independently across } m, x.
\end{aligned} \tag{6.27}$$

Proof. By construction, Θ is an a.s. discrete random measure with a countable infinity of atoms. Without loss of generality, suppose that for every (non-zero) value of an atom weight θ, there is a non-zero probability of generating an atom with non-zero weight x in the likelihood. Now suppose we generate $X_1, X_2, \ldots$. Then, for every atom $\theta\delta_{\psi}$ of Θ, there exists some finite n with an atom at ψ. Therefore, we can enumerate all of the atoms of Θ by enumerating

- Each atom $\theta\delta_{\psi}$ such that there is an atom in X_1 at ψ.
- Each atom $\theta\delta_{\psi}$ such that there is an atom in X_2 at ψ but there is not an atom in X_1 at ψ.

  $\vdots$
- Each atom $\theta\delta_{\psi}$ such that there is an atom in X_m at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$

Moreover, on the mth round of this enumeration, we can further break down the enumeration by the value of the observation X_m at the atom location:

- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** 1 at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.
- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** 2 at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$
- Each atom $\theta\delta_\psi$ such that there is an atom in X_m **of weight** x at ψ but there is not an atom in any of $X_1, X_2, \ldots, X_{m-1}$ at ψ.

  $\vdots$

Recall that the values θ_k that form the weights of Θ are generated according to a Poisson point process with rate measure $\nu(d\theta)$. So, on the first round, the values of θ_k such that $x_{1,k} = x$ also holds are generated according to a thinned Poisson point process with rate measure

$$\nu(d\theta)h(x|\theta).$$

In particular, since the rate measure has finite total mass by Assumption A2, we can define

$$M_{1,x} := \int_\theta \nu(d\theta)h(x|\theta),$$

which will be finite. Then the number of atoms θ_k for which $x_{1,k} = x$ is

$$\rho_{1,x} \sim \text{Poisson}(\rho|M_{1,x}).$$

And each such θ_k has weight with distribution

$$F_{size,1,x}(d\theta) \propto \nu(d\theta)h(x|\theta).$$

Finally, note from Theorem 6.3.1 that the posterior $\Theta|X_1$ has weight rate measure

$$\nu_1(d\theta) := \nu(d\theta)h(0|\theta).$$

Now take any $m > 1$. Suppose, inductively, that the ordinary component of the posterior $\Theta|X_1, \ldots, X_{m-1}$ has weight rate measure

$$\nu_{m-1}(d\theta) := \nu(d\theta)h(0|\theta)^{m-1}.$$

The atoms in this ordinary component have been selected precisely because they have not appeared in any of $X_1, \ldots, X_{m-1}$. As for $m = 1$, we have that the atoms θ_k in this ordinary component with corresponding weight in X_m equal to x are formed by a thinned Poisson point process, with rate measure

$$\nu_{m-1}(d\theta)h(x|\theta) = \nu(d\theta)h(0|\theta)^{m-1}h(x|\theta).$$

Since the rate measure has finite total mass by Assumption A2, we can define

$$M_{m,x} := \int_\theta \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta),$$

which will be finite. Then the number of atoms θ_k for which $x_{1,k} = x$ is

$$\rho_{m,x} \sim \text{Poisson}(\rho | M_{m,x}).$$

And each such θ_k has weight

$$F_{size,m,x} \propto \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta).$$

Finally, note from Theorem 6.3.1 that the posterior $\Theta | X_{1:m}$, which can be thought of as generated by prior $\Theta | X_{1:(m-1)}$ and likelihood $X_m | \Theta$, has weight rate measure

$$\nu(d\theta) h(0|\theta)^{m-1} h(0|\theta) = \nu_m(d\theta),$$

confirming the inductive hypothesis.

Recall that every atom of Θ is found in exactly one of these rounds and that $x \in \mathbb{Z}_+$. Also recall that the atom locations may be generated independently and identically across atoms, and independently from all the weights, according to proper distribution G (Section 6.2). To summarize, we have then

$$\Theta = \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}},$$

where

$$\begin{aligned}
\psi_{m,x,k} &\overset{iid}{\sim} G \text{ iid across } m, x, j \\
M_{m,x} &= \int_\theta \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \text{ across } m, x \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}(\rho | M_{m,x}) \text{ across } m, x \\
F_{size,m,x}(d\theta) &\propto \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \text{ across } m, x \\
\theta_{m,x,j} &\overset{indep}{\sim} F_{size,m,x}(d\theta) \text{ iid across } j \text{ and independently across } m, x,
\end{aligned}$$

as was to be shown. □

The following corollary gives a more detailed recipe for the calculations in Theorem 6.5.1 when the prior is in a conjugate exponential CRM to the likelihood.

Corollary 6.5.2 (Exponential CRM size-biased representations). *Let Θ be an exponential CRM with no fixed-location atoms (thereby trivially satisfying Assumption A0) such that Assumption A1 holds.*

Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. Let the distribution of the weight $x_{n,k}$ at ψ_k have probability mass function

$$h(x|\theta_k) = \kappa(x) \exp\left\{\langle \eta(\theta_k), \phi(x) \rangle - A(\theta_k)\right\}.$$

Suppose that Θ and X jointly satisfy Assumption A2. And let Θ be conjugate to X as in Theorem 6.4.2. Then we can write

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}} \\
\psi_{m,x,j} &\overset{iid}{\sim} G \quad \textit{iid across } m,x,j \\
M_{m,x} &= \gamma \cdot \kappa(0)^{m-1} \kappa(x) \cdot \exp\left\{B(\xi + (m-1)\phi(0) + \phi(x), \lambda + m)\right\} \\
\rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho | M_{m,x}\right) \\
&\quad \textit{independently across } m,x \\
\theta_{m,x,j} &\overset{indep}{\sim} f_{size,m,x}(\theta) \; d\theta \\
&\quad = \exp\left\{\langle \xi + (m-1)\phi(0) + \phi(x), \eta(\theta) \rangle + (\lambda + m)[-A(\theta)] \right. \\
&\quad\quad \left. - B(\xi + (m-1)\phi(0) + \phi(x), \lambda + m)\right\} \\
&\quad \textit{iid across } j \textit{ and independently across } m,x.
\end{aligned} \tag{6.28}$$

Proof. The corollary follows from Theorem 6.5.1 by plugging in the particular forms for $\nu(d\theta)$ and $h(x|\theta)$.

In particular,

$$\begin{aligned}
M_{m,x} &= \int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(0|\theta)^{m-1} h(x|\theta) \\
&= \int_{\theta \in \mathbb{R}_+} \gamma \exp\left\{\langle \xi, \eta(\theta) \rangle + \lambda \left[-A(\theta)\right]\right\} \\
&\quad \cdot \left[\kappa(0) \exp\left\{\langle \eta(\theta), \phi(0) \rangle - A(\theta)\right\}\right]^{m-1} \\
&\quad \cdot \kappa(x) \exp\left\{\langle \eta(\theta), \phi(x) \rangle - A(\theta)\right\} \; d\theta \\
&= \gamma \kappa(0)^{m-1} \kappa(x) \exp\left\{B\left(\xi + (m-1)\phi(0) + \phi(x), \lambda + m\right)\right\},
\end{aligned}$$

□

Corollary 6.5.2 can be used to find the known size-biased representation of the beta process (Thibaux and Jordan, 2007); we demonstrate this derivation in detail in Example 6.B.1 in Appendix 6.B. Here we use Corollary 6.5.2 to discover a new size-biased representation of the gamma process.

Example 6.5.3. Let Θ be a gamma process, and let X_n be iid Poisson likelihood processes conditioned on Θ for each n as in Example 6.4.3. That is, we have

$$\nu(d\theta) = \gamma \theta^{\xi} e^{-\lambda\theta} \, d\theta.$$

And

$$h(x|\theta_k) = \frac{1}{x!} \theta_k^x e^{-\theta_k}$$

with

$$\begin{aligned} \gamma &> 0 \\ \xi &\in (-2, -1] \\ \lambda &> 0 \\ \xi_{fix,k} &> -1 \text{ and } \lambda_{fix,k} > 0 \quad \text{for all } k \in [K_{prior,fix}] \end{aligned}$$

by Example 6.4.3.

We can pick out the following components of h:

$$\begin{aligned} \kappa(x) &= \frac{1}{x!} \\ \phi(x) &= x \\ \eta(\theta) &= \log(\theta) \\ A(\theta) &= \theta. \end{aligned}$$

Thus, by Corollary 6.5.2, we have

$$\begin{aligned} f_{size,m,x}(\theta) &\propto \theta^{\xi+x} e^{-(\lambda+m)\theta} \\ &\propto \text{Gamma}\left(\theta \,|\xi + x + 1, \lambda + m\,\right). \end{aligned}$$

We summarize the representation that follows from Corollary 6.5.2 in the following result.

Corollary 6.5.4. *Let the gamma process be a CRM Θ with fixed-location atom weight distributions as in Eq. (6.19) and ordinary component weight measure as in Eq. (6.20). Then we may write*

$$\begin{aligned} \Theta &= \sum_{m=1}^{\infty} \sum_{x=1}^{\infty} \sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j} \delta_{\psi_{m,x,j}} \\ \psi_{m,x,j} &\overset{iid}{\sim} G \quad \textit{iid across } m, x, j \\ M_{m,x} &= \gamma \cdot \frac{1}{x!} \cdot \Gamma(\xi + x + 1) \cdot (\lambda + m)^{-(\xi+x+1)} \textit{ across } m, x \\ \rho_{m,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho | M_{m,x}\right) \textit{ across } m, x \\ \theta_{m,x,j} &\overset{indep}{\sim} \text{Gamma}\left(\theta \,|\xi + x + 1, \lambda + m\,\right) \\ &\quad \textit{iid across } j \textit{ and independently across } m, x. \end{aligned}$$

■

6.6 Marginal processes

In Section 6.5, although we conceptually made use of the observations $\{X_1, X_2, \ldots\}$, we focused on a representation of the prior Θ: cf. Eqs. (6.27) and (6.28). In this section, we provide a representation of the marginal of $X_{1:N}$, with Θ integrated out.

The canonical example of a marginal process again comes from the Dirichlet process (DP). In this case, the full model consists of the DP-distributed prior on Θ_{DP} (as in Eq. (6.24)) together with the likelihood for $X_{mult,n}$ conditional on Θ_{DP} (iid across n) described by Eq. (6.25). Then the marginal distribution of $X_{mult,1:N}$ is described by the *Chinese restaurant process*. This marginal takes the following form.

For each $n = 1, 2, \ldots, N$,

1. Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_{mult,1}, \ldots, X_{mult,n-1}$. Then

$$X_{mult,n} | X_{mult,1}, \ldots, X_{mult,n-1}$$

has a single atom at ψ, where

$$\begin{aligned} \psi &= \begin{cases} \psi_k & \text{with probability} \propto \sum_{k=1}^{K_{n-1}} X_{mult,m}(\{\psi_k\}) \\ \psi_{new} & \text{with probability} \propto c \end{cases} \\ \psi_{new} &\sim G \end{aligned}$$

In the case of CRMs, the canonical example of a marginal process is the Indian buffet process (Griffiths and Ghahramani, 2006). Both the Chinese restaurant process and Indian buffet process have proven popular for inference since the underlying infinite-dimensional prior is integrated out in these processes and only the finite-dimensional marginal remains. Indeed, by Assumption A2, we know that this will generally be the case for our CRM Bayesian models. And thus we have the following general marginal representations for such models.

Theorem 6.6.1 (Marginal representations)**.** *Let Θ be a completely random measure that satisfies Assumptions A0 and A1; that is, Θ is a CRM with K_{fix} fixed atoms such that $K_{fix} < \infty$ and such that the kth atom can be written $\theta_{fix,k}\delta_{\psi_{fix,k}}$. The ordinary component of Θ has rate measure*

$$\mu(d\theta \times d\psi) = \nu(d\theta) \cdot G(d\psi),$$

where G is a proper distribution and $\nu(\mathbb{R}_+) = \infty$. Write $\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$, and let X_n be generated iid given Θ according to $X_n = \sum_{k=1}^{\infty} x_{n,k}\delta_{\psi_k}$ with $x_{n,k} \overset{indep}{\sim} h(x|\theta_k)$ for proper, discrete probability mass function h. And suppose X_n and Θ jointly satisfy Assumption A2 so that

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) < \infty.$$

Then the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1, 2, \ldots, N$,

1. *Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_1, \ldots, X_{n-1}$. Let $x_{m,k} := X_m(\{\psi_k\})$. Let $x_{n,k}$ denote the weight of $X_n|X_1, \ldots, X_{n-1}$ at ψ_k. Then $x_{n,k}$ has distribution described by the following probability mass function:*

$$h_{cond}\left(x_{n,k} = x \left| x_{1:(n-1),k}\right.\right) = \frac{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)}{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)}.$$

2. *For each $x = 1, 2, \ldots$*
   - *X_n has $\rho_{n,x}$ new atoms. That is, X_n has atoms at locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, where*

   $$\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \quad a.s.$$

   Moreover,

   $$\begin{aligned} \rho_{n,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho \left| \int_\theta \nu(d\theta) h(0|\theta)^{n-1} h(x|\theta)\right.\right) \text{ across } n, x \\ \psi_{n,x,j} &\overset{iid}{\sim} G(d\psi) \text{ across } n, x, j. \end{aligned}$$

Proof. We saw in the proof of Theorem 6.5.1 that the marginal for X_1 can be expressed as follows. For each $x \in \mathbb{Z}_+$, there are $\rho_{1,x}$ atoms of X_1 with weight x, where

$$\rho_{1,x} \overset{indep}{\sim} \text{Poisson}\left(\int_\theta \nu(d\theta) h(x|\theta)\right) \text{ across } x.$$

These atoms have locations $\{\psi_{1,x,j}\}_{j=1}^{\rho_{1,x}}$, where

$$\psi_{1,x,j} \overset{iid}{\sim} G(d\psi) \text{ across } x, j.$$

For the upcoming induction, let $K_1 := \sum_{x=1}^\infty \rho_{1,x}$. And let $\{\psi_k\}_{k=1}^{K_1}$ be the (a.s. disjoint by assumption) union of the sets $\{\psi_{1,x,j}\}_{j=1}^{\rho_{1,x}}$ across x. Note that K_1 is finite by Assumption A2.

We will also find it useful in the upcoming induction to let $\Theta_{post,1}$ have the distribution of $\Theta|X_1$. Let $\theta_{post,1,x,j} = \Theta_{post,1}(\{\psi_{1,x,j}\})$. By Theorem 6.3.1 or the proof of Theorem 6.5.1, we have that

$$\begin{aligned} \theta_{post,1,x,j} &\overset{indep}{\sim} F_{post,1,x,j}(d\theta) \propto \nu(d\theta) h(x|\theta) \\ &\text{independently across } x \text{ and iid across } j. \end{aligned}$$

Now take any $n > 1$. Inductively, we assume $\{\psi_{n-1,k}\}_{k=1}^{K_{n-1}}$ is the union of all the atom locations of $X_1, \ldots, X_{n-1}$. Further assume K_{n-1} is finite. Let $\Theta_{post,n-1}$ have the distribution of $\Theta | X_1, \ldots, X_{n-1}$. Let $\theta_{n-1,k}$ be the weight of $\Theta_{post,n-1}$ at $\psi_{n-1,k}$. And, for any $m \in [n-1]$, let $x_{m,k}$ be the weight of X_m at $\psi_{n-1,k}$. We inductively assume that

$$\begin{aligned} \theta_{n-1,k} \overset{indep}{\sim} F_{n-1,k}(d\theta) \propto \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta) \\ \text{independently across } k. \end{aligned} \tag{6.29}$$

Now let $\psi_{n,k}$ equal $\psi_{n-1,k}$ for $k \in [K_{n-1}]$. Let $x_{n,k}$ denote the weight of X_n at $\psi_{n,k}$ for $k \in [K_{n-1}]$. Conditional on the atom weight of Θ at $\psi_{n,k}$, the atom weights of $X_1, \ldots, X_{n-1}, X_n$ are independent. Since the atom weights of Θ are independent as well, we have that $x_{n,k} | X_1, \ldots, X_{n-1}$ has the same distribution as $x_{n,k} | x_{1,k}, \ldots, x_{n-1,k}$. We can write the probability mass function of this distribution as follows.

$$\begin{aligned} &h_{cond}\left(x_{n,k} = x \,|x_{1,k}, \ldots, x_{n-1,k}\right) \\ &= \int_{\theta \in \mathbb{R}_+} F_{n-1,k}(d\theta) h(x|\theta) \\ &= \frac{\int_{\theta \in \mathbb{R}_+} \left[\nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)\right] \cdot h(x|\theta)}{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)}, \end{aligned}$$

where the last line follows from Eq. (6.29).

We next show the inductive hypothesis in Eq. (6.29) holds for n and $k \in [K_{n-1}]$. Let $x_{n,k}$ denote the weight of X_n at $\psi_{n,k}$ for $k \in [K_{n-1}]$. Let $F_{n,k}(d\theta)$ denote the distribution of $x_{n,k}$ and note that

$$\begin{aligned} F_{n,k}(d\theta) &\propto F_{n-1,k}(d\theta) \cdot h(x_{n,k}|\theta) \\ &= \nu(d\theta) \prod_{m=1}^{n} h(x_{m,k}|\theta), \end{aligned}$$

which agrees with Eq. (6.29) for n when we assume the result for $n-1$.

The previous development covers atoms that are present in at least one of $X_1, \ldots, X_{n-1}$. Next we consider new atoms in X_n; that is, we consider atoms in X_n for which there are no atoms at the same location in any of $X_1, \ldots, X_{n-1}$.

We saw in the proof of Theorem 6.5.1 that, for each $x \in \mathbb{Z}_+$, there are $\rho_{n,x}$ new atoms of X_n with weight x such that

$$\rho_{n,x} \overset{indep}{\sim} \text{Poisson}\left(\rho \,\middle|\, \int_\theta \nu(d\theta) h(0|\theta)^{n-1} h(x|\theta)\right) \text{ across } x.$$

These new atoms have locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$ with

$$\psi_{n,x,j} \overset{iid}{\sim} G(d\psi) \text{ across } x, j.$$

By Assumption A2, $\sum_{x=1}^{\infty} \rho_{n,x} < \infty$. So

$$K_n := K_{n-1} + \sum_{x=1}^{\infty} \rho_{n,x}$$

remains finite. Let $\psi_{n,k}$ for $k \in \{K_{n-1}+1, \ldots, K_n\}$ index these new locations. Let $\theta_{n,k}$ be the weight of $\Theta_{post,n}$ at $\psi_{n,k}$ for $k \in \{K_{n-1}+1, \ldots, K_n\}$. And let $x_{n,k}$ be the value of X at $\psi_{n,k}$.

We check that the inductive hypothesis holds. By repeated application of Theorem 6.3.1, the ordinary component of $\Theta | X_1, \ldots, X_{n-1}$ has rate measure

$$\nu(d\theta) h(0|\theta)^{n-1}.$$

So, again by Theorem 6.3.1, we have that

$$\theta_{n,k} \overset{indep}{\sim} F_{n.k}(d\theta) \propto \nu(d\theta) h(0|\theta)^{n-1} h(x_{n,k}|\theta).$$

Since X_m has value 0 at $\psi_{n,k}$ for $m \in \{1, \ldots, n-1\}$ by construction, we have that the inductive hypothesis holds. □

As in the case of size-biased representations (Section 6.5 and Corollary 6.5.2), we can find a more detailed recipe when the prior is in a conjugate exponential CRM to the likelihood.

Corollary 6.6.2 (Exponential CRM marginal representations). *Let Θ be an exponential CRM with no fixed-location atoms (thereby trivially satisfying Assumption A0) such that Assumption A1 holds.*

Let X be generated conditional on Θ according to an exponential CRM with fixed-location atoms at $\{\psi_k\}_{k=1}^{\infty}$ and no ordinary component. Let the distribution of the weight $x_{n,k}$ at ψ_k have probability mass function

$$h(x|\theta_k) = \kappa(x) \exp\left\{ \langle \eta(\theta_k), \phi(x) \rangle - A(\theta_k) \right\}.$$

Suppose that Θ and X jointly satisfy Assumption A2. And let Θ be conjugate to X as in Theorem 6.4.2. Then the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1, 2, \ldots, N$,

1. *Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_1, \ldots, X_{n-1}$. Let $x_{m,k} := X_m(\{\psi_k\})$. Let $x_{n,k}$ denote the weight of $X_n | X_1, \ldots, X_{n-1}$ at ψ_k. Then $x_{n,k}$ has distribution described by the following probability mass function:*

$$\begin{aligned} & h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) \\ &= \kappa(x) \exp\left\{ -B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + x, \lambda + n) \right\}. \end{aligned}$$

2. For each $x = 1, 2, \ldots$

- *X_n has $\rho_{n,x}$ new atoms. That is, X_n has atoms at locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, where*

$$\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \qquad a.s.$$

Moreover,

$$\begin{aligned} M_{n,x} &:= \gamma \cdot \kappa(0)^{n-1}\kappa(x) \cdot \exp\left\{B(\xi + (n-1)\phi(0) + \phi(x), \lambda + n)\right\} \\ &\quad \textit{across } n, x \\ \rho_{n,x} &\overset{indep}{\sim} \text{Poisson}\left(\rho \,\middle|\, M_{n,x}\right) \textit{ across } n, x \\ \psi_{n,x,j} &\overset{iid}{\sim} G(d\psi) \textit{ across } n, x, j. \end{aligned}$$

Proof. The corollary follows from Theorem 6.6.1 by plugging in the forms for $\nu(d\theta)$ and $h(x|\theta)$.

In particular,

$$\begin{aligned} &\int_{\theta \in \mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n} h(x_{m,k}|\theta) \\ &= \int_{\theta \in \mathbb{R}_+} \gamma \exp\left\{\langle \xi, \eta(\theta) \rangle + \lambda\left[-A(\theta)\right]\right\} \cdot \left[\prod_{m=1}^{n} \kappa(x_{m,k}) \exp\left\{\langle \eta(\theta), \phi(x_{m,k}) \rangle - A(\theta)\right\}\right] \\ &= \gamma \left[\prod_{m=1}^{n} \kappa(x_{m,k})\right] B\left(\xi + \sum_{m=1}^{n} \phi(x_{m,k}), \lambda + n\right). \end{aligned}$$

So

$$\begin{aligned} &h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) \\ &= \frac{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) h(x|\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)}{\int_{\theta \in \mathbb{R}_+} \nu(d\theta) \prod_{m=1}^{n-1} h(x_{m,k}|\theta)} \\ &= \kappa(x) \exp\left\{-B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + x, \lambda + n)\right\}. \end{aligned}$$

□

In Example 6.C.1 in Appendix 6.C we show that Corollary 6.6.2 can be used to recover the Indian buffet process marginal from a beta process prior together with a Bernoulli process likelihood. In the following example, we discover a new marginal for the Poisson likelihood process with gamma process prior.

Example 6.6.3. Let Θ be a gamma process, and let X_n be iid Poisson likelihood processes conditioned on Θ for each n as in Example 6.4.3. That is, we have

$$\nu(d\theta) = \gamma\theta^{\xi}e^{-\lambda\theta}\ d\theta.$$

And

$$h(x|\theta_k) = \frac{1}{x!}\theta_k^x e^{-\theta_k}$$

with

$$\begin{aligned}\gamma &> 0\\ \xi &\in (-2,-1]\\ \lambda &> 0\\ \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} &> 0 \quad \text{for all } k \in [K_{prior,fix}]\end{aligned}$$

by Example 6.4.3.

We can pick out the following components of h:

$$\begin{aligned}\kappa(x) &= \frac{1}{x!}\\ \phi(x) &= x\\ \eta(\theta) &= \log(\theta)\\ A(\theta) &= \theta.\end{aligned}$$

And we calculate

$$\begin{aligned}\exp\left\{B(\xi,\lambda)\right\} &= \int_{\theta\in\mathbb{R}_+} \exp\left\{\langle\xi,\eta(\theta)\rangle + \lambda[-A(\theta)]\right\}\ d\theta\\ &= \int_{\theta\in\mathbb{R}_+} \theta^{\xi}e^{-\lambda\theta}\\ &= \Gamma(\xi+1)\lambda^{-(\xi+1)}.\end{aligned}$$

So, for $k \in \mathbb{Z}_*$, we have

$$\begin{aligned}\mathbb{P}(x_n = x) &= \kappa(x)\exp\left\{-B(\xi + \sum_{m=1}^{n-1} x_m, \lambda + n - 1) + B(\xi + \sum_{m=1}^{n-1} x_m + x, \lambda + n)\right\}\\ &= \frac{1}{x!}\cdot\frac{(\lambda+n-1)^{\xi+\sum_{m=1}^{n-1} x_m + 1}}{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1)}\\ &\quad\cdot\frac{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + x + 1)}{(\lambda+n)^{\xi+\sum_{m=1}^{n-1} x_m + x + 1}}\\ &= \frac{\Gamma(\xi + \sum_{m=1}^{n-1} x_m + x + 1)}{\Gamma(x+1)\Gamma(\xi + \sum_{m=1}^{n-1} x_m + 1)}\end{aligned}$$

$$
\begin{aligned}
&\cdot \left(\frac{\lambda + n - 1}{\lambda + n}\right)^{\xi + \sum_{m=1}^{n} x_m + 1} \left(\frac{1}{\lambda + n}\right)^{x} \\
&= \mathrm{NegBin}\left(x \,\middle|\, \xi + \sum_{m=1}^{n-1} x_m + 1, (\lambda + n)^{-1}\right).
\end{aligned}
$$

And

$$
\begin{aligned}
M_{n,x} &:= \gamma \cdot \kappa(0)^{n-1}\kappa(x) \cdot \exp\left\{B(\xi + (n-1)\phi(0) + \phi(x), \lambda + n)\right\} \\
&= \gamma \cdot \frac{1}{x!} \cdot \Gamma(\xi + x + 1)(\lambda + n)^{-(\xi + x + 1)}.
\end{aligned}
$$

We summarize the marginal distribution representation of $X_{1:N}$ that follows from Corollary 6.6.2 in the following result.

Corollary 6.6.4. *Let Θ be a gamma process with fixed-location atom weight distributions as in Eq. (6.19) and ordinary component weight measure as in Eq. (6.20). Let X_n be drawn, iid across n, conditional on Θ according to a Poisson likelihood process with fixed-location atom weight distributions as in Eq. (6.18). Then $X_{1:N}$ has the same distribution as the following construction.*

For each $n = 1, 2, \ldots, N$,

1. *Let $\{\psi_k\}_{k=1}^{K_{n-1}}$ be the union of atom locations in $X_1, \ldots, X_{n-1}$. Let $x_{m,k} := X_m(\{\psi_k\})$. Let $x_{n,k}$ denote the weight of $X_n | X_1, \ldots, X_{n-1}$ at ψ_k. Then $x_{n,k}$ has distribution described by the following probability mass function:*

$$
\begin{aligned}
&h_{cond}\left(x_{n,k} = x \,\middle|\, x_{1:(n-1),k}\right) \\
&= \mathrm{NegBin}\left(x \,\middle|\, \xi + \sum_{m=1}^{n-1} x_{m,k} + 1, (\lambda + n)^{-1}\right).
\end{aligned}
$$

2. *For each $x = 1, 2, \ldots$*

   - *X_n has $\rho_{n,x}$ new atoms. That is, X_n has atoms at locations $\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}}$, where*

$$
\{\psi_{n,x,j}\}_{j=1}^{\rho_{n,x}} \cap \{\psi_k\}_{k=1}^{K_{n-1}} = \emptyset \qquad a.s.
$$

   Moreover,

$$
\begin{aligned}
M_{n,x} &:= \gamma \cdot \frac{1}{x!} \cdot \frac{\Gamma(\xi + x + 1)}{(\lambda + n)^{\xi + x + 1}} \\
&\quad \textit{across } n, x \\
\rho_{n,x} &\overset{indep}{\sim} \mathrm{Poisson}\left(\rho \,\middle|\, M_{n,x}\right) \textit{ independently across } n, x \\
\psi_{n,x,j} &\overset{iid}{\sim} G(d\psi) \textit{ independently across } n, x \textit{ and iid across } j.
\end{aligned}
$$

■

6.7 Discussion

In the preceding sections, we have shown how to calculate posteriors for general CRM-based priors and likelihoods for Bayesian nonparametric models. We have also shown how to represent Bayesian nonparametric priors as a sequence of finite draws, and full Bayesian nonparametric models via finite marginals. We have introduced a notion of exponential families for CRMs, which we call exponential CRMs, that has allowed us to specify automatic Bayesian nonparametric conjugate priors for exponential CRM likelihoods. And we have demonstrated that our exponential CRMs allow particularly straightforward recipes for size-biased and marginal representations of Bayesian nonparametric models. Along the way, we have proved that the gamma process is a conjugate prior for the Poisson likelihood process and the beta prime process is a conjugate prior for the odds Bernoulli process. We have discovered a size-biased representation of the gamma process and a marginal representation of the gamma process coupled with a Poisson likelihood process.

All of this work has relied heavily on the description of Bayesian nonparametric models in terms of completely random measures. As such, we have worked very particularly with pairings of real values—the CRM atom weights, which we have interpreted as trait frequencies or rates—together with trait descriptors—the CRM atom locations. However, all of our proofs broke into essentially two parts: the fixed-location atom part and the ordinary component part. The fixed-location atom development essentially translated into the usual finite version of Bayes Theorem and could easily be extended to full Bayesian models where the prior describes a random element that need not be real-valued. Moreover, the ordinary component development relied entirely on its generation as a Poisson point process over a product space. It seems reasonable to expect that our development might carry through when the first element in this tuple need not be real-valued. And thus we believe our results are suggestive of broader results over more general spaces.

6.A Further automatic conjugate priors

We use Theorem 6.4.2 to calculate automatic conjugate priors for further exponential CRMs.

Example 6.A.1. Let X be generated according to a Bernoulli process as in Example 6.2.1. That is, X has an exponential CRM distribution with $K_{like,fix}$ fixed-location atoms, where $K_{like,fix} < \infty$ in accordance with Assumption A0:

$$X = \sum_{k=1}^{K_{like,fix}} x_{like,k} \delta_{\psi_{like,k}}.$$

The weight of the kth atom, $x_{like,k}$, has support on $\{0, 1\}$ and has a Bernoulli density with parameter $\theta_k \in (0, 1]$:

$$h(x|\theta_k) = \theta_k^x (1 - \theta_k)^{1-x}$$

$$= \exp\left\{x \log(\theta_k/(1-\theta_k)) + \log(1-\theta_k)\right\}.$$

The final line is rewritten to emphasize the exponential family form of this density, with

$$\begin{aligned}\kappa(x) &= 1\\ \phi(x) &= x\\ \eta(\theta) &= \log\left(\frac{\theta}{1-\theta}\right)\\ A(\theta) &= -\log(1-\theta).\end{aligned}$$

Then, by Theorem 6.4.2, X has a Bayesian nonparametric conjugate prior for

$$\Theta := \sum_{k=1}^{K_{like,fix}} \theta_k \delta_{\psi_k}.$$

This conjugate prior has two parts.

First, Θ has a set of $K_{prior,fix}$ fixed-location atoms at some subset of the $K_{like,fix}$ fixed locations of X. The kth such atom has random weight $\theta_{fix,k}$ with density

$$\begin{aligned}f_{prior,fix,k}(\theta) &= \exp\left\{\langle \xi_{fix,k}, \eta(\theta)\rangle + \lambda_{fix,k}\left[-A(\theta)\right] - B(\xi_{fix,k}, \lambda_{fix,k})\right\}\\ &= \theta^{\xi_{fix,k}}(1-\theta)^{\lambda_{fix,k}-\xi_{fix,k}} \exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\}\\ &= \mathrm{Beta}\left(\theta \left| \xi_{fix,k}+1, \lambda_{fix,k}-\xi_{fix,k}+1\right.\right),\end{aligned}$$

where $\mathrm{Beta}(\theta|a,b)$ denotes the beta density with shape parameters $a > 0$ and $b > 0$. So we must have fixed hyperparameters $\xi_{fix,k} > -1$ and $\lambda_{fix,k} > \xi_{fix,k} - 1$. Further,

$$\exp\left\{-B(\xi_{fix,k}, \lambda_{fix,k})\right\} = \frac{\Gamma(\lambda_{fix,k}+2)}{\Gamma(\xi_{fix,k}+1)\Gamma(\lambda_{fix,k}-\xi_{fix,k}+1)}$$

to ensure normalization.

Second, Θ has an ordinary component characterized by any proper distribution G and weight rate measure

$$\begin{aligned}\nu(d\theta) &= \gamma \exp\left\{\langle \xi, \eta(\theta)\rangle + \lambda\left[-A(\theta)\right]\right\}\ d\theta\\ &= \gamma \theta^{\xi}(1-\theta)^{\lambda-\xi}\ d\theta.\end{aligned}$$

Finally, we need to choose the allowable hyperparameter ranges for γ, ξ, and λ. $\gamma > 0$ ensures ν is a measure. By Assumption A1, we must have $\nu(\mathbb{R}_+) = \infty$, so ν must represent an improper beta distribution. As such, we require either $\xi + 1 \leq 0$ or $\lambda - \xi \leq 0$. By Assumption A2, we must have

$$\sum_{x=1}^{\infty} \int_{\theta \in \mathbb{R}_+} \nu(d\theta) \cdot h(x|\theta)$$

$$= \int_{\theta \in (0,1]} \nu(d\theta) h(1|\theta)$$

since the support of x is $\{0,1\}$ and the support of θ is $(0,1]$

$$= \gamma \int_{\theta \in (0,1]} \theta^{\xi}(1-\theta)^{\lambda-\xi} \, d\theta \cdot \theta$$

$$< \infty$$

Since the integrand is the kernel of a beta distribution, the integral is finite if and only if $\xi + 2 > 0$ and $\lambda - \xi + 1 > 0$.

Finally, then the hyperparameter restrictions can be summarized as:

$$\begin{aligned} \gamma &> 0 \\ \xi &\in (-2,-1] \\ \lambda &> \xi - 1 \\ \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} &> \xi_{fix,k} - 1 \quad \text{for all } k \in [K_{prior,fix}] \end{aligned}$$

By setting $\alpha = \xi + 1$, $c = \lambda + 2$, $\rho_{fix,k} = \xi_{fix,k} + 1$, and $\sigma_{fix,k} = \lambda_{fix,k} - \xi_{fix,k} + 1$, we recover the hyperparameters of Eq. (6.11) in Example 6.2.1. Here, by contrast to Example 6.2.1, we found the conjugate prior and its hyperparameter settings given just the Bernoulli process likelihood. Henceforth, we use the parameterization of the beta process above. ■

6.B Further size-biased representations

Example 6.B.1. Let Θ be a beta process, and let X_n be iid Bernoulli processes conditioned on Θ for each n as in Example 6.A.1. That is, we have

$$\nu(d\theta) = \gamma \theta^{\xi}(1-\theta)^{\lambda-\xi} \, d\theta.$$

And

$$h(x|\theta_k) = \theta_k^x (1-\theta_k)^{1-x}$$

with

$$\begin{aligned} \gamma &> 0 \\ \xi &\in (-2,-1] \\ \lambda &> \xi - 1 \\ \xi_{fix,k} > -1 \text{ and } \lambda_{fix,k} &> \xi_{fix,k} - 1 \quad \text{for all } k \in [K_{prior,fix}] \end{aligned}$$

by Example 6.A.1.

We can pick out the following components of h:

$$\kappa(x) = 1$$

$$\begin{aligned}
\phi(x) &= x \\
\eta(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \\
A(\theta) &= -\log(1-\theta).
\end{aligned}$$

Thus, by Corollary 6.5.2,

$$\begin{aligned}
\Theta &= \sum_{m=1}^{\infty}\sum_{x=1}^{\infty}\sum_{j=1}^{\rho_{m,x}} \theta_{m,x,j}\delta_{\psi_{m,x,j}} \\
\psi_{m,x,j} &\overset{iid}{\sim} G \quad \text{iid across } m, x, j \\
\theta_{m,x,j} &\overset{indep}{\sim} f_{size,m,x}(\theta)\; d\theta \\
&\propto \theta^{\xi+x}(1-\theta)^{\lambda+m-\xi-x}\; d\theta \\
&\propto \mathrm{Beta}\left(\theta\,|\xi+x, \lambda-\xi+m-x\right)\; d\theta \\
&\text{iid across } j \text{ and independently across } m, x \\
M_{m,x} &:= \gamma \cdot \frac{\Gamma(\xi+x+1)\Gamma(\lambda-\xi+m-x+1)}{\Gamma(\lambda+m+2)} \\
\rho_{m,x} &\overset{indep}{\sim} \mathrm{Poisson}\left(M_{m,x}\right) \\
&\text{across } m, x
\end{aligned}$$

Broderick, Jordan, and Pitman (2012) and Paisley, Blei, and Jordan (2012) have previously noted that this size-biased representation of the beta process arises from the Poisson point process. ■

6.C Further marginals

Example 6.C.1. Let Θ be a beta process, and let X_n be iid Bernoulli processes conditioned on Θ for each n as in Examples 6.A.1 and 6.B.1.

We calculate the main components of Corollary 6.6.2 for this pair of processes. In particular, we have

$$\begin{aligned}
\mathbb{P}(x_n = 1) &= \kappa(k)\exp\left\{-B(\xi + \sum_{m=1}^{n-1} x_m, \lambda+n-1) + B(\xi + \sum_{m=1}^{n-1} x_m + 1, \lambda+n)\right\} \\
&= \frac{\Gamma(\lambda+n-1+2)}{\Gamma(\xi+\sum_{m=1}^{n-1} x_m + 1)\Gamma(\lambda+n-1-\xi-\sum_{m=1}^{n-1} x_m+1)} \\
&\quad \cdot \frac{\Gamma(\xi+\sum_{m=1}^{n-1} x_m+1+1)\Gamma(\lambda+n-\xi-\sum_{m=1}^{n-1} x_m-1+1)}{\Gamma(\lambda+n+2)} \\
&= \frac{\xi+\sum_{m=1}^{n-1} x_m+1}{\lambda+n+1}
\end{aligned}$$

And

$$\begin{aligned} M_{n,1} &:= \gamma \cdot \kappa(0)^{n-1}\kappa(1) \cdot \exp\left\{B(\xi + (n-1)\phi(0) + \phi(1), \lambda + n)\right\} \\ &= \gamma \cdot \frac{\Gamma(\xi + 1 + 1)\Gamma(\lambda + n - \xi - 1 + 1)}{\Gamma(\lambda + n + 2)} \end{aligned}$$

Thus, the marginal distribution of $X_{1:N}$ is the same as that provided by the following construction.

For each $n = 1, 2, \ldots, N$,

1. At any location ψ for which there is some atom in $X_1, \ldots, X_{n-1}$, let x_m be the weight of X_m at ψ for $m \in [n-1]$. Then we have that $X_n | X_1, \ldots, X_{n-1}$ has weight x_n at ψ, where

$$\mathbb{P}(dx_n) = \text{Bern}\left(x_n \left| \frac{\xi + \sum_{m=1}^{n-1} x_m + 1}{\lambda + n + 1}\right.\right)$$

2. X_n has $\rho_{n,1}$ atoms at locations $\{\psi_{n,1,j}\}$ with $j \in [\rho_{n,1}]$ where there have not yet been atoms in any of $X_1, \ldots, X_{n-1}$. Moreover,

$$\begin{aligned} M_{n,1} &:= \gamma \cdot \frac{\Gamma(\xi + 1 + 1)\Gamma(\lambda + n - \xi - 1 + 1)}{\Gamma(\lambda + n + 2)} \\ &\quad \text{across } n \\ \rho_{n,1} &\overset{indep}{\sim} \text{Poisson}\left(M_{n,1}\right) \text{ across } n, x \\ \psi_{n,1,j} &\overset{iid}{\sim} G(d\psi) \text{ across } n, j \end{aligned}$$

Here, we have recovered the three-parameter extension of the Indian buffet process (Teh and Görür, 2009; Broderick, Jordan, and Pitman, 2013). ■

Part II

Scaling inference

Chapter 7

Streaming variational Bayes

We present SDA-Bayes, a framework for (S)treaming, (D)istributed, (A)synchronous computation of a Bayesian posterior. The framework makes streaming updates to the estimated posterior according to a user-specified approximation batch primitive. We demonstrate the usefulness of our framework, with variational Bayes (VB) as the primitive, by fitting the latent Dirichlet allocation model to two large-scale document collections. We demonstrate the advantages of our algorithm over stochastic variational inference (SVI) by comparing the two after a single pass through a known amount of data—a case where SVI may be applied—and in the streaming setting, where SVI does not apply.

7.1 Introduction

Large, streaming data sets are increasingly the norm in science and technology. Simple descriptive statistics can often be readily computed with a constant number of operations for each data point in the streaming setting, without the need to revisit past data or have advance knowledge of future data. But these time and memory restrictions are not generally available for the complex, hierarchical models that practitioners often have in mind when they collect large data sets. Significant progress on scalable learning procedures has been made in recent years (e.g., Niu et al., 2011; Kleiner et al., 2012). But the underlying models remain simple, and the inferential framework is generally non-Bayesian. The advantages of the Bayesian paradigm (e.g., hierarchical modeling, coherent treatment of uncertainty) currently seem out of reach in the Big Data setting.

An exception to this statement is provided by (Hoffman, Blei, and Bach, 2010; Hoffman, Blei, Paisley, et al., 2013; C. Wang, Paisley, and Blei, 2011), who have shown that a class of approximation methods known as *variational Bayes* (VB) (Wainwright and Jordan, 2008) can be usefully deployed for large-scale data sets. They have applied their approach, referred to as *stochastic variational inference* (SVI), to the domain of topic modeling of document collections, an area with a major need for scalable inference algorithms. VB traditionally uses the variational lower bound on the marginal likelihood as an objective function, and the idea of SVI is to apply a variant of stochastic gradient descent to this objective. Notably, this objective is based on the conceptual existence of a full data set involving D data points (i.e., documents in the topic model setting), for a fixed value of D. Although the stochastic gradient is computed for a single, small subset of data points (documents) at a time, the posterior being targeted is a posterior for D data points. This value of D must be specified in advance and is used by the algorithm at each step. Posteriors for D' data points, for $D' \neq D$, are not obtained as part of the analysis.

We view this lack of a link between the number of documents that have been processed thus far and the posterior that is being targeted as undesirable in many settings involving streaming data. In this chapter we aim at an approximate Bayesian inference algorithm that is scalable like SVI but is also truly a streaming procedure, in that it yields an approximate posterior for each processed collection of D' data points—and not just a pre-specified "final" number of data points D. To that end, we return to the classical perspective of Bayesian updating, where the recursive application of Bayes theorem provides a sequence of posteriors, not a sequence of approximations to a fixed posterior. To this classical recursive perspective we bring the VB framework; our updates need not be exact Bayesian updates but rather may be approximations such as VB. This approach is similar in spirit to assumed density filtering or expectation propagation (Minka, 2001b; Minka, 2001a; Opper, 1998), but each step of those methods involves a moment-matching step that can be computationally costly for models such as topic models. We are able to avoid the moment-matching step via the use of VB. We also note other related work in this general vein: MCMC approximations have been explored by (Canini, Shi, and Griffiths, 2009), and VB or VB-like approximations have also been explored by (Honkela and Valpola, 2003; Luts, Broderick, and Wand, 2012).

Although the empirical success of SVI is the main motivation for our work, we are also motivated by recent developments in computer architectures, which permit distributed and asynchronous computations in addition to streaming computations. As we will show, a streaming VB algorithm naturally lends itself to distributed and asynchronous implementations.

7.2 Streaming, distributed, asynchronous Bayesian updating

Streaming Bayesian updating. Consider data $x_1, x_2, \ldots$ generated iid according to a distribution $p(x \,|\, \Theta)$ given parameter(s) Θ. Assume that a prior $p(\Theta)$ has also been specified. Then Bayes theorem gives us the *posterior distribution* of Θ given a collection of S data points, $C_1 := (x_1, \ldots, x_S)$:

$$p(\Theta \mid C_1) = p(C_1)^{-1}\, p(C_1 \mid \Theta)\, p(\Theta),$$

where $p(C_1 \mid \Theta) = p(x_1, \ldots, x_S \mid \Theta) = \prod_{s=1}^{S} p(x_s \mid \Theta)$.

Suppose we have seen and processed $b-1$ collections, sometimes called *minibatches*, of data. Given the posterior $p(\Theta \mid C_1, \ldots, C_{b-1})$, we can calculate the posterior after the bth minibatch:

$$p(\Theta \mid C_1, \ldots, C_b) \propto p(C_b \mid \Theta)\, p(\Theta \mid C_1, \ldots, C_{b-1}). \tag{7.1}$$

That is, we treat the posterior after $b-1$ minibatches as the new prior for the incoming data points. If we can save the posterior from $b-1$ minibatches and calculate the normalizing constant for the bth posterior, repeated application of Eq. (7.1) is streaming; it automatically gives us the new posterior without needing to revisit old data points.

In complex models, it is often infeasible to calculate the posterior exactly, and an approximation must be used. Suppose that, given a prior $p(\Theta)$ and data minibatch C, we have an approximation algorithm $\mathcal{A}$ that calculates an approximate posterior q: $q(\Theta) = \mathcal{A}(C, p(\Theta))$. Then, setting $q_0(\Theta) = p(\Theta)$, one way to recursively calculate an approximation to the posterior is

$$p(\Theta \mid C_1, \ldots, C_b) \approx q_b(\Theta) = \mathcal{A}\left(C_b, q_{b-1}(\Theta)\right). \tag{7.2}$$

When $\mathcal{A}$ yields the posterior from Bayes theorem, this calculation is exact. This approach already differs from that of (Hoffman, Blei, and Bach, 2010; C. Wang, Paisley, and Blei, 2011; Hoffman, Blei, Paisley, et al., 2013), which we will see (Section 7.3) directly approximates $p(\Theta \mid C_1, \ldots, C_B)$ for fixed B without making intermediate approximations for b strictly between 1 and B.

Distributed Bayesian updating. The sequential updates in Eq. (7.2) handle streaming data in theory, but in practice, the $\mathcal{A}$ calculation might take longer than the time interval between minibatch arrivals or simply take longer than desired. Parallelizing computations increases algorithm throughput. And posterior calculations need not be sequential. Indeed, Bayes theorem yields

$$p(\Theta \mid C_1, \ldots, C_B) \propto \left[\prod_{b=1}^{B} p(C_b \mid \Theta)\right] p(\Theta) \propto \left[\prod_{b=1}^{B} p(\Theta \mid C_b)\, p(\Theta)^{-1}\right] p(\Theta). \tag{7.3}$$

That is, we can calculate the individual minibatch posteriors $p(\Theta \mid C_b)$, perhaps in parallel, and then combine them to find the full posterior $p(\Theta \mid C_1, \ldots, C_B)$.

Given an approximating algorithm $\mathcal{A}$ as above, the corresponding approximate update would be

$$p(\Theta \mid C_1, \ldots, C_B) \approx q(\Theta) \propto \left[\prod_{b=1}^{B} \mathcal{A}(C_b, p(\Theta))\, p(\Theta)^{-1}\right] p(\Theta), \tag{7.4}$$

for some approximating distribution q, provided the normalizing constant for the right-hand side of Eq. (7.4) can be computed.

Variational inference methods are generally based on exponential family representations (Wainwright and Jordan, 2008), and we will make that assumption here. In particular, we suppose $p(\Theta) \propto \exp\{\xi_0 \cdot T(\Theta)\}$; that is, $p(\Theta)$ is an exponential family distribution for Θ with sufficient statistic $T(\Theta)$ and natural parameter ξ_0. We suppose further that $\mathcal{A}$ always returns a distribution in the same exponential family; in particular, we suppose that there exists some parameter ξ_b such that

$$q_b(\Theta) \propto \exp\{\xi_b \cdot T(\Theta)\} \quad \text{for} \quad q_b(\Theta) = \mathcal{A}(C_b, p(\Theta)). \tag{7.5}$$

When we make these two assumptions, the update in Eq. (7.4) becomes

$$p(\Theta \mid C_1, \ldots, C_B) \approx q(\Theta) \propto \exp\left\{ \left[\xi_0 + \sum_{b=1}^{B} (\xi_b - \xi_0) \right] \cdot T(\Theta) \right\}, \tag{7.6}$$

where the normalizing constant is readily obtained from the exponential family form. In what follows we use the shorthand $\xi \leftarrow \mathcal{A}(C, \xi_0)$ to denote that $\mathcal{A}$ takes as input a minibatch C and a prior with exponential family parameter ξ_0 and that it returns a distribution in the same exponential family with parameter ξ.

So, to approximate $p(\Theta | C_1, \ldots, C_B)$, we first calculate ξ_b via the approximation primitive $\mathcal{A}$ for each minibatch C_b; note that these calculations may be performed in parallel. Then we sum together the quantities $\xi_b - \xi_0$ across b, along with the initial ξ_0 from the prior, to find the final exponential family parameter to the full posterior approximation q. We previously saw that the general Bayes sequential update can be made streaming by iterating with the old posterior as the new prior (Eq. (7.2)). Similarly, here we see that the full posterior approximation q is in the same exponential family as the prior, so one may iterate these parallel computations to arrive at a parallelized algorithm for streaming posterior computation.

We emphasize that while these updates are reminiscent of prior-posterior conjugacy, it is actually the approximate posteriors and single, original prior that we assume belong to the same exponential family. It is not necessary to assume any conjugacy in the generative model itself nor that any true intermediate or final posterior take any particular limited form.

Asynchronous Bayesian updating. Performing B computations in parallel can in theory speed up algorithm running time by a factor of B, but in practice it is often the case that a single computation thread takes longer than the rest. Waiting for this thread to finish diminishes potential gains from distributing the computations. This problem can be ameliorated by making computations *asynchronous*. In this case, processors known as *workers* each solve a subproblem. When a worker finishes, it reports its solution to a single *master* processor. If the master gives the worker a new subproblem without waiting for the other workers to finish, it can decrease downtime in the system.

Our asynchronous algorithm is in the spirit of Hogwild! (Niu et al., 2011). To present the algorithm we first describe an asynchronous computation that we will not use in practice, but which will serve as a conceptual stepping stone. Note in particular that the following scheme makes the computations in Eq. (7.6) asynchronous. Have each worker continuously iterate between three steps: (1) collect a new minibatch C, (2) compute the local approximate posterior $\xi \leftarrow \mathcal{A}(C, \xi_0)$, and (3) return $\Delta\xi := \xi - \xi_0$ to the master. The master, in turn, starts by assigning the posterior to equal the prior: $\xi^{(\text{post})} \leftarrow \xi_0$. Each time the master receives a quantity $\Delta\xi$ from any worker, it updates the posterior synchronously: $\xi^{(\text{post})} \leftarrow \xi^{(\text{post})} + \Delta\xi$. If $\mathcal{A}$ returns the exponential family parameter of the true posterior (rather than an approximation), then the posterior at the master is exact by Eq. (7.4).

A preferred asynchronous computation works as follows. The master initializes its posterior estimate to the prior: $\xi^{(\text{post})} \leftarrow \xi_0$. Each worker continuously iterates between four steps: (1) collect a new minibatch C, (2) copy the master posterior value locally $\xi^{(\text{local})} \leftarrow \xi^{(\text{post})}$, (3) compute the local approximate posterior $\xi \leftarrow \mathcal{A}(C, \xi^{(\text{local})})$, and (4) return $\Delta\xi := \xi - \xi^{(\text{local})}$ to the master. Each time the master receives a quantity $\Delta\xi$ from any worker, it updates the posterior synchronously: $\xi^{(\text{post})} \leftarrow \xi^{(\text{post})} + \Delta\xi$.

The key difference between the first and second frameworks proposed above is that, in the second, the latest posterior is used as a prior. This latter framework is more in line with the streaming update of Eq. (7.2) but introduces a new layer of approximation. Since $\xi^{(\text{post})}$ might change at the master while the worker is computing $\Delta\xi$, it is no longer the case that the posterior at the master is exact when $\mathcal{A}$ returns the exponential family parameter of the true posterior. Nonetheless we find that the latter framework performs better in practice, so we focus on it exclusively in what follows.

We refer to our overall framework as *SDA-Bayes*, which stands for (S)treaming, (D)istributed, (A)synchronous Bayes. The framework is intended to be general enough to allow a variety of local approximations $\mathcal{A}$. Indeed, SDA-Bayes works out of the box once an implementation of $\mathcal{A}$—and a prior on the global parameter(s) Θ—is provided. In the current chapter our preferred local approximation will be VB.

7.3 Case study: latent Dirichlet allocation

In what follows, we consider examples of the choices for the Θ prior and primitive $\mathcal{A}$ in the context of *latent Dirichlet allocation* (LDA) (Blei, Ng, and Jordan, 2003). LDA models the content of D documents in a corpus. Themes potentially shared by multiple documents are described by *topics*. The unsupervised learning problem is to learn the topics as well as discover which topics occur in which documents.

More formally, each topic (of K total topics) is a distribution over the V words in the vocabulary: $\beta_k = (\beta_{kv})_{v=1}^{V}$. Each document is an admixture of topics. The words in document d are assumed to be exchangeable. Each word w_{dn} belongs to a latent topic z_{dn} chosen according to a document-specific distribution of topics $\theta_d = (\theta_{dk})_{k=1}^{K}$. The full generative model, with Dirichlet priors for β_k and θ_d conditioned on respective parameters η_k and α, appears in (Blei, Ng, and Jordan, 2003).

To see that this model fits our specification in Section 7.2, consider the set of global parameters $\Theta = \beta$. Each document $w_d = (w_{dn})_{n=1}^{N_d}$ is distributed iid conditioned on the global topics. The full collection of data is a corpus $C = w = (w_d)_{d=1}^{D}$ of documents. The posterior for LDA, $p(\beta, \theta, z \mid C, \eta, \alpha)$, is equal to the following expression up to proportionality:

$$\propto \left[\prod_{k=1}^{K} \mathrm{Dirichlet}(\beta_k \mid \eta_k)\right] \cdot \left[\prod_{d=1}^{D} \mathrm{Dirichlet}(\theta_d \mid \alpha)\right] \cdot \left[\prod_{d=1}^{D}\prod_{n=1}^{N_d} \theta_{d z_{dn}} \beta_{z_{dn}, w_{dn}}\right]. \tag{7.7}$$

The posterior for just the global parameters $p(\beta|C, \eta, \alpha)$ can be obtained from $p(\beta, \theta, z|C, \eta, \alpha)$ by integrating out the local, document-specific parameters θ, z. As is common in complex models, the normalizing constant for Eq. (7.7) is intractable to compute, so the posterior must be approximated.

Posterior-approximation algorithms

To apply SDA-Bayes to LDA, we use the prior specified by the generative model. It remains to choose a posterior-approximation algorithm $\mathcal{A}$. We consider two possibilities here: variational Bayes (VB) and expectation propagation (EP). Both primitives take Dirichlet distributions as priors for β and both return Dirichlet distributions for the approximate posterior of the topic parameters β; thus the prior and approximate posterior are in the same exponential family. Hence both VB and EP can be utilized as a choice for $\mathcal{A}$ in the SDA-Bayes framework.

```
**Subroutine** LocalVB($d, \lambda$)
    **Output**: $(\gamma_d, \phi_d)$
    Initialize $\gamma_d$
    **while** $(\gamma_d, \phi_d)$ *not converged* **do**
        $\forall (k, v)$, set $\phi_{dvk} \propto \exp\left(\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kv}]\right)$ (normalized across $k$)
        $\forall k$, $\gamma_{dk} \leftarrow \alpha_k + \sum_{v=1}^{V} \phi_{dvk} n_{dv}$
    **end**
```

Algorithm 7.1: Subroutine LocalUpdate(d, λ), used by the global variational algorithms. Here, n_{dv} represents the number of words v in document d.

Mean-field variational Bayes. We use the shorthand p_D for Eq. (7.7), the posterior given D documents. We assume the approximating distribution, written q_D for shorthand, takes the form

$$q_D(\beta, \theta, z \mid \lambda, \gamma, \phi) = \left[\prod_{k=1}^{K} q_D(\beta_k \mid \lambda_k)\right] \cdot \left[\prod_{d=1}^{D} q_D(\theta_d \mid \gamma_d)\right] \cdot \left[\prod_{d=1}^{D}\prod_{n=1}^{N_d} q_D(z_{dn} \mid \phi_{d w_{dn}})\right] \tag{7.8}$$

for parameters $(\lambda_{kv}), (\gamma_{dk}), (\phi_{dvk})$ with $k \in \{1, \ldots, K\}, v \in \{1, \ldots, V\}, d \in \{1, \ldots, D\}$. Moreover, we set $q_D(\beta_k \mid \lambda_k) = \mathrm{Dirichlet}_V(\beta_k \mid \lambda_k)$, $q_D(\theta_d \mid \gamma_d) = \mathrm{Dirichlet}_K(\theta_d \mid \gamma_d)$, and $q_D(z_{dn} \mid \phi_{d w_{dn}}) = \mathrm{Categorical}_K(z_{dn} \mid \phi_{d w_{dn}})$. The subscripts on Dirichlet and Categorical indicate the dimensions of the distributions (and of the parameters).

```
Input: Data $(n_d)_{d=1}^{D}$; hyperparameters $\eta, \alpha$
Output: $\lambda$
Initialize $\lambda$
while $(\lambda, \gamma, \phi)$ not converged do
    for $d = 1, \ldots, D$ do
        $(\gamma_d, \phi_d) \leftarrow$ LocalVB($d, \lambda$)
    end
    $\forall(k, v)$, $\lambda_{kv} \leftarrow \eta_{kv} + \sum_{d=1}^{D} \phi_{dvk} n_{dv}$
end
```

Algorithm 7.2: VB for LDA. Iterates multiple times through the data. Here, n_{dv} represents the number of words v in document d. See Alg. 7.1 for the subroutine `LocalVB`.

```
Input: Hyperparameters $\eta, \alpha$
Output: A sequence $\lambda^{(1)}, \lambda^{(2)}, \ldots$
Initialize $\forall(k, v)$, $\lambda_{kv}^{(0)} \leftarrow \eta_{kv}$
for $b = 1, 2, \ldots$ do
    Collect new data minibatch $C$
    foreach document indexed $d$ in $C$ do
        $(\gamma_d, \phi_d) \leftarrow$ LocalVB($d, \lambda$)
    end
    $\forall(k, v)$, $\lambda_{kv}^{(b)} \leftarrow \lambda_{kv}^{(b-1)} + \sum_{d \text{ in } C} \phi_{dvk} n_{dv}$
end
```

Algorithm 7.3: SSU for LDA (streaming). Here, n_{dv} represents the number of words v in document d. See Alg. 7.1 for the subroutine `LocalVB`.

```
Input: Hyperparameters $\eta, \alpha, D, (\rho_t)_{t=1}^{T}$
Output: $\lambda$
Initialize $\lambda$
for $t = 1, \ldots, T$ do
    Collect new data minibatch $C$
    foreach document indexed $d$ in $C$ do
        $(\gamma_d, \phi_d) \leftarrow$ LocalVB($d, \lambda$)
    end
    $\forall(k, v)$, $\tilde{\lambda}_{kv} \leftarrow \eta_{kv} + \frac{D}{|C|} \sum_{d \text{ in } C} \phi_{dvk} n_{dv}$
    $\forall(k, v)$, $\lambda_{kv} \leftarrow (1 - \rho_t)\lambda_{kv} + \rho_t \tilde{\lambda}_{kv}$
end
```

Algorithm 7.4: SVI for LDA (single-pass). Here, n_{dv} represents the number of words v in document d. See Alg. 7.1 for the subroutine `LocalVB`.

The problem of VB is to find the best approximating q_D, defined as the collection of variational parameters λ, γ, ϕ that minimize the KL divergence from the true posterior: $\mathrm{KL}\left(q_D \parallel p_D\right)$. Even finding the minimizing parameters is a difficult optimization problem. Typically the solution is approximated by coordinate descent in each parameter (Blei, Ng, and Jordan, 2003; Wainwright and Jordan, 2008) as in Alg. 7.2. The derivation of VB for LDA can be found in (Blei, Ng, and Jordan, 2003; Hoffman, Blei, Paisley, et al., 2013) and Appendix 7.A.

Expectation propagation. An EP (Minka, 2001b) algorithm for approximating the LDA posterior appears in Alg. 7.7 of Appendix 7.B. Alg. 7.7 differs from (Minka and Lafferty, 2002), which does not provide an approximate posterior for the topic parameters, and is instead our own derivation. Our version of EP, like VB, learns factorized Dirichlet distributions over topics.

Other single-pass algorithms for approximate LDA posteriors

The algorithms in Section 7.3 pass through the data multiple times and require storing the data set in memory—but are useful as primitives for SDA-Bayes in the context of the processing of minibatches of data. Next, we consider two algorithms that can pass through a data set just one time (*single pass*) and to which we compare in the evaluations (Section 7.4).

Stochastic variational inference. VB uses coordinate descent to find a value of q_D, Eq. (7.8), that locally minimizes the KL divergence, $\mathrm{KL}\left(q_D \parallel p_D\right)$. *Stochastic variational inference* (SVI) (Hoffman, Blei, and Bach, 2010; Hoffman, Blei, Paisley, et al., 2013) is exactly the application of a particular version of stochastic gradient descent to the same optimization problem. While stochastic gradient descent can often be viewed as a streaming algorithm, the optimization problem itself here depends on D via p_D, the posterior on D data points. We see that, as a result, D must be specified in advance, appears in each step of SVI (see Alg. 7.4), and is independent of the number of data points actually processed by the algorithm. Nonetheless, while one may choose to visit $D' \neq D$ data points or revisit data points when using SVI to estimate p_D (Hoffman, Blei, and Bach, 2010; Hoffman, Blei, Paisley, et al., 2013), SVI can be made single-pass by visiting each of D data points exactly once and then has constant memory requirements. We also note that two new parameters, $\tau_0 > 0$ and $\kappa \in (0.5, 1]$, appear in SVI, beyond those in VB, to determine a learning rate ρ_t as a function of iteration t: $\rho_t := (\tau_0 + t)^{-\kappa}$.

Sufficient statistics. On each round of VB (Alg. 7.2), we update the local parameters for all documents and then compute $\lambda_{kv} \leftarrow \eta_{kv} + \sum_{d=1}^{D} \phi_{dvk} n_{dv}$. An alternative single-pass (and indeed streaming) option would be to update the local parameters for each minibatch of documents as they arrive and then add the corresponding terms $\phi_{dvk} n_{dv}$ to the current estimate of λ for each document d in the minibatch. This essential idea has been proposed previously for models other than LDA by (Honkela and Valpola, 2003; Luts, Broderick, and Wand, 2012) and forms the basis of what we call the *sufficient statistics update algorithm* (SSU): Alg. 7.3. This algorithm is equivalent to SDA-Bayes with $\mathcal{A}$ chosen to be a single iteration over the global variable λ of VB (i.e., updating λ exactly once instead of iterating until convergence).

7.4 Evaluation

We follow (Hoffman, Blei, Paisley, et al., 2013) (and further (Teh, D. Newman, and Welling, 2006; Asuncion et al., 2009)) in evaluating our algorithms by computing (approximate) predictive probability. Under this metric, a higher score is better, as a better model will assign a higher probability to the held-out words.

We calculate predictive probability by first setting aside held-out testing documents $C^{(\text{test})}$ from the full corpus and then further setting aside a subset of held-out testing words $W_{d,\text{test}}$ in each testing document d. The remaining (training) documents $C^{(\text{train})}$ are used to estimate the global parameter posterior $q(\beta)$, and the remaining (training) words $W_{d,\text{train}}$ within the dth testing document are used to estimate the document-specific parameter posterior $q(\theta_d)$.[1] To calculate predictive probability, an approximation is necessary since we do not know the predictive distribution—just as we seek to learn the posterior distribution. Specifically, we calculate the normalized predictive distribution and report "log predictive probability" as

$$\frac{\sum_{d \in C^{(\text{test})}} \log p(W_{d,\text{test}} \mid C^{(\text{train})}, W_{d,\text{train}})}{\sum_{d \in C^{(\text{test})}} |W_{d,\text{test}}|} = \frac{\sum_{d \in C^{(\text{test})}} \sum_{w_{\text{test}} \in W_{d,\text{test}}} \log p(w_{\text{test}} \mid C^{(\text{train})}, W_{d,\text{train}})}{\sum_{d \in C^{(\text{test})}} |W_{d,\text{test}}|},$$

where we use the approximation

$$\begin{aligned} p(w_{\text{test}} \mid C^{(\text{train})}, W_{d,\text{train}}) &= \int_{\beta} \int_{\theta_d} \left(\sum_{k=1}^{K} \theta_{dk} \beta_{k w_{\text{test}}} \right) p(\theta_d \mid W_{d,\text{train}}, \beta)\, p(\beta \mid C^{(\text{train})})\, d\theta_d\, d\beta \\ &\approx \int_{\beta} \int_{\theta_d} \left(\sum_{k=1}^{K} \theta_{dk} \beta_{k w_{\text{test}}} \right) q(\theta_d)\, q(\beta)\, d\theta_d\, d\beta = \sum_{k=1}^{K} \mathbb{E}_q[\theta_{dk}]\, \mathbb{E}_q[\beta_{k w_{\text{test}}}]. \end{aligned}$$

To facilitate comparison with SVI, we use the Wikipedia and Nature corpora of (Hoffman, Blei, and Bach, 2010; C. Wang, Paisley, and Blei, 2011) in our experiments. These two corpora represent a range of sizes (3,611,558 training documents for Wikipedia and 351,525 for Nature) as well as different types of topics. We expect words in Wikipedia to represent an extremely broad range of topics whereas we expect words in Nature to focus more on the sciences. We further use the vocabularies of (Hoffman, Blei, and Bach, 2010; C. Wang, Paisley, and Blei, 2011) and SVI code available online at (Hoffman, 2010). We hold out 10,000 Wikipedia documents and 1,024 Nature documents (not included in the counts above) for testing. In the results presented in the main text, we follow (Hoffman, Blei, and Bach, 2010; Hoffman, Blei, Paisley, et al., 2013) in fitting an LDA model with $K = 100$ topics

[1] In all cases, we estimate $q(\theta_d)$ for evaluative purposes using VB since direct EP estimation takes prohibitively long.

	Wikipedia				Nature			
	32-SDA	1-SDA	SVI	SSU	32-SDA	1-SDA	SVI	SSU
Log pred prob	$\mathbf{-7.31}$	-7.43	-7.32	-7.91	-7.11	-7.19	$\mathbf{-7.08}$	-7.82
Time (hours)	**2.09**	43.93	7.87	8.28	**0.55**	10.02	1.22	1.27

Table 7.1: A comparison of (1) log predictive probability of held-out data and (2) running time of four algorithms: SDA-Bayes with 32 threads, SDA-Bayes with 1 thread, SVI, and SSU.

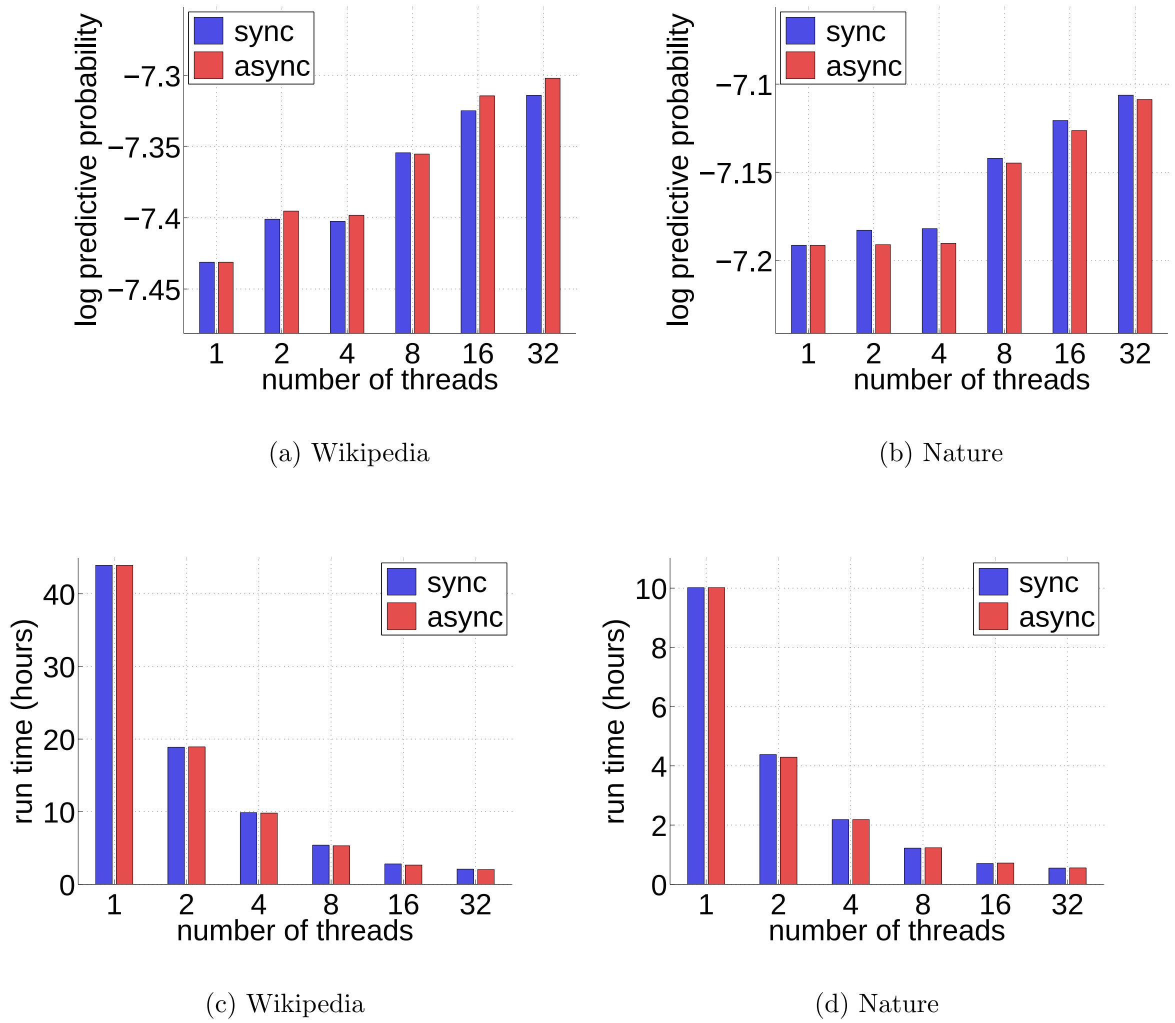
and hyperparameters chosen as: $\forall k, \alpha_k = 1/K$, $\forall (k,v), \eta_{kv} = 0.01$. For both Wikipedia and Nature, we set the parameters in SVI according to the optimal values of the parameters described in Table 1 of (Hoffman, Blei, and Bach, 2010) (number of documents D correctly set in advance, step size parameters $\kappa = 0.5$ and $\tau_0 = 64$).

Figures 7.4 and 7.4 demonstrate that both SVI and SDA are sensitive to minibatch size when $\eta_{kv} = 0.01$, with generally superior performance at larger batch sizes. Interestingly, both SVI and SDA performance improve and are steady across batch size when $\eta_{kv} = 1$ (Figures 7.4 and 7.4). Nonetheless, we use $\eta_{kv} = 0.01$ in what follows in the interest of consistency with (Hoffman, Blei, and Bach, 2010; Hoffman, Blei, Paisley, et al., 2013). Moreover, in the remaining experiments, we use a large minibatch size of $2^{15} = 32{,}768$. This size is the largest before SVI performance degrades in the Nature data set (Figure 7.4).

Performance and timing results are shown in Table 7.1. One would expect that with additional streaming capabilities, SDA-Bayes should show a performance loss relative to SVI. We see from Table 7.1 that such loss is small in the single-thread case, while SSU performs much worse. SVI is faster than single-thread SDA-Bayes in this single-pass setting.

Full SDA-Bayes improves run time with no performance cost. We handicap SDA-Bayes in the above comparisons by utilizing just a single thread. In Table 7.1, we also report performance of SDA-Bayes with 32 threads and the same minibatch size. In the synchronous case, we consider minibatch size to equal the total number of data points processed per round; therefore, the minibatch size equals the number of data points sent to each thread per round times the total number of threads. In the asynchronous case, we analogously report minibatch size as this product.

Figure 7.1 shows the performance of SDA-Bayes when we run with $\{1, 2, 4, 8, 16, 32\}$ threads while keeping the minibatch size constant. The goal in such a distributed context is to improve run time while not hurting performance. Indeed, we see dramatic run time improvement as the number of threads grows and in fact some slight performance improvement as well. We tried both a parallel version and a full distributed, asynchronous version of the algorithm; Figure 7.1 indicates that the speedup and performance improvements we see here come from parallelizing—which is theoretically justified by Eq. (7.3) when $\mathcal{A}$ is Bayes rule. Our experiments indicate that our Hogwild!-style asynchrony does not hurt performance. In our experiments, the processing time at each thread seems to be approximately equal across



(a) Wikipedia

(b) Nature

(c) Wikipedia

(d) Nature

Figure 7.1: SDA-Bayes log predictive probability (*two upper plots*) and run time (*two lower plots*) as a function of number of threads.

threads and dominate any communication time at the master, so synchronous and asynchronous performance and running time are essentially identical. In general, a practitioner might prefer asynchrony since it is more robust to node failures.

SVI is sensitive to the choice of total data size D. The evaluations above are for a single posterior over D data points. Of greater concern to us in this work is the evaluation of algorithms in the streaming setting. We have seen that SVI is designed to find the posterior for a particular, pre-chosen number of data points D. In practice, when we run SVI on the full data set but change the input value of D in the algorithm, we can see degradations in performance. In particular, we try values of D equal to $\{0.01, 0.1, 1, 10, 100\}$ times the true

D in Figure 7.4 for the Wikipedia data set and in Figure 7.4 for the Nature data set.

A practitioner in the streaming setting will typically not know D in advance, or multiple values of D may be of interest. Figures 7.4 and 7.4 illustrate that an estimate may not be sufficient. Even in the case where D is known in advance, it is reasonable to imagine a new influx of further data. One might need to run SVI again from the start (and, in so doing, revisit the first data set) to obtain the desired performance.

SVI is sensitive to learning step size. (Hoffman, Blei, and Bach, 2010; C. Wang, Paisley, and Blei, 2011) use cross-validation to tune step-size parameters (τ_0, κ) in the stochastic gradient descent component of the SVI algorithm. This cross-validation requires multiple runs over the data and thus is not suited to the streaming setting. Figures 7.4 and 7.4 demonstrate that the parameter choice does indeed affect algorithm performance. In these figures, we keep D at the true training data size.

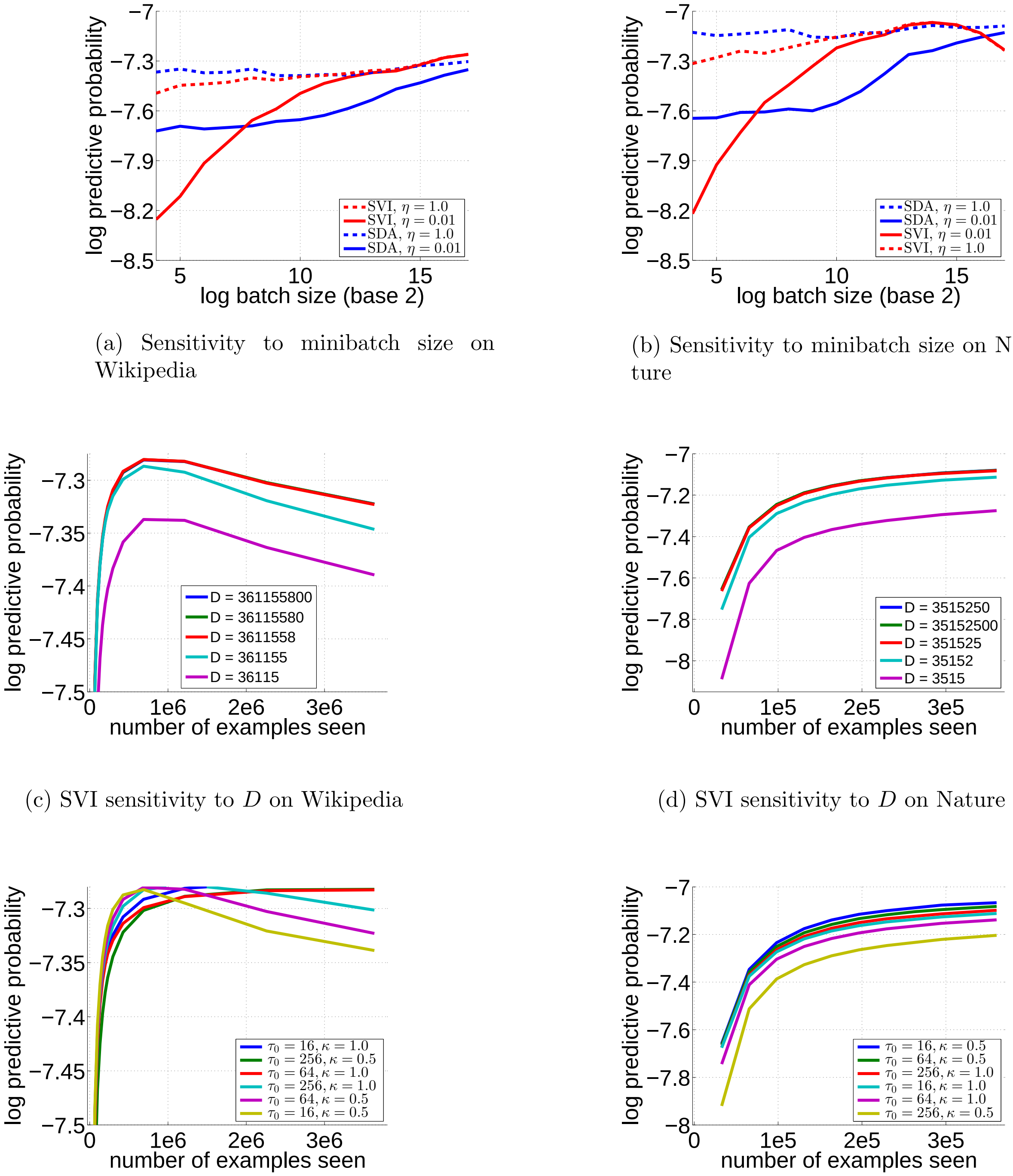
(Hoffman, Blei, and Bach, 2010) have observed that the optimal (τ_0, κ) may interact with minibatch size, and we further observe that the optimal values may vary with D as well. We also note that recent work has suggested a way to update (τ_0, κ) adaptively during an SVI run (Ranganath et al., 2013).

EP is not suited to LDA. Earlier attempts to apply EP to the LDA model in the non-streaming setting have had mixed success, with (Buntine and Jakulin, 2004) in particular finding that EP performance can be poor for LDA and, moreover, that EP requires "unrealistic intermediate storage requirements." We found this to also be true in the streaming setting. We were not able to obtain competitive results with EP; based on an 8-thread implementation of SDA-Bayes with an EP primitive[2], after over 91 hours on Wikipedia (and 6.7×10^4 data points), log predictive probability had stabilized at around -7.95 and, after over 97 hours on Nature (and 9.7×10^4 data points), log predictive probability had stabilized at around -8.02. Although SDA-Bayes with the EP primitive is not effective for LDA, it remains to be seen whether this combination may be useful in other domains where EP is known to be effective.

7.5 Discussion

We have introduced SDA-Bayes, a framework for streaming, distributed, asynchronous computation of an approximate Bayesian posterior. Our framework makes streaming updates to the estimated posterior according to a user-specified approximation primitive. We have demonstrated the usefulness of our framework, with variational Bayes as the primitive, by fitting the latent Dirichlet allocation topic model to the Wikipedia and Nature corpora. We have demonstrated the advantages of our algorithm over stochastic variational inference and the sufficient statistics update algorithm, particularly with respect to the key issue of obtaining approximations to posterior probabilities based on the number of documents seen thus far, not posterior probabilities for a fixed number of documents.

[2] We chose 8 threads since any fewer was too slow to get results and anything larger created too high of a memory demand on our system.



(a) Sensitivity to minibatch size on Wikipedia

(b) Sensitivity to minibatch size on Nature

(c) SVI sensitivity to D on Wikipedia

(d) SVI sensitivity to D on Nature

(e) SVI sensitivity to stepsize parameters on Wikipedia

(f) SVI sensitivity to stepsize parameters on Nature

Figure 7.2: Sensitivity of SVI and SDA-Bayes to some respective parameters. Legends have the same top-to-bottom order as the rightmost curve points.

7.A Variational Bayes

Batch VB

As described in the main text, the idea of VB is to find the distribution q_D that best approximates the true posterior, p_D. More specifically, the optimization problem of VB is defined as finding a q_D to minimize the KL divergence between the approximating distribution and the posterior:

$$\mathrm{KL}\left(q_D \parallel p_D\right) := \mathbb{E}_{q_D}\left[\log\left(q_D/p_D\right)\right]$$

Typically q_D takes a particular, constrained form, and finding the optimal q_D amounts to finding the optimal parameters for q_D. Moreover, the optimal parameters usually cannot be expressed in closed form, so often a coordinate descent algorithm is used.

For the LDA model, we have q_D in the form of Eq. (7.8) and p_D defined by Eq. (7.7). We wish to find the following variational parameters (i.e., parameters to q_D): λ (describing each topic), γ (describing the topic proportions in each document), and ϕ (describing the assignment of each word in each document to a topic).

Evidence lower bound

Finding q_D to minimize the KL divergence between q_D and p_D is equivalent to finding q_D to maximize the *evidence lower bound* (ELBO),

$$\begin{aligned}
\mathrm{ELBO} &:= \mathbb{E}_{q_D}\left[\log p(\Theta, x_{1:D})\right] - \mathbb{E}_{q_D}\left[\log q_D\right] \\
&= \mathbb{E}_{q_D}\left[\log p_D\right] + p(x_{1:D}) - \mathbb{E}_{q_D}\left[\log q_D\right] \\
&= -\mathrm{KL}\left(q_D \parallel p_D\right) + p(x_{1:D}),
\end{aligned}$$

since $p(x_{1:D})$ is constant in q_D. The VB optimization problem is often phrased in terms of the ELBO instead of the KL divergence.

The ELBO for LDA can be written as follows, where the model parameters are β, θ, z and the data is w; η and α are fixed hyperparameters.

$$\begin{aligned}
\mathrm{ELBO}(\lambda, \gamma, \phi) &= \mathbb{E}_q\left[\log p(\beta, \theta, z, w \mid \eta, \alpha)\right] - \mathbb{E}_q\left[\log q(\beta, \theta, z \mid \lambda, \gamma, \phi)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_q\left[\log \mathrm{Dirichlet}(\beta_k \mid \eta_k)\right] + \sum_{d=1}^{D} \mathbb{E}_q\left[\log \mathrm{Dirichlet}(\theta_d \mid \alpha)\right] \\
&\quad + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q\left[\log \mathrm{Multinomial}(z_{dn} \mid \theta_d)\right] \\
&\quad + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q\left[\log \mathrm{Multinomial}(w_{dn} \mid \beta_{z_{dn}})\right] \\
&\quad - \sum_{k=1}^{K} \mathbb{E}_q\left[\log \mathrm{Dirichlet}(\beta_k \mid \lambda_k)\right] - \sum_{d=1}^{D} \mathbb{E}_q\left[\log \mathrm{Dirichlet}(\theta_d \mid \gamma_d)\right]
\end{aligned}$$

$$- \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q \left[\log \text{Multinomial}(z_{dn} \mid \phi_{dw_{dn}}) \right].$$

The expectations in q in the previous equation can be evaluated as follows. The equations below make use of the *digamma function* ψ and *trigamma function* ψ_1. Here,

$$\begin{aligned} \psi(x) &= \frac{d}{dx} \log \Gamma(x) = \left[\frac{d}{dx} \Gamma(x) \right] / \Gamma(x) \\ \psi_1(x) &= \frac{d^2}{dx^2} \log \Gamma(x) = \frac{d}{dx} \psi(x). \end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}_q \left[\log \text{Dirichlet}(\beta_k \mid \eta_k) \right] \\
&\quad = \log \Gamma \left(\sum_{v=1}^{V} \eta_{kv} \right) - \sum_{v=1}^{V} \log \Gamma(\eta_{kv}) + \sum_{v=1}^{V} (\eta_{kv} - 1)\, \mathbb{E}_q[\log \beta_{kv}] \\
&\quad = \log \Gamma \left(\sum_{v=1}^{V} \eta_{kv} \right) - \sum_{v=1}^{V} \log \Gamma(\eta_{kv}) + \sum_{v=1}^{V} (\eta_{kv} - 1) \left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) \right) \\
&\mathbb{E}_q \left[\log \text{Dirichlet}(\theta_d \mid \alpha) \right] \\
&\quad = \log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) + \sum_{k=1}^{K} (\alpha_k - 1)\, \mathbb{E}_q[\log \theta_{dk}] \\
&\quad = \log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) + \sum_{k=1}^{K} (\alpha_k - 1) \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) \\
&\mathbb{E}_q \left[\log \text{Multinomial}(z_{dn} \mid \theta_d) \right] \\
&\quad = \sum_{k=1}^{K} \phi_{dw_{dn}k} \mathbb{E}_q[\log \theta_{dk}] \\
&\quad = \sum_{k=1}^{K} \phi_{dw_{dn}k} \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) \\
&\mathbb{E}_q \left[\log \text{Multinomial}(w_{dn} \mid \beta_{z_{dn}}) \right] \\
&\quad = \sum_{v=1}^{V} \mathbb{1}\{w_{dn} = v\}\, \mathbb{E}_q[\log \beta_{z_{dn},v}] \\
&\quad = \sum_{v=1}^{V} \mathbb{1}\{w_{dn} = v\} \sum_{k=1}^{K} \phi_{dw_{dn}k} \mathbb{E}_q[\log \beta_{kv}] \\
&\quad = \sum_{v=1}^{V} \sum_{k=1}^{K} \mathbb{1}\{w_{dn} = v\}\, \phi_{dw_{dn}k} \left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) \right)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_q\left[\log \mathrm{Dirichlet}(\beta_k \mid \lambda_k)\right] \\
&\qquad = \log\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right) - \sum_{v=1}^{V}\log\Gamma(\lambda_{kv}) + \sum_{v=1}^{V}(\lambda_{kv}-1)\,\mathbb{E}_q[\log\beta_{kv}] \\
&\qquad = \log\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right) - \sum_{v=1}^{V}\log\Gamma(\lambda_{kv}) + \sum_{v=1}^{V}(\lambda_{kv}-1)\left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right) \\
&\mathbb{E}_q\left[\log \mathrm{Dirichlet}(\theta_d \mid \gamma_d)\right] \\
&\qquad = \log\Gamma\left(\sum_{k=1}^{K}\gamma_{dk}\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_{dk}) + \sum_{k=1}^{K}(\gamma_{dk}-1)\,\mathbb{E}_q[\log\theta_{dk}] \\
&\qquad = \log\Gamma\left(\sum_{k=1}^{K}\gamma_{dk}\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_{dk}) + \sum_{k=1}^{K}(\gamma_{dk}-1)\left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K}\gamma_{dj}\Big)\right) \\
&\mathbb{E}_q\left[\log \mathrm{Multinomial}(z_{dn} \mid \phi_{dn})\right] \\
&\qquad = \sum_{k=1}^{K}\phi_{dw_{dn}k}\log\phi_{dw_{dn}k}.
\end{aligned}$$

Coordinate ascent

We maximize the ELBO via coordinate ascent in each dimension of the variational parameters: λ, γ, and ϕ.

Variational parameter λ. Choose a topic index k. Fix γ, ϕ, and each λ_j for $j \neq k$. Then we can write the ELBO's functional dependence on λ_k as follows, where "const" is a constant in λ_k.

$$\begin{aligned}
\mathrm{ELBO}(\lambda_k) = &\sum_{v=1}^{V}(\eta_{kv}-1)\left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right) \\
&+ \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{v=1}^{V}\mathbb{1}\{w_{dn}=v\}\,\phi_{dw_{dn}k}\left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right) \\
&- \log\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right) + \sum_{v=1}^{V}\log\Gamma(\lambda_{kv}) \\
&- \sum_{v=1}^{V}(\lambda_{kv}-1)\left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right) + \mathrm{const} \\
= &\sum_{v=1}^{V}\left(\eta_{kv} - \lambda_{kv} + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=v\}\,\phi_{dw_{dn}k}\right)\left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right)
\end{aligned}$$

$$-\log\Gamma\left(\sum_{v=1}^{V}\lambda_{kv}\right)+\sum_{v=1}^{V}\log\Gamma(\lambda_{kv})+\text{const}$$

The partial derivative of $\text{ELBO}(\lambda_k)$ with respect to one of the dimensions of λ_k, say λ_{kv}, is

$$\begin{aligned}
&\frac{\partial}{\partial\lambda_{kv}}\text{ELBO}(\lambda_k)\\
&\quad=-\left(\psi(\lambda_{kv})-\psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right)\\
&\qquad+\left(\eta_{kv}-\lambda_{kv}+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=v\}\,\phi_{dw_{dn}k}\right)\left(\psi_1(\lambda_{kv})-\psi_1\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\right)\\
&\qquad-\sum_{t:t\neq v}\left(\eta_{kt}-\lambda_{kt}+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=t\}\,\phi_{dw_{dn}k}\right)\,\psi_1\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)-\psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)+\psi(\lambda_{kv})\\
&\quad=\psi_1(\lambda_{kv})\,\left(\eta_{kv}-\lambda_{kv}+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=v\}\,\phi_{dw_{dn}k}\right)\\
&\qquad-\psi\Big(\sum_{u=1}^{V}\lambda_{ku}\Big)\,\sum_{u=1}^{V}\left(\eta_{ku}-\lambda_{ku}+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=u\}\,\phi_{dw_{dn}k}\right).
\end{aligned}$$

From the last line of the previous equation, we see that one can set the gradient of $\text{ELBO}(\lambda_k)$ to zero by setting

$$\lambda_{kv}\leftarrow\eta_{kv}+\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}\{w_{dn}=v\}\,\phi_{dw_{dn}k}\qquad\text{for } v=1,\ldots,V.$$

Equivalently, if n_{dv} is the number of occurrences (tokens) of word type v in document d, then the update may be written

$$\lambda_{kv}\leftarrow\eta_{kv}+\sum_{d=1}^{D}n_{dv}\,\phi_{dvk}\qquad\text{for } v=1,\ldots,V.$$

Variational parameter γ. Now choose a document d. Fix λ, ϕ, and γ_c for $c\neq d$. Then we can express the functional dependence of the ELBO on γ_d as follows.

$$\begin{aligned}
\text{ELBO}(\gamma_d)=&\sum_{k=1}^{K}(\alpha_k-1)\left(\psi(\gamma_{dk})-\psi\Big(\sum_{j=1}^{K}\gamma_{dj}\Big)\right)+\sum_{n=1}^{N_d}\sum_{k=1}^{K}\phi_{dw_{dn}k}\left(\psi(\gamma_{dk})-\psi\Big(\sum_{j=1}^{K}\gamma_{dj}\Big)\right)\\
&-\log\Gamma\left(\sum_{k=1}^{K}\gamma_{dk}\right)+\sum_{k=1}^{K}\log\Gamma(\gamma_{dk})-\sum_{k=1}^{K}(\gamma_{dk}-1)\left(\psi(\gamma_{dk})-\psi\Big(\sum_{j=1}^{K}\gamma_{dj}\Big)\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \text{const} \\
&= \sum_{k=1}^{K} \left(\alpha_k - \gamma_{dk} + \sum_{n=1}^{N_d} \phi_{d w_{dn} k} \right) \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) \\
&\quad - \log \Gamma \left(\sum_{k=1}^{K} \gamma_{dk} \right) + \sum_{k=1}^{K} \log \Gamma(\gamma_{dk}) + \text{const}
\end{aligned}
$$

The partial derivative of $\text{ELBO}(\gamma_d)$ with respect to one of the dimensions of γ_d, say γ_{dk}, is

$$
\begin{aligned}
&\frac{\partial}{\partial \gamma_{dk}} \text{ELBO}(\gamma_d) \\
&= - \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) + \left(\alpha_k - \gamma_{dk} + \sum_{n=1}^{N_d} \phi_{d w_{dn} k} \right) \left(\psi_1(\gamma_{dk}) - \psi_1\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) \\
&\quad - \sum_{i : i \neq k} \left(\alpha_i - \gamma_{di} + \sum_{n=1}^{N_d} \phi_{d w_{dn} i} \right) \ \psi_1\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) + \psi(\gamma_{dk}) \\
&= \psi_1(\gamma_{dk}) \left(\alpha_k - \gamma_{dk} + \sum_{n=1}^{N_d} \phi_{d w_{dn} k} \right) - \psi_1\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \sum_{j=1}^{K} \left(\alpha_j - \gamma_{dj} + \sum_{n=1}^{N_d} \phi_{d w_{dn} j} \right).
\end{aligned}
$$

As for the λ case above, one obvious way to achieve a gradient of $\text{ELBO}(\gamma_d)$ equal to zero is to set

$$
\gamma_{dk} \leftarrow \alpha_k + \sum_{n=1}^{N_d} \phi_{d w_{dn} k} \quad \text{for } k = 1, \ldots, K.
$$

Equivalently,

$$
\gamma_{dk} \leftarrow \alpha_k + \sum_{v=1}^{V} n_{dv} \ \phi_{dvk} \quad \text{for } k = 1, \ldots, K.
$$

Variational parameter ϕ. Finally, consider fixing λ, γ, and ϕ_{cu} for $(c, u) \neq (d, v)$. In this case, the dependence of the ELBO on ϕ_{dv} can be written as follows.

$$
\begin{aligned}
&\text{ELBO}(\phi_{dv}) \\
&= \sum_{k=1}^{K} n_{dv} \ \phi_{dvk} \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) \right) \\
&\quad + \sum_{k=1}^{K} n_{dv} \ \phi_{dvk} \left(\psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) \right) - \sum_{k=1}^{K} n_{dv} \ \phi_{dvk} \log \phi_{dvk} + \text{const} \\
&= \sum_{k=1}^{K} n_{dv} \ \phi_{dvk} \left(- \log \phi_{dvk} + \psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) + \psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) \right)
\end{aligned}
$$

$$+ \text{const}$$

The partial derivative of $\text{ELBO}(\phi_{dv})$ with respect to one of the dimensions of ϕ_{dv}, say ϕ_{dvk}, is

$$\begin{aligned}
&\frac{\partial}{\partial \phi_{dvk}} \text{ELBO}(\phi_{dv}) \\
&= n_{dv} \left(-\log \phi_{dvk} + \psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) + \psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) - 1 \right).
\end{aligned}$$

Using the method of Lagrange multipliers to incorporate the constraint that $\sum_{k=1}^{K} \phi_{dvk} = 1$, we wish to find ρ and ϕ_{dvk} such that

$$0 = \frac{\partial}{\partial \phi_{dvk}} \left[\text{ELBO}(\phi_{dv}) - \rho \left(\sum_{k=1}^{K} \phi_{dvk} - 1 \right) \right]. \tag{7.9}$$

Setting

$$\phi_{dvk} \propto_k \exp \left(\psi(\gamma_{dk}) - \psi\Big(\sum_{j=1}^{K} \gamma_{dj} \Big) + \psi(\lambda_{kv}) - \psi\Big(\sum_{u=1}^{V} \lambda_{ku} \Big) \right)$$

achieves the desired outcome in Eq. (7.9). Here, $\propto_k$ indicates that the proportionality is across k. The optimal choice of ρ is expressed via this proportionality. The above assignment may also be written as

$$\phi_{dvk} \propto_k \exp \left(\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kv}] \right)$$

The coordinate-ascent algorithm iteratively updates the parameters λ, γ, and ϕ. In practice, we usually iterate the updates for the "local" parameters ϕ and γ until they converge, then update the "global" parameter λ, and repeat. The resulting batch variational Bayes algorithm is presented in Alg. 7.2.

SDA-Bayes VB

For a fixed hyperparameter α, we can think of BatchVB as an algorithm that takes input in the form of a prior on topic parameters β and a minibatch of documents. In particular, let C_b be the bth minibatch of documents; for documents with indices in $\mathcal{D}_b$, these documents can be summarized by the word counts $(n_d)_{d \in \mathcal{D}_b}$. Then, in the notation of Eq. (7.2), we have $\Theta = \beta$, $\mathcal{A} = \text{BatchVB}$, and

$$q_0(\beta) = \prod_{k=1}^{K} \text{Dirichlet}(\beta_k | \eta_k).$$

In general, the bth posterior takes the same form and therefore can be summarized by its parameters $\lambda^{(b)}$:

$$q_b(\beta) = \prod_{k=1}^{K} \mathrm{Dirichlet}(\beta_k|\lambda_k^{(b)}).$$

In this case, if we set the prior parameters to $\lambda_k^{(0)} := \eta_k$, Eq. (7.2) becomes Alg. 7.5.

```
**Input**: Hyperparameter $\eta$
Initialize $\lambda^{(0)} \leftarrow \eta$
**foreach** *Minibatch $C_b$ of documents* **do**
    $\lambda^{(b)} \leftarrow \mathrm{BatchVB}\Big(C_b, \lambda^{(b-1)}\Big)$
    $q_b(\beta) = \prod_{k=1}^{K} \mathrm{Dirichlet}(\beta_k|\lambda_k^{(b)})$
**end**
```

Algorithm 7.5: Streaming VB for LDA.

Next, we apply the asynchronous, distributed updates described in the "Asynchronous Bayesian updating" portion of Section 7.2 to the batch VB primitive and LDA model. In this case, $\lambda^{(\mathrm{post})}$ is the posterior parameter estimate maintained at the master, and each worker updates this value after a local computation. The posterior after seeing a collection of minibatches is $q(\beta) = \prod_{k=1}^{K} \mathrm{Dirichlet}(\beta_k|\lambda_k^{(\mathrm{post})})$. The resulting algorithm is Alg. 7.6.

```
**Input**: Hyperparameter $\eta$
Initialize $\lambda^{(\mathrm{post})} \leftarrow \eta$
**foreach** *Minibatch $C_b$ of documents, at a worker* **do**
    Copy master value locally: $\lambda^{(local)} \leftarrow \lambda^{(\mathrm{post})}$ $\lambda \leftarrow \mathrm{BatchVB}\Big(C_b, \lambda^{(\mathrm{local})}\Big)$
    $\Delta\lambda \leftarrow \lambda - \lambda^{(\mathrm{local})}$
    Update the master value synchronously: $\lambda^{(\mathrm{post})} \leftarrow \lambda^{(\mathrm{post})} + \Delta\lambda$
**end**
```

Algorithm 7.6: SDA-Bayes with VB primitive for LDA.

7.B Expectation propagation

Batch EP

Our batch expectation propagation (EP) algorithm for LDA learns a posterior for both the document-specific topic mixing proportions $(\theta_d)_{d=1}^{D}$ and the topic distributions over words

$(\beta_k)_{k=1}^K$. By contrast, the algorithm in (Minka and Lafferty, 2002) learns only the former and so is not appropriate to the model in Section 7.3.

For consistency, we also follow Minka and Lafferty (2002) in making a distinction between token and type word updates, where a token refers to a particular word instance and a type refers to all words with the same vocabulary value. Let $C = (w_d)_{d=1}^D$ denote the set of documents that we observe, and for each word v in the vocabulary, let n_{dv} denote the number of times v appears in document d.

Collapsed posterior. We begin by collapsing (i.e., integrating out) the word assignments z in the posterior (7.7) of LDA. We can express the collapsed posterior as

$$p(\beta, \theta \mid C, \eta, \alpha) \propto \left[\prod_{k=1}^K \text{Dirichlet}_V(\beta_k \mid \eta_k)\right] \cdot \prod_{d=1}^D \left[\text{Dirichlet}_K(\theta_d \mid \alpha) \cdot \prod_{v=1}^V \left(\sum_{k=1}^K \theta_{dk}\, \beta_{kv}\right)^{n_{dv}}\right].$$

For each document-word pair (d, v), consider approximating the term $\sum_{k=1}^K \theta_{dk}\beta_{kv}$ above by

$$\left[\prod_{k=1}^K \text{Dirichlet}_V(\beta_k \mid \chi_{kdv} + \mathbf{1}_V)\right] \cdot \text{Dirichlet}_K(\theta_d \mid \zeta_{dv} + \mathbf{1}_K),$$

where $\chi_{kdv} \in \mathbb{R}^V$, $\zeta_{dv} \in \mathbb{R}^K$, and $\mathbf{1}_M$ is a vector of all ones of length M. This proposal serves as inspiration for taking the approximating variational distribution for $p(\beta, \theta \mid C, \eta, \alpha)$ to be of the form

$$q(\beta, \theta \mid \lambda, \gamma) := \left[\prod_{k=1}^K q(\beta_k \mid \lambda_k)\right] \cdot \prod_{d=1}^D q(\theta_d \mid \gamma_d), \tag{7.10}$$

where $q(\beta_k \mid \lambda_k) = \text{Dirichlet}(\beta_k \mid \lambda_k)$ and $q(\theta_d \mid \gamma_d) = \text{Dirichlet}(\theta_d \mid \gamma_d)$, with the parameters

$$\lambda_k = \eta_k + \sum_{d=1}^D \sum_{v=1}^V n_{dv}\chi_{kdv}, \qquad \gamma_d = \alpha + \sum_{v=1}^V n_{dv}\zeta_{dv}, \tag{7.11}$$

and the constraints $\lambda_k \in \mathbb{R}_+^V$ and $\gamma_d \in \mathbb{R}_+^K$ for each k and d. We assume this form in the remainder of the analysis and write $q(\beta, \theta \mid \chi, \zeta)$ for $q(\beta, \theta \mid \lambda, \gamma)$, where $\chi = (\chi_{kdv})$, $\zeta = (\zeta_{dv})$.

Optimization problem. We seek to find the optimal parameters (χ, ζ) by minimizing the (reverse) KL divergence:

$$\min_{\chi, \zeta} \text{KL}\left(p(\beta, \theta \mid C, \eta, \alpha) \parallel q(\beta, \theta \mid \chi, \zeta)\right).$$

This joint minimization problem is not tractable, and the idea of EP is to proceed iteratively by fixing most of the factors in Eq. (7.10) and minimizing the KL divergence over the parameters related to a single word.

More formally, suppose we already have a set of parameters (χ, ζ). Consider a document d and word v that occurs in document d (i.e., $n_{dv} \geq 1$). We start by removing the component of q related to (d, v) in Eq. (7.10). Following Minka (2001b), we subtract out the effect of one occurrence of word v in document d, but at the end of this process we update the distribution on the type level. In doing so, we use the following shorthand for the remaining global parameters:

$$\lambda_k^{\backslash(d,v)} = \lambda_k - \chi_{kdv} = \eta_k + (n_{dv} - 1)\chi_{kdv} + \sum_{(d',v'):(d',v')\neq(d,v)} n_{d'v'}\chi_{kd'v'}$$
$$\gamma_d^{\backslash(d,v)} = \gamma_d - \zeta_{dv} = \alpha + (n_{dv} - 1)\zeta_{dv} + \sum_{v':v'\neq v} n_{dv'}\zeta_{dv'}.$$

We replace this removed part of q by the term $\sum_{k=1}^{K} \theta_{dk}\beta_{kv}$, which corresponds to the contribution of one occurrence of word v in document d to the true posterior p. Call the resulting normalized distribution $\tilde{q}_{dv}$, so $\tilde{q}_{dv}(\beta, \theta \mid \lambda^{\backslash(d,v)}, \gamma_{\backslash d}, \gamma_d^{\backslash(d,v)})$ satisfies

$$\propto \left[\prod_{k=1}^{K} \text{Dirichlet}(\beta_k \mid \lambda_k^{\backslash(d,v)})\right] \cdot \left[\prod_{d'\neq d} \text{Dirichlet}(\theta_{d'} \mid \gamma_{d'})\right] \cdot \text{Dirichlet}(\theta_d \mid \gamma_d^{\backslash(d,v)}) \cdot \sum_{k=1}^{K} \theta_{dk}\, \beta_{kv}.$$

We obtain an improved estimate of the posterior q by updating the parameters from (λ, γ) to $(\hat{\lambda}, \hat{\gamma})$, where

$$(\hat{\lambda}, \hat{\gamma}) = \arg\min_{\lambda',\gamma'} \text{KL}\left(\tilde{q}_{dv}(\beta, \theta \mid \lambda^{\backslash(d,v)}, \gamma_{\backslash d}, \gamma_d^{\backslash(d,v)}) \parallel q(\beta, \theta \mid \lambda', \gamma')\right). \tag{7.12}$$

Solution to the optimization problem. First, note that for $d' : d' \neq d$, we have $\hat{\gamma}_{d'} = \gamma_{d'}$.

Now consider the index d chosen on this iteration. Since β and θ are Dirichlet-distributed under q, the minimization problem in Eq. (7.12) reduces to solving the moment-matching equations (Minka, 2001b; Seeger, 2005)

$$\mathbb{E}_{\tilde{q}_{dv}}[\log \beta_{ku}] = \mathbb{E}_{\hat{\lambda}_k}[\log \beta_{ku}] \qquad \text{for } 1 \leq k \leq K,\ 1 \leq u \leq V,$$
$$\mathbb{E}_{\tilde{q}_{dv}}[\log \theta_{dk}] = \mathbb{E}_{\hat{\gamma}_d}[\log \theta_{dk}] \qquad \text{for } 1 \leq k \leq K.$$

These can be solved via Newton's method though Minka (2001b) recommends solving exactly for the first and "average second" moments of β_{ku} and θ_{dk}, respectively, instead. We choose the latter approach for consistency with Minka (2001b); our own experiments also suggested taking the approach of Minka (2001b) was faster than Newton's method with no noticeable performance loss. The resulting moment updates are

$$\hat{\lambda}_{ku} = \frac{\sum_{y=1}^{V} \left(\mathbb{E}_{\tilde{q}_{dv}}[\beta_{ky}^2] - \mathbb{E}_{\tilde{q}_{dv}}[\beta_{ky}]\right)}{\sum_{y=1}^{V} \left(\mathbb{E}_{\tilde{q}_{dv}}[\beta_{ky}]^2 - \mathbb{E}_{\tilde{q}_{dv}}[\beta_{ky}^2]\right)} \cdot \mathbb{E}_{\tilde{q}_{dv}}[\beta_{ku}] \tag{7.13}$$

$$\hat{\gamma}_{dk} = \frac{\sum_{j=1}^{K} \left(\mathbb{E}_{\tilde{q}_{dv}}[\theta_{dj}^2] - \mathbb{E}_{\tilde{q}_{d,n}}[\theta_{dj}]\right)}{\sum_{j=1}^{K} \left(\mathbb{E}_{\tilde{q}_{dv}}[\theta_{dj}]^2 - \mathbb{E}_{\tilde{q}_{dv}}[\theta_{dj}^2]\right)} \cdot \mathbb{E}_{\tilde{q}_{dv}}[\theta_{dk}]. \tag{7.14}$$

We then set $(\chi_{kdv})_{k=1}^{K}$ and ζ_{dv} such that the new global parameters $(\lambda_k)_{k=1}^{K}$ and γ_d are equal to the optimal parameters $(\hat{\lambda}_k)_{k=1}^{K}$ and $\hat{\gamma}_d$. The resulting algorithm is presented in Alg. 7.7.

```
Input: Data $C = (w_d)_{d=1}^{D}$; hyperparameters $\eta, \alpha$
Output: $\lambda$
Initialize $\forall(k, d, v)$, $\chi_{kdv} \leftarrow 0$ and $\zeta_{dv} \leftarrow 0$
while $(\chi, \zeta)$ not converged do
    foreach $(d, v)$ with $n_{dv} \geq 1$ do
        /* Variational distribution without the word token $(d, v)$          */
        $\forall k,\ \lambda_k^{\setminus(d,v)} \leftarrow \eta_k + (n_{dv} - 1)\chi_{kdv} + \sum_{(d',v') \neq (d,v)} n_{d'v'}\chi_{kd'v'}$
            $\gamma_d^{\setminus(d,v)} \leftarrow \alpha + (n_{dv} - 1)\zeta_{dv} + \sum_{v' \neq v} n_{dv'}\zeta_{dv'}$
            If any of $\lambda_{ku}^{\setminus(d,v)}$ or $\gamma_{dk}^{\setminus(d,v)}$ are non-positive, skip updating this $(d, v)$   (†)
            /* Variational parameters from moment-matching                    */
        $\forall(k, u)$, compute $\hat{\lambda}_{ku}$ from Eq. (7.13)
            $\forall k$, compute $\hat{\gamma}_{dk}$ from Eq. (7.14)
            /* Type-level updates to parameter values                         */
        $\forall k,\ \chi_{kdv} \leftarrow n_{dv}^{-1}\left(\hat{\lambda}_k - \lambda_k^{\setminus(d,v)}\right) + \left(1 - n_{dv}^{-1}\right)\chi_{kdv}$
        $\zeta_{dv} \leftarrow n_{dv}^{-1}\left(\hat{\gamma}_d - \gamma_d^{\setminus(d,v)}\right) + \left(1 - n_{dv}^{-1}\right)\zeta_{dv}$
        Other $\chi, \zeta$ remain unchanged
    end
end
/* Global variational parameters                                              */
$\forall k,\ \lambda_k \leftarrow \eta_k + \sum_{d=1}^{D}\sum_{v=1}^{V} n_{dv}\chi_{kdv}$
```

Algorithm 7.7: EP for LDA.

The results in the main text (Section 7.4) are reported for Alg. 7.7. We also tried a slightly modified EP algorithm that makes token-level updates to parameter values, rather than type-level updates. This modified version iterates through each word *placeholder* in document d; that is, through pairs (d, n) rather than pairs (d, v) corresponding to word *values*. Since there are always at least as many (d, n) pairs as (d, v) pairs with $n_{dv} \geq 1$ (and usually many more of the former), the modified algorithm requires many more iterations. In practice, we find better experimental performance for the modified EP algorithm in terms of log predictive probability as a function of number of data points in the training set seen so far: e.g., leveling off at about −7.96 for Nature vs. −8.02. However, the modified algorithm is also much slower, and still returns much worse results than SDA-Bayes or SVI, so we do not report these results in the main text.[3]

SDA-Bayes EP

Putting a batch EP algorithm for LDA into the SDA-Bayes framework is almost identical to putting a batch VB algorithm for LDA into the SDA-Bayes framework. This similarity is to be expected since SDA-Bayes works out of the box with a batch approximation algorithm in the correct form.

For a fixed hyperparameter α, we can think of BatchEP as an algorithm (just like BatchVB) that takes input in the form of a prior on topic parameters β and a minibatch of documents. The same setup and notation from Appendix 7.A applies. In this case, Eq. (7.2) becomes Alg. 7.8. This algorithm is exactly the same as Alg. 7.5 but with a batch EP primitive instead of a batch VB primitive.

```
Input: Hyperparameter $\eta$
Initialize $\lambda^{(0)} \leftarrow \eta$
foreach Minibatch $C_b$ of documents do
    $\lambda^{(b)} \leftarrow \mathrm{BatchEP}\Big(C_b, \lambda^{(b-1)}\Big)$
    $q_b(\beta) = \prod_{k=1}^{K} \mathrm{Dirichlet}(\beta_k | \lambda_k^{(b)})$
end
```

Algorithm 7.8: Streaming EP for LDA.

Next, we apply the asynchronous, distributed updates described in the "Asynchronous Bayesian updating" portion of Section 7.2 to the batch EP primitive and LDA model. Again, the setup and notation from Appendix 7.A applies, and we find Alg. 7.9. Indeed, the recipe outlined here applies more generally to other primitives besides EP and VB.

```
Input: Hyperparameter $\eta$
Initialize $\lambda^{(\mathrm{post})} \leftarrow \eta$
foreach Minibatch $C_b$ of documents, at a worker do
    Copy master value locally: $\lambda^{(local)} \leftarrow \lambda^{(\mathrm{post})}$ $\lambda \leftarrow \mathrm{BatchEP}\Big(C_b, \lambda^{(\mathrm{local})}\Big)$
    $\Delta\lambda \leftarrow \lambda - \lambda^{(\mathrm{local})}$
    Update the master value synchronously: $\lambda^{(\mathrm{post})} \leftarrow \lambda^{(\mathrm{post})} + \Delta\lambda$
end
```

Algorithm 7.9: SDA-Bayes with EP primitive for LDA.

[3] Here and in the main text we run EP with $\eta = 1$. We also tried EP with $\eta = 0.01$, but the positivity check for $\lambda_{ku}^{\setminus(d,v)}$ and $\gamma_{dk}^{\setminus(d,v)}$ on line (†) in Alg. 7.7 always failed and as a result none of the parameters were updated.

Chapter 8

MAD-Bayes: MAP-based asymptotic derivations from Bayes

The classical mixture of Gaussians model is related to K-means via *small-variance asymptotics*: as the covariances of the Gaussians tend to zero, the negative log-likelihood of the mixture of Gaussians model approaches the K-means objective, and the EM algorithm approaches the K-means algorithm. Kulis and Jordan (2012) used this observation to obtain a novel K-means-like algorithm from a Gibbs sampler for the Dirichlet process (DP) mixture. We instead consider applying small-variance asymptotics directly to the posterior in Bayesian nonparametric models. This framework is independent of any specific Bayesian inference algorithm, and it has the major advantage that it generalizes immediately to a range of models beyond the DP mixture. To illustrate, we apply our framework to the feature learning setting, where the beta process and Indian buffet process provide an appropriate Bayesian nonparametric prior. We obtain a novel objective function that goes beyond clustering to learn (and penalize new) groupings for which we relax the mutual exclusivity and exhaustivity assumptions of clustering. We demonstrate several other algorithms, all of which are scalable and simple to implement. Empirical results demonstrate the benefits of the new framework.

8.1 Introduction

Clustering is a canonical learning problem and arguably the dominant application of unsupervised learning. Much of the popularity of clustering revolves around the K-means algorithm; its simplicity and scalability make it the preferred choice in many large-scale unsupervised learning problems—even though a wide variety of more flexible algorithms, including those from Bayesian nonparametrics, have been developed since the advent of K-means (Steinley, 2006; Jain, 2010). Indeed, Berkhin (2006) writes that K-means is “by far the most popular clustering tool used nowadays in scientific and industrial applications.”

K-means does have several known drawbacks. For one, the K-means algorithm clusters data into mutually exclusive and exhaustive clusters, which may not always be the optimal or desired form of latent structure for a data set. For example, pictures on a photo-sharing website might each be described by multiple tags, or social network users might be described by multiple interests. In these examples, a *feature allocation* in which each data point can belong to any nonnegative integer number of groups—now called *features*—is a more appropriate description of the data (Griffiths and Ghahramani, 2006; Broderick, Jordan, and Pitman, 2013). Second, the K-means algorithm requires advance knowledge of the number of clusters, which may be unknown or grow with the number of data points in some applications. A vast literature exists just on how to choose a number of clusters using heuristics or extensions of K-means (Steinley, 2006; Jain, 2010). A recent algorithm called DP-means (Kulis and Jordan, 2012) provides another perspective on the choice of cluster cardinality. Recalling the small-variance asymptotic argument that takes the EM algorithm for mixtures of Gaussians and yields the K-means algorithm, the authors apply this argument to a Gibbs sampler for a Dirichlet process (DP) mixture (Antoniak, 1974; Escobar, 1994; Escobar and West, 1995) and obtain a K-means-like algorithm that does not fix the number of clusters upfront.

Notably, this derivation of DP-means is specific to the choice of the sampling algorithm and is also not immediately amenable to the feature learning setting. In this chapter, we provide a more general perspective on these small-variance asymptotics. We show that one can obtain the objective function for DP-means (independent of any algorithm) by applying asymptotics directly to the MAP estimation problem of a Gaussian mixture model with a Chinese Restaurant Process (CRP) prior (Blackwell and MacQueen, 1973; Aldous, 1985) on the latent clustering. The key is to express the posterior in terms of the exchangeable partition probability function (EPPF) of the CRP (Pitman, 1995).

A critical advantage of this more general view of small-variance asymptotics is that it provides a framework for extending beyond the DP mixture. The Bayesian nonparametric toolbox contains many models that may yield—via small-variance asymptotics—a range of new algorithms that to the best of our knowledge have not been discovered in the K-means literature. We thus view our major contribution as providing new directions for researchers working on K-means and related discrete optimization problems.

To highlight this generality, we show how the framework may be used in the feature learning setting. We take as our point of departure the beta process (BP) (Hjort, 1990; Thibaux and Jordan, 2007), which is the feature learning counterpart of the DP, and the Indian Buffet Process (IBP) (Griffiths and Ghahramani, 2006), which is the feature learning counterpart of the CRP. We show how to express the corresponding MAP inference problem via an analogue of the EPPF that we refer to as an “exchangeable feature probability function” (EFPF) (Broderick, Pitman, and Jordan, 2013). Taking an asymptotic limit we obtain a novel objective function for feature learning, as well as a simple and scalable algorithm for learning features in a data set. The resulting algorithm, which we call *BP-means*, is similar to the DP-means algorithm, but allows each data point to be assigned to more than one feature. We also use our framework to derive several additional algorithms, including algorithms based on the Dirichlet-multinomial prior as well as extensions to the marginal MAP problem in which the cluster/feature means are integrated out. We compare our algorithms to existing Gibbs sampling methods as well as existing hard clustering methods in order to highlight the benefits of our approach.

8.2 MAP asymptotics for clusters

We begin with the problem setting of Kulis and Jordan (2012) but diverge in our treatment of the small-variance asymptotics. We consider a Bayesian nonparametric framework for generating data via a prior on clusterings and a likelihood that depends on the (random) clustering. Prior and likelihood yield a posterior distribution. A point estimate of the clustering (i.e., a hard clustering) may be achieved by choosing a clustering that maximizes the posterior; the result is a *maximum a posteriori* (MAP) estimate.

Consider a data set $x_1, \ldots, x_N$, where x_n is a D-component vector. Let K^+ denote the (random) number of clusters. Let z_{nk} equal one if data index n belongs to cluster k and 0 otherwise, so there is exactly one value of k for each n such that $z_{nk} = 1$. We can order the cluster labels k so that the first K^+ clusters are non-empty (i.e., $z_{nk} = 1$ for some n for each such k). Together K^+ and $z_{1:N,1:K^+}$ describe a clustering.

The Chinese restaurant process (CRP) (Blackwell and MacQueen, 1973; Aldous, 1985) gives a prior on K^+ and $z_{1:N,1:K^+}$ as follows. Let $\theta > 0$ be a hyperparameter of the model. The first customer (data index 1) starts a new table in the restaurant; i.e., $z_{1,1} = 1$. Recursively, the nth customer (data index n) sits at an existing table k with probability in proportion to the number of people sitting there (i.e., in proportion to $S_{n-1,k} := \sum_{m=1}^{n-1} z_{mk}$) and at a new table with probability proportional to θ.

Suppose the final restaurant has K^+ tables with N total customers sitting according to $z_{1:N,1:K^+}$. Then the probability of this clustering is found from the above recursion:

$$\mathbb{P}(z_{1:N,1:K^+}) = \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k} - 1)!, \tag{8.1}$$

a formula that is known as an exchangeable partition probability function (EPPF) Pitman, 1995.

A common choice for the likelihood is to assume that data in cluster k are Gaussian with cluster-specific mean μ_k and shared variance $\sigma^2 I_D$ (where I_D is the $D \times D$ identity matrix and $\sigma^2 > 0$). Then the likelihood of data $x = x_{1:N}$ given clustering $z = z_{1:N,1:K^+}$ and means $\mu = \mu_{1:K^+}$ is:

$$\mathbb{P}(x|z,\mu) = \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \mathcal{N}(x_n|\mu_k, \sigma^2 I_D).$$

Further suppose the μ_k are drawn iid Gaussian from a prior with mean 0 in every dimension and variance $\rho^2 I_D$ for hyperparameter $\rho^2 > 0$: $\mathbb{P}(\mu_{1:K^+}) = \prod_{k=1}^{K^+} \mathcal{N}(\mu_k|0, \rho^2 I_D)$.

The posterior distribution over the clustering given the observed data, $\mathbb{P}(z,\mu|x)$, is calculated from the prior and likelihood using Bayes theorem: $\mathbb{P}(z,\mu|x) \propto \mathbb{P}(x|z,\mu)\mathbb{P}(\mu)\mathbb{P}(z)$. We find the MAP point estimate for the clustering and cluster means by maximizing the posterior: $\text{argmax}_{K^+,z,\mu}\,\mathbb{P}(z,\mu|x)$. Note that the point estimate will be the same if we instead minimize the negative log joint likelihood: $\text{argmin}_{K^+,z,\mu} -\log \mathbb{P}(z,\mu,x)$.

In general, calculating the posterior or MAP estimate is difficult and usually requires approximation, e.g. via Markov chain Monte Carlo or a variational method. A different approximation can be obtained by taking the limit of the objective function above as the cluster variances decrease to zero: $\sigma^2 \to 0$. Since the prior allows an unbounded number of clusters, taking this limit will result in each data point being assigned to its own cluster in the MAP. To arrive at a limiting objective function that favors a non-trivial cluster assignment, we modulate the number of clusters via the hyperparameter θ, which varies linearly with the expected number of clusters in the prior. In particular, we choose some constant $\lambda^2 > 0$ and let $\theta = \exp(-\lambda^2/(2\sigma^2))$, so that, e.g., $\theta \to 0$ as $\sigma^2 \to 0$.

Substituting θ as a function of σ^2 and letting $\sigma^2 \to 0$, we find that $-2\sigma^2 \log \mathbb{P}(z,\mu,x)$ satisfies

$$\sim \sum_{k=1}^{K^+} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K^+ - 1)\lambda^2, \tag{8.2}$$

where $f(\sigma^2) \sim g(\sigma^2)$ here denotes $f(\sigma^2)/g(\sigma^2) \to 1$ as $\sigma^2 \to 0$. The double sum originates from the exponential function in the Gaussian data likelihood, and the penalty term—reminiscent of an AIC penalty (Akaike, 1974)—originates from the CRP prior (Appendix 8.A).

From Eq. (8.2), we see that finding the MAP estimate of the CRP Gaussian mixture model is asymptotically equivalent to the following optimization problem:

$$\underset{K^+,z,\mu}{\text{argmin}} \sum_{k=1}^{K^+} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K^+ - 1)\lambda^2. \tag{8.3}$$

Kulis and Jordan (2012) derived a similar objective function, which they called the *DP-means objective function* (a name we retain for Eq. (8.3)), by first deriving a K-means-style algorithm from a DP Gibbs sampler. Here, by contrast, we have found this objective function directly from the MAP problem, with no reference to any particular inference algorithm and thereby demonstrating a more fundamental link between the MAP problem and Eq. (8.3). In the following, we show that this focus on limits of a MAP estimate can yield useful optimization problems in diverse domains.

Notably, the objective in Eq. (8.3) takes the form of the K-means objective function (the double sum) plus a penalty of λ^2 for each cluster after the first; this offset penalty is natural since any partition of a non-empty set must have at least one cluster.[1] Once we

[1] The objective of Kulis and Jordan (2012) penalizes all K^+ clusters; the optimal arguments are the same in each case.

have Eq. (8.3), we may consider efficient solution methods; one candidate is the DP-means algorithm of Kulis and Jordan (2012).

8.3 MAP asymptotics for features

Once more consider a data set $x_{1:N}$, where x_n is a D-component vector. Now let K^+ denote the (random) number of features. Let z_{nk} equal one if data index n is in feature k and zero otherwise. In the feature case, while there must be a finite number of k values such that $z_{nk} = 1$ for any n, it is not required that there be exactly a single such k or even any such k. We order the feature labels k so that the first K^+ features are non-empty; i.e., we have $z_{nk} = 1$ for some n for each such k. Together K^+ and $z_{1:N,1:K^+}$ describe a feature allocation.

The Indian buffet process (IBP) (Griffiths and Ghahramani, 2006) is a prior on $z_{1:N,1:K^+}$ that places strictly positive probability on any finite, nonnegative value of K^+. Like the CRP, it is based on an analogy between the customers in a restaurant and the data indices. In the IBP, the dishes in the buffet correspond to features. Let $\gamma > 0$ be a hyperparameter of the model. The first customer (data index 1) samples $K_1^+ \sim \text{Poisson}(\gamma)$ dishes from the buffet. Recursively, when the nth customer (data index n) arrives at the buffet, $\sum_{m=1}^{n-1} K_m^+$ dishes have been sampled by the previous customers. Suppose dish k of these dishes has been sampled $S_{n-1,k}$ times by the first $n-1$ customers. The nth customer samples dish k with probability $S_{n-1,k}/n$. The nth customer also samples $K_n^+ \sim \text{Poisson}(\gamma/n)$ new dishes.

Suppose the buffet has been visited by N customers who sampled a total of K^+ dishes. Let $z = z_{1:N,1:K^+}$ represent the resulting feature allocation. Let H be the number of unique values of the $z_{1:N,k}$ vector across k; let $\tilde{K}_h$ be the number of k with the hth unique value of this vector. We calculate an "exchangeable feature probability function" (EFPF) (Broderick, Pitman, and Jordan, 2013) by multiplying together the probabilities from the N steps in the description and find that $\mathbb{P}(z)$ equals (Griffiths and Ghahramani, 2006)

$$\frac{\gamma^{K^+} \exp\left\{-\sum_{n=1}^{N} \frac{\gamma}{n}\right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} S_{N,k}^{-1} \binom{N}{S_{N,k}}^{-1}. \tag{8.4}$$

It remains to specify a probability for the observed data x given the latent feature allocation z. The linear Gaussian model of Griffiths and Ghahramani (2006) is a natural extension of the Gaussian mixture model to the feature case. As previously, we specify a prior on feature means $\mu_k \overset{iid}{\sim} \mathcal{N}(0, \rho^2 I_D)$ for some hyperparameter $\rho^2 > 0$. Now data point n is drawn independently with mean equal to the sum of its feature means, $\sum_{k=1}^{K^+} z_{nk}\mu_k$, and variance $\sigma^2 I_D$ for some hyperparameter $\sigma^2 > 0$. In the case where each data point belongs to exactly one feature, this model is just a Gaussian mixture. We often write the means as a $K \times D$ matrix A with kth row μ_k. Writing Z for the $N \times K$ matrix with (n, k) element z_{nk} and X for the $N \times D$ matrix with nth row x_n, we have $\mathbb{P}(X|Z, A)$ equal to

$$\frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{-\frac{\mathbf{tr}((X - ZA)'(X - ZA))}{2\sigma^2}\right\}. \tag{8.5}$$

As in the clustering case, we wish to find the joint MAP estimate of the structural component Z and group-specific parameters A. It is equivalent to find the values of Z and A that minimize $-\log \mathbb{P}(X, Z, A)$. Finally, we wish to take the limit of this objective as $\sigma^2 \to 0$. Lest every data point be assigned to its own separate feature, we modulate the number of features in the small-σ^2 limit by choosing some constant $\lambda^2 > 0$ and setting $\gamma = \exp(-\lambda^2/(2\sigma^2))$.

Letting $\sigma^2 \to 0$, we find that asymptotically (Appendix 8.B)

$$-2\sigma^2 \log \mathbb{P}(X, Z, A) \sim \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2,$$

The trace originates from the matrix Gaussian, and the penalty term originates from the IBP prior.

It follows that finding the MAP estimate for the feature learning problem is asymptotically equivalent to solving:

$$\operatorname*{argmin}_{K^+, Z, A} \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2. \tag{8.6}$$

We follow Kulis and Jordan (2012) in referring to the underlying random measure when naming objective functions derived from Bayesian nonparametric priors. Recalling that the beta process (BP) (Hjort, 1990; Thibaux and Jordan, 2007) is the random measure underlying the IBP, we call the objective in Eq. (8.6) the *BP-means objective*. The trace term in Eq. (8.6) forms a K-means-style objective on a feature matrix Z and feature means A when the number of features (i.e., the number of columns of Z or rows of A) is fixed. The second term enforces a penalty of λ^2 for each feature. In contrast to the DP-means objective, even the first feature is penalized since $K^+ = 0$ is allowed here.

We formulate a *BP-means algorithm* to solve the optimization problem in Eq. (8.6) and discuss its convergence properties. In Alg. 8.1, note that $Z'Z$ is invertible so long as no two features have the same collection of indices. If that is not the case, we simply combine the two features into a single feature before performing the inversion.

```
Iterate until no changes are made:
1. For $n = 1, \ldots, N$
    • For $k = 1, \ldots, K^+$, choose the optimal value (0 or 1) of $z_{nk}$.
    • Let $Z'$ equal $Z$ but with one new feature (labeled $K^+ + 1$) containing only data
      index $n$. Set $A' = A$ but with one new row: $A'_{K^+ + 1, \cdot} \leftarrow X_{n,\cdot} - Z_{n,\cdot}A$.
    • If the triplet $(K^+ + 1, Z', A')$ lowers the objective from the triplet $(K^+, Z, A)$,
      replace the latter triplet with the former.
2. Set $A \leftarrow (Z'Z)^{-1}Z'X$.
```

Algorithm 8.1: BP-means.

Proposition 8.3.1. *The BP-means algorithm converges after a finite number of iterations to a local minimum of the BP-means objective in Eq. (8.6).*

See Appendix 8.G for the proof. Though the proposition guarantees convergence, it does not guarantee convergence to the global optimum—an analogous result to those available for the K-means and DP-means algorithms (Kulis and Jordan, 2012). Many authors have noted the problem of local optima in the clustering literature (Steinley, 2006; Jain, 2010). One expects that the issue of local optima is only exacerbated in the feature domain, where the combinatorial landscape is much more complex. In clustering, this issue is often addressed by multiple random restarts and careful choice of cluster initialization; in Section 8.5 below, we also make use of random algorithm restarts and propose a feature initialization akin to one with provable guarantees for K-means clustering (Arthur and Vassilvitskii, 2007).

8.4 Extensions

We demonstrate our methodology using different priors on Z below and using different likelihoods in Appendix 8.F.

Collapsed objectives. It is believed that *collapsing* out the cluster or feature means from a Bayesian model by calculating instead the marginal structural posterior can improve MCMC sampler mixing in many scenarios (Liu, 1994). In the clustering case, collapsing translates to forming the posterior $\mathbb{P}(z|x) = \int_\mu \mathbb{P}(z, \mu|x)$. Note that even in the cluster case, we may use the matrix representations Z, X, and A so long as we make the additional assumption that $\sum_{k=1}^{K^+} z_{nk} = 1$ for each n. Finding the MAP estimate $\operatorname{argmax}_Z \mathbb{P}(Z|X)$ may, as usual, be accomplished by minimizing the negative log joint distribution with respect to Z. $\mathbb{P}(Z)$ is given by the CRP (Eq. (8.1)). $\mathbb{P}(X|Z)$ takes the form:

$$\frac{\exp\left\{-\frac{\mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X\right)}{2\sigma^2}\right\}}{(2\pi\sigma^2)^{ND/2}(\rho^2/\sigma^2)^{K^+D/2}|Z'Z + \frac{\sigma^2}{\rho^2} I_D|^{D/2}}. \tag{8.7}$$

Eq. (8.7) was derived by Griffiths and Ghahramani (2006) for linear-Gaussian features but applies to Gaussian clusters when Z encodes a clustering. Using the same asymptotics in σ^2 and θ as before, we find the limiting optimization problem (Appendix 8.C):

$$\operatorname*{argmin}_{K^+, Z} \mathbf{tr}(X'(I_N - Z(Z'Z)^{-1}Z')X) + (K^+ - 1)\lambda^2. \tag{8.8}$$

The first term in this objective was proposed, via independent considerations, by Gordon and Henderson (1977).

Simple algebraic manipulations allow us to rewrite the objective in a more intuitive format (Appendix 8.C):

$$\operatorname*{argmin}_{K^+, Z} \sum_{k=1}^{K^+} \sum_{n: z_{nk}=1} \|x_{n,\cdot} - \bar{x}^{(k)}\|_2^2 + (K^+ - 1)\lambda^2, \tag{8.9}$$

where $\bar{x}^{(k)} := S_{N,k}^{-1} \sum_{m:z_{mk}=1} x_{m,\cdot}$ is the kth empirical cluster mean, i.e., the mean of all data points assigned to cluster k. This *collapsed DP-means objective* is just the original DP-means objective in Eq. (8.3) with the cluster means replaced by empirical cluster means. A corresponding optimization algorithm appears in Alg. 8.2. A similar proof to that of Kulis and Jordan (2012) shows that this algorithm converges in a finite number of iterations to a local minimum of the objective.

```
Iterate until no changes are made:
1. For $n = 1, \ldots, N$
    • Assign $x_n$ to the closest cluster if the contribution to the objective in Eq. (8.9) from the squared distance is at most $\lambda^2$.
    • Otherwise, form a new cluster with just $x_n$.
```

Algorithm 8.2: Collapsed DP-means.

We have already noted that the likelihood associated with the Gaussian mixture model conditioned on a clustering is just a special case of the linear Gaussian model conditioned on a feature matrix. Therefore, it is not surprising that Eq. (8.7) also describes $\mathbb{P}(X|Z)$ when Z is a feature matrix. Now, $\mathbb{P}(Z)$ is given by the IBP (Eq. (8.4)). Using the same asymptotics in σ^2 and γ as in the joint MAP case, the MAP problem for feature allocation Z asymptotically becomes (Appendix 8.D):

$$\underset{K^+,Z}{\operatorname{argmin}} \, \mathbf{tr}(X'(I_N - Z(Z'Z)^{-1}Z')X) + K^+\lambda^2. \tag{8.10}$$

The key difference with Eq. (8.8) is that here Z may have any finite number of ones in each row. We call the objective in Eq. (8.10) the *collapsed BP-means objective.*

```
Repeat the following step until no changes are made:
1. For $n = 1, \ldots, N$
    • Choose $z_{n,1:K^+}$ to minimize the objective in Eq. (8.10). Delete any redundant features.
    • Add a new feature (indexed $K^+ + 1$) with only data index $n$ if doing so decreases the objective and if the feature would not be redundant.
```

Algorithm 8.3: Collapsed BP-means.

Just as the collapsed DP-means objective has an empirical cluster means interpretation, so does the collapsed BP-means objective have an interpretation in which the feature means matrix A in Eq. (8.6) is replaced by its empirical estimate $(Z'Z)^{-1}ZX$ (cf. Appendix 8.G). In particular, we can rewrite the objective in Eq. (8.10) as $\mathbf{tr}[(X - Z(Z'Z)^{-1}Z'X)'(X - Z(Z'Z)^{-1}Z'X)] + K^+\lambda^2$. A corresponding optimization algorithm appears in Alg. 8.3. A similar proof to that of Proposition 8.3.1 shows that this algorithm converges in a finite number of iterations to a local minimum of the objective.

Parametric objectives. The generative models studied so far are *nonparametric* in the usual Bayesian sense; there is no a priori bound on the number of cluster or feature parameters. The objectives above are similarly nonparametric. Parametric models, with a fixed bound on the number of clusters or features, are often useful as well. See Appendix 8.E for derivations of objectives for clustering and feature learning in the parametric case. Since below we apply the parametric version for the feature learning setting, which we call *K-features* (analogous to K-means but for feature learning), we include its description in Alg. 8.4.

```
Repeat until no changes are made:
1. For $n = 1, \ldots, N$
    • For $k = 1, \ldots, K$, set $z_{n,k}$ to minimize $\|x_{n,1:K} - z_{n,1:K}A\|^2$.
2. Set $A = (Z'Z)^{-1}Z'X$.
```

Algorithm 8.4: K-features.

8.5 Experiments

We examine collections of unlabeled data to discover latent shared features. We have already seen the BP-means and collapsed BP-means algorithms when the number of features is unknown. A third algorithm that we evaluate here involves running the K-features algorithm for different values of K and choosing the joint values of K, Z, A that minimize the BP-means objective in Eq. (8.6); we call this the *stepwise K-features algorithm.* If we assume the plot of the minimized K-features objective (Eq. (8.14)) as a function of K has increasing increments (i.e., decreasing negative increments), then we need only run the K-features algorithm for increasing K until the objective increases.

It is well known that the K-means algorithm is sensitive to the choice of cluster initialization (Peña, Lozano, and Larrañaga, 1999). Potential methods of addressing this issue include multiple random initializations and choosing initial, random cluster centers according to the K-means++ algorithm (Arthur and Vassilvitskii, 2007). In the style of K-means++, we introduce a similar feature means initialization.

We first consider fixed K. In K-means++, the initial cluster center is chosen uniformly at random from the data set. However, we note that empirically, the various feature algorithms discussed tend to prefer the creation of a *base feature*, shared amongst all the data. So start by assigning every data index to the first feature, and let the first feature mean be the mean of all the data points. Recursively, for feature k with $k > 1$, calculate the distance from each data point $x_{n,\cdot}$ to its feature representation $z_{n,\cdot}A$ for the construction thus far. Choose a data index n with probability proportional to this distance squared. Assign $A_{k,\cdot}$ to be the nth distance. Assign $z_{m,k}$ for all $m = 1, \ldots, N$ to optimize the K-features objective. In the case where K is not known in advance, we repeat the recursive step as long as doing so decreases the objective.

Another important consideration in running these algorithms without a fixed number of clusters or features is choosing the relative penalty effect λ^2. One option is to solve for λ^2 from a proposed K value via a heuristic (Kulis and Jordan, 2012) or validation on a data subset. Rather than assume K and return to it in this roundabout way, in the following we aim merely to demonstrate that there exist reasonable values of λ^2 that return meaningful results. More carefully examining the translation from a discrete (K) to continuous (λ^2) parameter space may be a promising direction for future work.

Tabletop data. Using a LogiTech digital webcam, Griffiths and Ghahramani (2006) took 100 pictures of four objects (a prehistoric handaxe, a Klein bottle, a cellular phone, and a \$20 bill) placed on a tabletop. The images are in JPEG format with 240 pixel height, 320 pixel width, and 3 color channels. Each object may or may not appear in a given picture; the experimenters endeavored to place each object (by hand) in a respective fixed location across pictures.

This setup lends itself naturally to the feature allocation domain. We expect to find a base feature depicting the tabletop and four more features, respectively corresponding to each of the four distinct objects. Conversely, clustering on this data set would yield either a cluster for each distinct feature combination—a much less parsimonious and less informative representation than the feature allocation—or some averages over feature combinations. The latter case again fails to capture the combinatorial nature of the data.

We emphasize a further point about identifiability within this combinatorial structure. One "true" feature allocation for this data is the one described above. But an equally valid allocation, from a combinatorial perspective, is one in which the base feature contains all four objects and the tabletop. There are four further features, each of which deletes an object and replaces it with tabletop so that every possible combination of objects on the tabletop can be constructed from the features. Indeed, any combination of objects on the tabletop could equally well serve as a base feature; the four remaining features serve to add or delete objects as necessary.

We run PCA on the data and keep the first $D = 100$ principal components to form the data vector for each image. This pre-processing is the same as that performed by Griffiths and Ghahramani (2006), except the authors in that case first average the three color channels of the images.

We consider the Gibbs sampling algorithm of Griffiths and Ghahramani (2006) with initialization (mass parameter 1 and feature mean variance 0.5) and number of sampling steps (1000) determined by the authors; we explore alternative initializations below. We compare to the three feature means algorithms described above—all with $\lambda^2 = 1$. Each of the final three algorithms uses the appropriate variant of greedy initialization analogous to K-means++. We run 1000 random initializations of the collapsed and BP-means algorithms to mitigate local minima. We run 300 random initializations of K-features for each value of K and note that $K = 2, \ldots, 6$ are (dynamically) explored by the algorithm. All code was run in Matlab on the same computer. Timing and feature count results are on the left of Figure 8.1.

While it is notoriously difficult to compare computation times for deterministic, hard-

Alg	Per run	Total	#
Gibbs	$8.5 \cdot 10^3$	—	10
Collap	11	$1.1 \cdot 10^4$	5
BP-m	0.36	$3.6 \cdot 10^2$	6
FeatK	0.10	$1.55 \cdot 10^2$	5

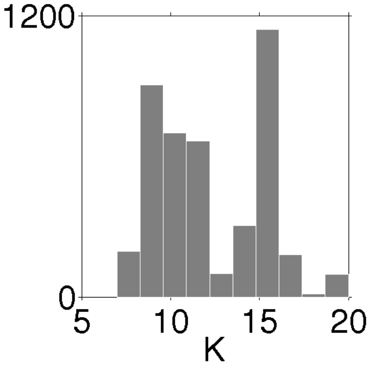


Figure 8.1: *Left*: A comparison of results for the IBP Gibbs sampler (Griffiths and Ghahramani, 2006), the collapsed BP-means algorithm, the basic BP-means algorithm, and the stepwise K-features algorithm. The first column shows the time for each run of the algorithm in seconds; the second column shows the total running time of the algorithm (i.e., over multiple repeated runs for the final three); and the third column shows the final number of features learned (the IBP # is stable for > 900 final iterations). *Right*: A histogram of collections of the final K values found by the IBP for a variety of initializations and parameter starting values.

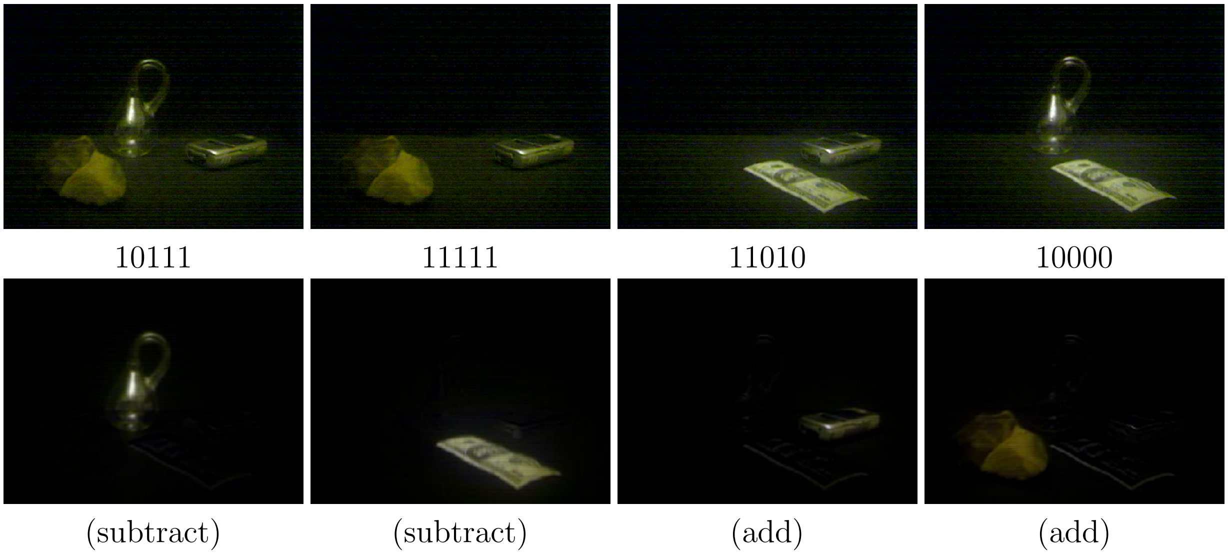


Figure 8.2: *Upper row*: Four example images in the tabletop data set. *Second row*: Feature assignments of each image. The first feature is the base feature, which depicts the Klein bottle and $20 bill on a tabletop and is almost identical to the fourth picture in the first row. The remaining four features are shown in order in the *third row*. The *fourth row* indicates whether the picture is added or subtracted when the feature is present.

assignment algorithms such K-means to stochastic algorithms such as Gibbs sampling, particularly given the practical need for reinitialization to avoid local minima in the former, and difficult-to-assess convergence in the latter, it should be clear from the first column in the lefthand table of Figure 8.1 that there is a major difference in computation time between Gibbs sampling and the new algorithms. Even when the BP-means algorithm is run 1000 times in a reinitialization procedure, the total time consumed is still an order of magnitude less than that for a single run of Gibbs sampling. Stepwise K-features is the fastest of the new algorithms.

We further note that if we were to take advantage of parallelism, additional drastic advantages could be obtained for the new algorithms. The Gibbs sampler requires each Gibbs iteration to be performed sequentially whereas the random initializations of the various feature means algorithms can be performed in parallel. A certain level of parallelism may even be exploited for the steps within each iteration of the collapsed and BP-means algorithms while the $z_{n,1:K}$ optimizations of K-features may all be performed in parallel across n

Another difficulty in comparing algorithms is that there is no clear single criterion with which to measure accuracy of the final model in unsupervised learning problems such as these. We do note, however, that theoretical considerations suggest that the IBP is not designed to find either a fixed number of features as N varies nor roughly equal sizes in those features it does find (Broderick, Jordan, and Pitman, 2012). This observation may help explain the distribution of observed feature counts over a variety of IBP runs with the given data. To obtain feature counts from the IBP, we tried running in a variety of different scenarios—combining different initializations (one shared feature, 5 random features, 10 random features, initialization with the BP-means result) and different starting parameter values[2] (mass parameter values ranging logarithmically from 0.01 to 1 and mean-noise parameter values ranging logarithmically from 0.1 to 10). The final 100 K draws for each of these combinations are aggregated and summarized in a histogram on the right of Figure 8.1. Feature counts lower than 7 were not obtained in our experiments, which suggests these values are, at least, difficult to obtain using the IBP with the given hyperpriors.

On the other hand, the feature counts for the new K-means-style algorithms suggest parsimony is more easily achieved in this case. The lower picture and text rows of Figure 8.2 show the features (after the base feature) found by stepwise K-features: as desired, there is one feature per tabletop object. The upper text row of Figure 8.2 shows the features to which each of the example images in the top row are assigned by the optimal feature allocation. For comparison, the collapsed algorithm also finds an optimal feature encoding. The BP-means algorithm adds an extra, superfluous feature containing both the Klein bottle and $20 bill.

Faces data. Next, we analyze the FEI face database, consisting of 400 pre-aligned images of faces (Thomaz and Giraldi, 2010). 200 different individuals are pictured, each with one smiling and one neutral expression. Each picture has height 300 pixels, width 250 pixels, and one grayscale channel. Four example pictures appear in the first row of Figure 8.3. This

[2] We found convergence failed for some parameter initializations outside this range.

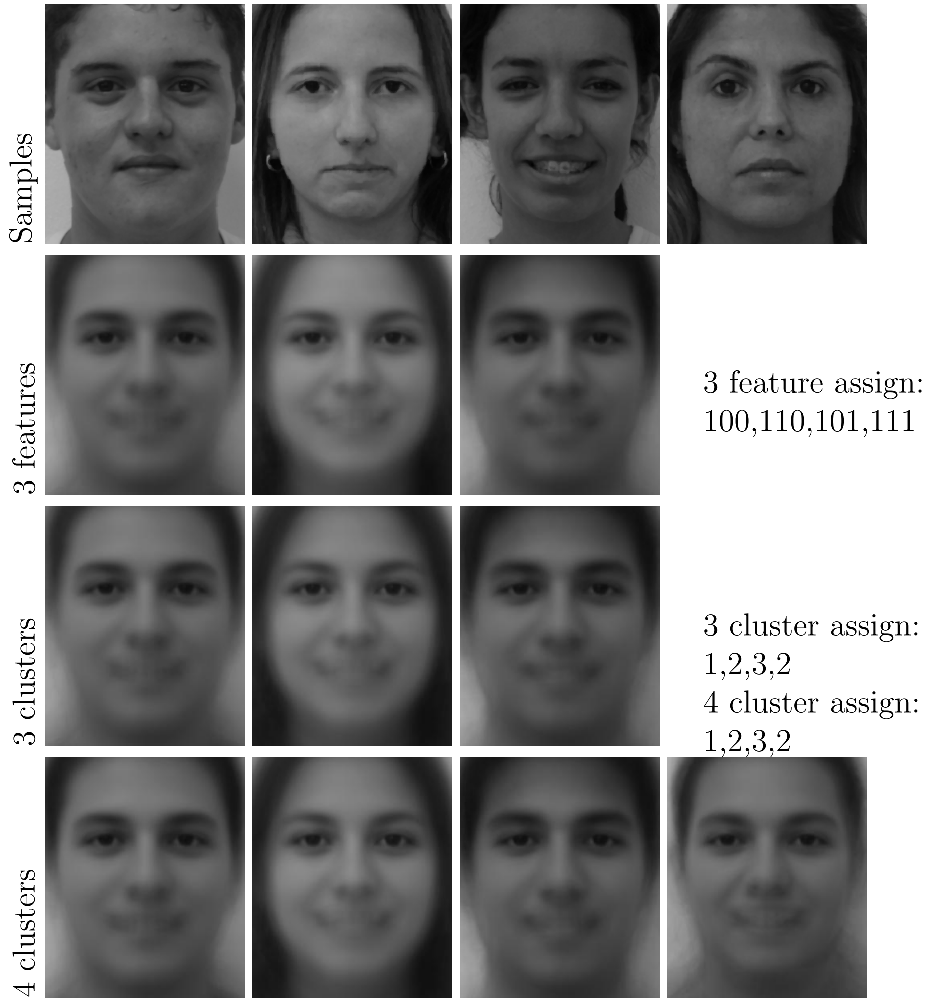


Figure 8.3: *1st row*: Four sample faces. *2nd row*: The base feature (left) and other 2 features returned by stepwise K-features with $\lambda^2 = 5$. The final pictures are the cluster means from K-means with $K = 3$ (*3rd row*) and $K = 4$ (*4th row*). The righthand text shows how the sample pictures are assigned to features/clusters by each algorithm.

time, we compare the stepwise K-features algorithm to classic K-means. We keep the top 100 principal components to form the data vectors for both algorithms.

Given $\lambda^2 = 5$, stepwise K-features chooses one base feature (lefthand picture in the second row of Figure 8.3) plus two additional features as optimal; the central and righthand pictures in the second row of Figure 8.3 depict the sum of the base feature plus the corresponding feature. The second feature codes for longer hair and a shorter chin relative to the base feature. The third feature codes for darker skin and slightly different facial features. The feature combinations of each picture in the first row appear in the first text row on the right;

all four possible combinations are represented.

K-means with 2 clusters and K-features with 2 features both encode exactly 2 distinct, disjoint groups. For larger numbers of groups though, the two representations diverge. For instance, consider a 3-cluster model of the face data, which has the same number of parameters as the 3-feature model. The resulting cluster means appear in the third row of Figure 8.3. While the cluster means appear similar to the feature means, the assignment of faces to clusters is quite different. The second righthand text row in Figure 8.3 shows to which cluster each of the four first-row faces is assigned. The feature allocation of the fourth picture in the top row tells us that the subject has long hair and certain facial features, roughly, whereas the clustering tells us that the subject's hair is more dominant than facial structure in determining grouping. Globally, the counts of faces for clusters (1,2,3) are (154,151,95) while the counts of faces for feature combinations (100,110,101,111) are (139,106,80,75).

We might also consider a clustering of size 4 since there are 4 groups specified by the 3-feature model. The resulting cluster means are in the bottom row of Figure 8.3, and the cluster assignments of the sample pictures are in the bottom, righthand text row. None of the sample pictures falls in cluster 4. Again, the groupings provided by the feature allocation and the clustering are quite different. Notably, the clustering has divided up the pictures with shorter hair into 3 separate clusters. In this case, the counts of faces for clusters (1,2,3,4) are (121,150,74,55). The feature allocation here seems to provide a sparser and more interpretable representation relative to both cluster cardinalities.

8.6 Discussion

We have developed a general methodology for obtaining hard-assignment objective functions from Bayesian MAP problems. The key idea is to include the structural variables explicitly in the posterior using combinatorial functions such as the EPPF and the EFPF. We apply this methodology to a number of generative models for unsupervised learning, with particular emphasis on latent feature models. We show that the resulting algorithms are capable of modeling latent structure out of reach of clustering algorithms but are also much faster than existing feature allocation learners from Bayesian nonparametrics.

8.A DP-means objective derivation

First consider the generative model in Section 8.2. The joint distribution of the observed data x, cluster indicators z, and cluster means μ can be written as follows.

$$\begin{aligned}\mathbb{P}(x, z, \mu) &= \mathbb{P}(x|z, \mu)\mathbb{P}(z)\mathbb{P}(\mu) \\ &= \prod_{k=1}^{K^+} \prod_{n: z_{n,k}=1} \mathcal{N}(x_n|\mu_k, \sigma^2 I_D)\end{aligned}$$

$$
\begin{aligned}
&\cdot \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k}-1)! \\
&\cdot \prod_{k=1}^{K^+} \mathcal{N}(\mu_k | 0, \rho^2 I_D).
\end{aligned}
$$

Then set $\theta := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. In the following, $f(\sigma^2) = O(g(\sigma^2))$ denotes that there exist some constants $c, s^2 > 0$ such that $|f(\sigma^2)| \leq c|g(\sigma^2)|$ for all $\sigma^2 < s^2$.

$$
\begin{aligned}
&- \log \mathbb{P}(x, z, \mu) \\
&= \sum_{k=1}^{K^+} \sum_{n : z_{n,k}=1} \left[O(\log \sigma^2) + \frac{1}{2\sigma^2} \|x_n - \mu_k\|^2 \right] \\
&+ (K^+ - 1) \frac{\lambda^2}{2\sigma^2} + O(1) \\
&+ O(1).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
-2\sigma^2 \log \mathbb{P}(x, z, \mu) = &\sum_{k=1}^{K^+} \sum_{n : z_{n,k}=1} \|x_n - \mu_k\|^2 \\
&+ (K^+ - 1)\lambda^2 + O(\sigma^2 \log(\sigma^2)).
\end{aligned}
$$

But since $\sigma^2 \log(\sigma^2) \to 0$ as $\sigma^2 \to 0$, we have that the remainder of the righthand side is asymptotically equivalent (as $\sigma^2 \to 0$) to the lefthand side (Eq. (8.2)).

8.B BP-means objective derivation

The recipe is the same as in Appendix 8.A. This time we start with the generative model in Section 8.3. The joint distribution of the observed data X, feature indicators Z, and feature means A can be written as follows.

$$
\begin{aligned}
&\mathbb{P}(X, Z, A) = \mathbb{P}(X|Z, A)\mathbb{P}(Z)\mathbb{P}(A) \\
&= \frac{1}{(2\pi\sigma^2)^{ND/2}} \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}((X - ZA)'(X - ZA)) \right\} \\
&\cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^{N} \frac{\gamma}{n} \right\}}{\prod_{h=1}^{H} \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N - S_{N,k})!}{N!} \\
&\cdot \frac{1}{(2\pi\rho^2)^{K^+D/2}} \exp\left\{ -\frac{1}{2\rho^2} \mathbf{tr}(A'A) \right\}.
\end{aligned}
$$

Now set $\gamma := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}
&-\log \mathbb{P}(X, Z, A) \\
&= O(\log \sigma^2) + \frac{1}{2\sigma^2}\mathbf{tr}((X - ZA)'(X - ZA)) \\
&+ K^+ \frac{\lambda^2}{2\sigma^2} + \exp(-\lambda^2/(2\sigma^2)) \sum_{n=1}^{N} n^{-1} + O(1) \\
&+ O(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
-2\sigma^2 \log \mathbb{P}(X, Z, A) &= \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2 \\
&+ O\left(\sigma^2 \exp(-\lambda^2/(2\sigma^2))\right) + O(\sigma^2 \log(\sigma^2)).
\end{aligned}$$

But since $\exp(-\lambda^2/(2\sigma^2)) \to 0$ and $\sigma^2 \log(\sigma^2) \to 0$ as $\sigma^2 \to 0$, we have that $-2\sigma^2 \log \mathbb{P}(X, Z, A) \sim \mathbf{tr}[(X - ZA)'(X - ZA)] + K^+\lambda^2$.

8.C Collapsed DP-means objective derivation

We apply the usual recipe as in Appendix 8.A. The generative model for collapsed DP-means is described in Section 8.4. The joint distribution of the observed data X and cluster indicators Z can be written as follows:

$$\begin{aligned}
&\mathbb{P}(X, Z) = \mathbb{P}(X|Z)\mathbb{P}(Z) \\
&= \left((2\pi)^{ND/2} (\sigma^2)^{(N-K^+)D/2} (\rho^2)^{K^+D/2} |Z'Z + \frac{\sigma^2}{\rho^2} I_D|^{D/2} \right)^{-1} \\
&\cdot \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \right\} \\
&\cdot \theta^{K^+-1} \frac{\Gamma(\theta + 1)}{\Gamma(\theta + N)} \prod_{k=1}^{K^+} (S_{N,k} - 1)!.
\end{aligned}$$

Now set $\theta := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}
&-\log \mathbb{P}(X, Z) = O(\log(\sigma^2)) \\
&+ \frac{1}{2\sigma^2} \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \\
&+ (K^+ - 1) \frac{\lambda^2}{2\sigma^2} + O(1).
\end{aligned}$$

It follows that

$$-2\sigma^2 \log \mathbb{P}(X, Z)$$

$$
\begin{aligned}
&= \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \\
&+ (K^+ - 1)\lambda^2 + O(\sigma^2 \log(\sigma^2)).
\end{aligned}
$$

We note that $\sigma^2 \log(\sigma^2) \to 0$ as $\sigma^2 \to 0$. Further note that $Z'Z$ is a diagonal $K \times K$ matrix with (k,k) entry (call it $S_{N,k}$) equal to the number of indices in cluster k. $Z'Z$ is invertible since we assume no empty clusters are represented in Z. Then

$$
\begin{aligned}
&-2\sigma^2 \log \mathbb{P}(X, Z) \\
&\sim \mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1} Z')X \right) + (K^+ - 1)\lambda^2
\end{aligned}
$$

as $\sigma^2 \to 0$.

More interpretable objective

The objective for the collapsed Dirichlet process is more interpretable after some algebraic manipulation. We describe here how the opaque $\mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1} Z')X \right)$ term can be written in a form more reminiscent of the $\sum_{k=1}^{K^+} \sum_{n: z_{n,k}=1} \|x_n - \mu_k\|^2$ term in the uncollapsed objective. First, recall that $C := Z'Z$ is a $K \times K$ matrix with $C_{k,k} = S_{N,k}$ and $C_{j,k} = 0$ for $j \neq k$. Then $C' := Z(Z'Z)^{-1}Z'$ is an $N \times N$ matrix with $C'_{n,m} = S_{N,k}^{-1}$ if and only if $z_{n,k} = z_{m,k} = 1$ and $C'_{n,m} = 0$ if $z_{n,k} \neq z_{m,k}$.

$$
\begin{aligned}
&\mathbf{tr}(X'(I_N - Z(Z'Z)^{-1} Z')X) \\
&= \mathbf{tr}(X'X) - \mathbf{tr}(X'Z(Z'Z)^{-1} Z'X) \\
&= \mathbf{tr}(XX') - \sum_{d=1}^{D} \sum_{k=1}^{K^+} \sum_{n: z_{n,k}=1} \sum_{m: z_{m,k}=1} S_{N,k}^{-1} X_{n,d} X_{m,d} \\
&= \sum_{k=1}^{K^+} \Bigg[\sum_{n: z_{n,k}=1} x_n x_n' - 2 S_{N,k}^{-1} \sum_{n: z_{n,k}=1} x_n \sum_{m: z_{m,k}=1} x_m' \\
&\quad + S_{N,k}^{-1} \sum_{n: z_{n,k}=1} x_n \sum_{m: z_{m,k}=1} x_m' \Bigg] \\
&= \sum_{k=1}^{K^+} \sum_{n: z_{n,k}=1} \|x_n - S_{N,k}^{-1} \sum_{m: z_{m,k}=1} x_{m,k}\|^2 \\
&= \sum_{k=1}^{K^+} \sum_{n: z_{n,k}=1} \|x_n - \bar{x}^{(k)}\|^2,
\end{aligned}
$$

for the cluster-specific empirical mean defined as $\bar{x}^{(k)} := S_{N,k}^{-1} \sum_{m: z_{m,k}=1} x_{m,k}$, as in the main text.

8.D Collapsed BP-means objective derivation

We continue to apply the usual recipe as in Appendix 8.A. The generative model for collapsed BP-means is described in Section 8.4. The joint distribution of the observed data X and feature indicators Z can be written as follows:

$$\begin{aligned}
\mathbb{P}(X,Z) &= \mathbb{P}(X|Z)\mathbb{P}(Z) \\
&= \left((2\pi)^{ND/2} (\sigma^2)^{(N-K^+)D/2} (\rho^2)^{K^+ D/2} |Z'Z + \frac{\sigma^2}{\rho^2} I_D|^{D/2} \right)^{-1} \\
&\cdot \exp\left\{ -\frac{1}{2\sigma^2} \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \right\} \\
&\cdot \frac{\gamma^{K^+} \exp\left\{ -\sum_{n=1}^N \frac{\gamma}{n} \right\}}{\prod_{h=1}^H \tilde{K}_h!} \prod_{k=1}^{K^+} \frac{(S_{N,k}-1)!(N-S_{N,k})!}{N!}.
\end{aligned}$$

Now set $\gamma := \exp(-\lambda^2/(2\sigma^2))$ and consider the limit $\sigma^2 \to 0$. Then

$$\begin{aligned}
-\log \mathbb{P}(X,Z) &= O(\log(\sigma^2)) \\
&+ \frac{1}{2\sigma^2} \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \\
&+ K^+ \frac{\lambda^2}{2\sigma^2} + \exp(-\lambda^2/(2\sigma^2)) \sum_{n=1}^N n^{-1} + O(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
-2\sigma^2 \log \mathbb{P}(X,Z) &= \mathbf{tr}\left(X'(I_N - Z(Z'Z + \frac{\sigma^2}{\rho^2} I_D)^{-1} Z')X \right) \\
&+ K^+\lambda^2 + O\left(\sigma^2 \exp(-\lambda^2/(2\sigma^2))\right) + O(\sigma^2 \log(\sigma^2)).
\end{aligned}$$

But $\exp(-\lambda^2/(2\sigma^2)) \to 0$ and $\sigma^2 \log(\sigma^2) \to 0$ as $\sigma^2 \to 0$. And $Z'Z$ is invertible so long as two features do not have identical membership (in which case we collect them into a single feature). So we have that $-2\sigma^2 \log \mathbb{P}(X,Z) \sim \mathbf{tr}\left(X'(I_N - Z(Z'Z)^{-1}Z')X\right) + K^+\lambda^2$.

8.E Parametric objectives

First, consider a clustering prior with some fixed maximum number of clusters K. Let $q_{1:K}$ represent a distribution over clusters. Suppose $q_{1:K}$ is drawn from a finite Dirichlet distribution with size $K > 1$ and parameter $\theta > 0$. Further, suppose the cluster for each data point is drawn iid according to $q_{1:K}$. Then, integrating out q, the marginal distribution of the clustering is Dirichlet-multinomial:

$$\mathbb{P}(z) = \frac{\Gamma(K\theta)}{\Gamma(N+K\theta)} \prod_{k=1}^K \frac{\Gamma(S_{N,k}+\theta)}{\Gamma(\theta)}. \tag{8.11}$$

We again assume a Gaussian mixture likelihood, only now the number of cluster means μ_k has an upper bound of K.

We can find the MAP estimate of z and μ under this model in the limit $\sigma^2 \to 0$. With θ fixed, the clustering prior has no effect, and the resulting optimization problem is $\operatorname{argmin}_{z,\mu} \sum_{k=1}^{K} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2$, which is just the usual K-means optimization problem.

We can also try scaling $\theta = \exp(-\lambda^2/(2\sigma^2))$ for some constant $\lambda^2 > 0$ as in the unbounded cardinality case. Then taking the $\sigma^2 \to 0$ limit of the log joint likelihood yields a term of λ^2 for each cluster containing at least one data index in the product in Eq. (8.11)—except for one such cluster. Call the number of such activated clusters K^+. The resulting optimization problem is

$$\operatorname*{argmin}_{K^+,z,\mu} \sum_{k=1}^{K} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K \wedge K^+ - 1)\lambda^2. \tag{8.12}$$

This objective caps the number of clusters at K but contains a penalty for each new cluster up to K.

A similar story holds in the feature case. Imagine that we have a fixed maximum of K features. In this finite case, we now let $q_{1:K}$ represent frequencies of each feature and let $q_k \overset{iid}{\sim} \text{Beta}(\gamma, 1)$. We draw $z_{nk} \sim \text{Bern}(q_k)$ iid across n and independently across k. The linear Gaussian likelihood model is as in Eq. (8.5) except that now the number of features is bounded. If we integrate out the $q_{1:K}$, the resulting marginal prior on Z is

$$\prod_{k=1}^{K} \left(\frac{\Gamma(S_{N,k} + \gamma)\Gamma(N - S_{N,k} + 1)}{\Gamma(N + \gamma + 1)} \frac{\Gamma(\gamma + 1)}{\Gamma(\gamma)\Gamma(1)} \right). \tag{8.13}$$

Then the limiting MAP problem as $\sigma^2 \to 0$ is

$$\operatorname*{argmin}_{Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)]. \tag{8.14}$$

This objective is analogous to the K-means objective but holds for the more general problem of feature allocations. Eq. (8.14) can be solved according to the *K features algorithm* (Alg. 8.4). Notably, all of the optimizations for n in the first step of the algorithm may be performed in parallel.

We can further set $\gamma = \exp(-\lambda^2/(2\sigma^2))$ as for the unbounded cardinality case before taking the limit $\sigma^2 \to 0$. Then a λ^2 term contributes to the limiting objective for each non-empty feature from the product in Eq. (8.13):

$$\operatorname*{argmin}_{K^+,Z,A} \mathbf{tr}[(X - ZA)'(X - ZA)] + (K \wedge K^+)\lambda^2, \tag{8.15}$$

reminiscent of the BP-means objective but with a cap of K possible features.

8.F General multivariate Gaussian likelihood

Above, we assumed a multivariate spherical Gaussian likelihood for each cluster. This assumption can be generalized in a number of ways. For instance, assume a general covariance matrix $\sigma^2 \Sigma_k$ for positive scalar σ^2 and positive definite $D \times D$ matrix Σ_k. Then we assume the following likelihood model for data points assigned to the kth cluster ($z_{n,k} = 1$): $x_n \sim \mathcal{N}(\mu_k, \sigma^2 \Sigma_k)$. Moreover, assume an inverse Wishart prior on the positive definite matrix Σ_k: $\Sigma_k \sim W^{-1}(\Phi, \nu)$ for Φ a positive definite matrix and $\nu > D - 1$. Assume a prior $\mathbb{P}(\mu)$ on μ that puts strictly positive density on all real-valued D-length vectors μ. For now we assume K is fixed and that $\mathbb{P}(z)$ puts a prior that has strictly positive density on all valid clusterings of the data points. Then

$$\begin{aligned}
&\mathbb{P}(x, z, \mu, \sigma^2 \Sigma) \\
&= \mathbb{P}(x | z, \mu, \sigma^2 \Sigma) \mathbb{P}(z) \mathbb{P}(\mu) \mathbb{P}(\Sigma) \\
&= \prod_{k=1}^{K^+} \prod_{n : z_{n,k}=1} \mathcal{N}(x_n | \mu_k, \sigma^2 \Sigma_k) \\
&\cdot \mathbb{P}(z) \mathbb{P}(\mu) \cdot \prod_{k=1}^{K} \left[\frac{|\Phi|^{\nu/2}}{2^{\nu D/2} \Gamma_D(\nu/2)} |\Sigma_k|^{-\frac{\nu+D+1}{2}} \right. \\
&\left. \cdot \exp\left\{ -\frac{1}{2} \mathbf{tr}(\Phi \Sigma_k^{-1}) \right\} \right],
\end{aligned}$$

where Γ_D is a multivariate gamma function. Consider the limit $\sigma^2 \to 0$. Set $\nu = \lambda^2/\sigma^2$ for some constant $\lambda^2 : \lambda^2 > 0$. Then

$$\begin{aligned}
&- \log \mathbb{P}(x, z, \mu, \sigma^2 \Sigma) \\
&= \sum_{k=1}^{K} \sum_{n : z_{n,k}=1} \left[O(\log \sigma^2) + \frac{1}{2\sigma^2} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k) \right] \\
&+ O(1) + \sum_{k=1}^{K} \left[-\frac{1}{2\sigma^2} \lambda^2 \log |\Phi| + \frac{D}{2\sigma^2} \lambda^2 \log 2 \right. \\
&\left. + \log \Gamma_D(\lambda^2/(2\sigma^2)) + \left(\frac{\lambda^2}{2\sigma^2} + \frac{D+1}{2} \right) \log |\Sigma_k| + O(1) \right].
\end{aligned}$$

So we find

$$\begin{aligned}
&-2\sigma^2 \left[\log \mathbb{P}(x, z, \mu, \sigma^2 \Sigma) + \log \Gamma_D(\lambda^2/(2\sigma^2)) \right] \\
&\sim \sum_{k=1}^{K} \sum_{n : z_{n,k}=1} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k)
\end{aligned}$$

$$+ \sum_{k=1}^{K} \lambda^2 \log |\Sigma_k| + c + O(\sigma^2),$$

where c is a constant in z, μ, σ^2, Σ. Letting $\sigma^2 \to 0$, the righthand side becomes

$$\sum_{k=1}^{K} \sum_{n: z_{n,k}=1} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k) + \sum_{k=1}^{K} \lambda^2 \log |\Sigma_k| + c.$$

It is equivalent to optimize the same quantity without c.

If the Σ_k are known, they may be inputted and the objective may be optimized over the cluster means and cluster assignments. For unknown Σ_k, though, the resulting optimization problem is

$$\min_{z,\mu,\Sigma} \sum_{k=1}^{K} \left[\sum_{n: z_{n,k}=1} (x_n - \mu_k)' \Sigma_k^{-1} (x_n - \mu_k) + \lambda^2 \log |\Sigma_k| \right].$$

That is, the squared Euclidean distance in the classic K-means objective function has been replaced with a Mahalanobis distance, and we have added a penalty term on the size of the Σ_k matrices (with λ^2 modulating the penalty as in previous examples). This objective is reminiscent of that proposed by Sung and Poggio (1998).

8.G Proof of BP-means local convergence

The proof of Proposition 8.3.1 is as follows.

Proof. By construction, the first step in any iteration does not increase the objective. The second step starts by deleting any features that have the same index collection as an existing feature. Suppose there are m such features with indices J and we keep feature k. By setting $A_{k,\cdot} \leftarrow \sum_{j \in J} A_{j,\cdot}$, the objective is unchanged.

Next, let $\|Y\|_F = \sqrt{\mathbf{tr}(Y'Y)}$ denote the Frobenius norm of a matrix Y. Then $\|Y\|_F^2$ is a convex function. We check that $f(A) = \mathbf{tr}[(X - ZA)'(X - ZA)]$ is convex. Take $\lambda \in [0, 1]$, and let A and B be $K \times D$ matrices; then,

$$\begin{aligned}
&f(\lambda A + (1-\lambda)B) \\
&= \|Z[\lambda A + (1-\lambda)B] - X\|_F^2 \\
&= \|\lambda(ZA - X) + (1-\lambda)(ZB - X)\|_F^2 \\
&\le \lambda \|ZA - X\|_F^2 + (1-\lambda)\|ZB - X\|_F^2 \\
&= \lambda f(A) + \lambda f(B)
\end{aligned}$$

We conclude that $f(A)$ is convex.

With this result in hand, note

$$\nabla_A \mathbf{tr}[(X - ZA)'(X - ZA)] = -2Z'(X - ZA). \tag{8.16}$$

Setting the gradient to zero, we find that $A = (Z'Z)^{-1}Z'X$ solves the equation for A and therefore minimizes the objective with respect to A when $Z'Z$ is invertible, as we have already guaranteed.

Finally, since there are only finitely many feature allocations in which each data point has at most one feature unique to only that data point and no features containing identical indices (any extra such features would only increase the objective due to the penalty), the algorithm cannot visit more than this many configurations and must finish in a finite number of iterations. □

Bibliography

Adams, R. P., Z. Ghahramani, and M. I. Jordan (2010). "Tree-structured stick breaking for hierarchical data". In: *Advances in Neural Information Processing Systems* 23, pp. 19–27.

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Aldous, D. (1985). "Exchangeability and related topics". In: *Ecole d'Eté de Probabilités de Saint-Flour 1983*, pp. 1–198.

Antoniak, C.E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". In: *The Annals of Statistics*, pp. 1152–1174.

Arthur, D. and S. Vassilvitskii (2007). "k-means++: The advantages of careful seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, pp. 1027–1035.

Asuncion, A., M. Welling, P. Smyth, and Y. Teh (2009). "On smoothing and inference for topic models". In: *Uncertainty in Artificial Intelligence*.

Berkhin, P. (2006). "A survey of clustering data mining techniques". In: *Grouping Multidimensional Data*, pp. 25–71.

Bertoin, J. (1998). *Lévy Processes*. Vol. 121. Cambridge University Press.

— (2000). *Subordinators, Lévy processes with no negative jumps, and branching processes*. Centre for Mathematical Physics and Stochastics, University of Aarhus.

— (2004). "Subordinators: examples and applications". In: *Lectures on Probability Theory and Statistics*, pp. 1–91.

Blackwell, D. and J. B. MacQueen (1973). "Ferguson distributions via Pólya Urn Schemes". In: *The Annals of Statistics* 1.2, pp. 353–355.

Blei, D. M. and P. Frazier (2010). "Distance dependent Chinese restaurant processes". In: *International Conference on Machine Learning*.

Blei, D. M., T. L. Griffiths, and M. I. Jordan (2010). "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies". In: *Journal of the ACM* 57.2, p. 7.

Blei, D. M. and M. I. Jordan (2006). "Variational inference for Dirichlet process mixtures". In: *Bayesian Analysis* 1.1, pp. 121–144.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent Dirichlet allocation". In: *The Journal of Machine Learning Research* 3, pp. 993–1022.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability.* University of California Press.

Broderick, T., N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan (2013). "Streaming variational Bayes". In: *Neural Information Processing Systems.*

Broderick, T., M. I. Jordan, and J. Pitman (2012). "Beta processes, stick-breaking, and power laws". In: *Bayesian Analysis* 7.2, pp. 439–476.

— (2013). "Cluster and feature modeling from combinatorial stochastic processes". In: *Statistical Science* 28.3, pp. 289–312.

Broderick, T., B. Kulis, and M. I. Jordan (2013). "MAD-Bayes: MAP-based asymptotic derivations from Bayes". In: *International Conference on Machine Learning.*

Broderick, T., L. Mackey, J. W. Paisley, and M. I. Jordan (2014). "Combinatorial clustering and the beta negative binomial process". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Broderick, T., J. Pitman, and M. I. Jordan (2013). "Feature allocations, probability functions, and paintboxes". In: *Bayesian Analysis* 8.4, pp. 801–836.

Broderick, T., A. C. Wilson, and M. I. Jordan (2014). "Posteriors, conjugacy, and exponential families for completely random measures". In: *Submitted.*

Buntine, W. L. and A. Jakulin (2004). "Applying Discrete PCA in Data Analysis". In: *Uncertainty in Artificial Intelligence.*

Canini, K. R., L. Shi, and T. L. Griffiths (2009). "Online inference of topics with latent Dirichlet allocation". In: *Artificial Intelligence and Statistics.* Vol. 5.

Cao, L. and F. Li (2007). "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes". In: *IEEE International Conference on Computer Vision*, pp. 1–8.

Chen, H., H. Xing, and N. R. Zhang (2011). "Stochastic segmentation models for allele-specific copy number estimation with SNP-array data". In: *PLoS Computational Biology* 7, e1001060.

Chen, S. X. and J. S. Liu (1997). "Statistical applications of the Poisson-binomial and conditional Bernoulli distributions". In: *Statistica Sinica* 7, pp. 875–892.

Chernoff, H. (1952). "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations". In: *The Annals of Mathematical Statistics*, pp. 493–507.

Dalal, N. and B. Triggs (2005). "Histograms of oriented gradients for human detection". In: *Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, pp. 886–893.

Damien, P., J. Wakefield, and S. G. Walker (1999). "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables". In: *Journal of the Royal Statistical Society: Series B* 61.2, pp. 331–344.

De Finetti, B. (1931). "Funzione caratteristica di un fenomeno aleatorio". In: *Atti della R. Academia Nazionale dei Lincei, Serie 6.* 4, pp. 251–299.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* John Wiley & Sons, Inc.

Doshi, F., K. T. Miller, J. Van Gael, and Y. W. Teh (2009). "Variational inference for the Indian buffet process". In: *AISTATS.*

Dunson, D. B. and J. H. Park (2008). "Kernel stick-breaking processes". In: *Biometrika* 95.2, pp. 307–323.

Erosheva, E. A. and S. E. Fienberg (2005). "Bayesian mixed membership models for soft clustering and classification". In: *Classification–The Ubiquitous Challenge*. New York: Springer, pp. 11–26.

Escobar, M. D. (1994). "Estimating normal means with a Dirichlet process prior". In: *Journal of the American Statistical Association*, pp. 268–277.

Escobar, M. D. and M. West (1995). "Bayesian density estimation and inference using mixtures". In: *Journal of the American Statistical Association*, pp. 577–588.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications, Vol. II*. New York: John Wiley.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems". In: *The Annals of Statistics* 1.2, pp. 209–230. ISSN: 0090-5364.

Fox, E.B., E.B. Sudderth, M.I. Jordan, and A.S. Willsky (2009). "Sharing features among dynamical systems with beta processes". In: *Advances in Neural Information Processing Systems* 22, pp. 549–557.

Fraley, C. and A. E. Raftery (2002). "Model-based clustering, discriminant analysis and density estimation". In: *Journal of the American Statistical Association* 97, pp. 611–631.

Franceschetti, M., O. Dousse, D. N. C. Tse, and P. Thiran (2007). "Closing the gap in the capacity of wireless networks via percolation theory". In: *Information Theory, IEEE Transactions on* 53.3, pp. 1009–1018.

Freedman, D. (1973). "Another note on the Borel-Cantelli lemma and the strong law, with the Poisson approximation as a by-product". In: *The Annals of Probability* 1.6, pp. 910–925.

Freedman, D. A. (1965). "Bernard Friedman's urn". In: *The Annals of Mathematical Statistics* 36.3, pp. 956–970.

Geman, S. and D. Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741.

Gnedin, A., B. Hansen, and J. Pitman (2007). "Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws". In: *Probability Surveys* 4, pp. 146–171.

Gnedin, A. and J. Pitman (2006). "Exchangeable Gibbs partitions and Stirling triangles". In: *Journal of Mathematical Sciences* 138.3, pp. 5674–5685.

Goldwater, S., T. L. Griffiths, and M. Johnson (2006). "Interpolating between types and tokens by estimating power-law generators". In: *Advances in Neural Information Processing Systems, 18*. Cambridge, MA: MIT Press.

Gordon, A. D. and J. T. Henderson (1977). "An algorithm for Euclidean sum of squares classification". In: *Biometrics*, pp. 355–362.

Griffiths, T. L. and Z. Ghahramani (2006). "Infinite latent feature models and the Indian buffet process". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. Platt. Cambridge, MA: MIT Press, pp. 475–482.

Griffiths, T. L. and Z. Ghahramani (2011). "The Indian buffet process: An introduction and review". In: *Journal of Machine Learning Research* 12, pp. 1185–1224.

Griffiths, T. L. and M. Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.Suppl. 1, pp. 5228–5235.

Hagerup, T. and C. Rub (1990). "A guided tour of Chernoff bounds". In: *Information Processing Letters* 33.6, pp. 305–308.

Hansen, B. and J. Pitman (May 1998). *Prediction rules for exchangeable sequences related to species sampling.* Tech. rep. 520. University of California, Berkeley.

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects.* Orlando, FL: Academic Press. ISBN: 0123357500.

Hewitt, E. and L. J. Savage (1955). "Symmetric measures on Cartesian products". In: *Transactions of the American Mathematical Society* 80.2, pp. 470–501.

Hjort, N. L. (1990). "Nonparametric Bayes estimators based on beta processes in models for life history data". In: *Annals of Statistics* 18.3, pp. 1259–1294.

Hoffman, M. (2010). *Online inference for LDA (Python code) at*
http://www.cs.princeton.edu/~blei/downloads/onlineldavb.tar.

Hoffman, M., D. M. Blei, and F. Bach (2010). "Online learning for latent Dirichlet allocation". In: *Neural Information Processing Systems.* Vol. 23, pp. 856–864.

Hoffman, M., D. M. Blei, J. W. Paisley, and C. Wang (2013). "Stochastic variational inference". In: *Journal of Machine Learning Research* 14, pp. 1303–1347.

Honkela, A. and H. Valpola (2003). "On-line variational Bayesian learning". In: *International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 803–808.

Hoppe, F. M. (1984). "Pólya-like urns and the Ewens' sampling formula". In: *Journal of Mathematical Biology* 20.1, pp. 91–94.

Ishwaran, H. and L. F. James (2001). "Gibbs sampling methods for stick-breaking priors". In: *Journal of the American Statistical Association* 96.453, pp. 161–173.

Ishwaran, H. and M. Zarepour (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models". In: *Biometrika* 87.2, pp. 371–390.

Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". In: *Pattern Recognition Letters* 31.8, pp. 651–666.

Johnson, O., I. Kontoyiannis, and M. Madiman (2011). "Log-concavity, ultra-log-concavity, and a maximum entropy property of discrete compound Poisson measures". In: *Discrete Applied Mathematics.*

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). "An introduction to variational methods for graphical models". In: *Machine Learning* 37.2, pp. 183–233.

Kallenberg, O. (2002). *Foundations of Modern Probability.* Springer.

Kalli, M., J. E. Griffin, and S. G. Walker (2011). "Slice sampling mixture models". In: *Statistics and Computing* 21.1, pp. 93–105.

Kim, Y. (1999a). "Nonparametric Bayesian estimators for counting processes". In: *Annals of Statistics*, pp. 562–588.

Kim, Y. (1999b). “Nonparametric Bayesian estimators for counting processes”. In: *Annals of Statistics* 27.2, pp. 562–588.

Kim, Y. and J. Lee (2001). “On posterior consistency of survival models”. In: *Annals of Statistics*, pp. 666–686.

Kingman, J. F. C. (1967). “Completely random measures”. In: *Pacific Journal of Mathematics* 21.1, pp. 59–78.

— (1978). “The representation of partition structures”. In: *Journal of the London Mathematical Society* 2.2, p. 374.

— (1993). *Poisson Processes.* Oxford University Press.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. Jordan (2012). “The big data bootstrap”. In: *International Conference on Machine Learning.*

Korwar, R. M. and M. Hollander (1973). “Contributions to the theory of Dirichlet processes”. In: *The Annals of Probability*, pp. 705–711.

Kulis, B. and M. I. Jordan (2012). “Revisiting k-means: New algorithms via Bayesian nonparametrics”. In: *Proceedings of the 23rd International Conference on Machine Learning.*

LeCun, Y. and C. Cortes (1998). *The MNIST database of handwritten digits.* URL: `http://yann.lecun.com/exdb/mnist/`.

Lee, J., F. A. Quintana, P. Müller, and L. Trippa (2008). *Defining predictive probability functions for species sampling models.* Tech. rep. Working Paper.

Li, W. and A. McCallum (2006). “Pachinko allocation: DAG-structured mixture models of topic correlations”. In: *Proceedings of the 23rd International Conference on Machine Learning.* ACM, pp. 577–584.

Lijoi, A., R. H. Mena, and I. Prünster (2007). “Controlling the reinforcement in Bayesian non-parametric mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4, pp. 715–740.

Liu, J. S. (1994). “The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem”. In: *Journal of the American Statistical Association* 89, pp. 958–966.

Lowe, D. G. (2004). “Distinctive image features from scale-invariant keypoints”. In: *International Journal of Computer Vision* 60.2, pp. 91–110.

Luts, J., T. Broderick, and M. P. Wand (2012). “Real-time semiparametric regression”. In: *Journal of Computational and Graphical Statistics, to appear. Preprint arXiv:1209.3550.*

MacEachern, S. N. (1994). “Estimating normal means with a conjugate style Dirichlet process prior”. In: *Communications in Statistics-Simulation and Computation* 23.3, pp. 727–741.

MacEachern, S. N. and P. Müller (1998). “Estimating mixture of Dirichlet process models”. In: *Journal of Computational and Graphical Statistics* 7, pp. 223–238.

MacEachern, S.N. (1999). “Dependent nonparametric processes”. In: *Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

McCloskey, J. W. (1965). “A model for the distribution of individuals by species in an environment”. PhD thesis. Michigan State University.

McCullagh, P., J. Pitman, and M. Winkel (2008). “Gibbs fragmentation trees”. In: *Bernoulli* 14.4, pp. 988–1002.

McLachlan, G. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

Minka, T. P. (2001a). “A family of algorithms for approximate Bayesian inference”. PhD thesis. Massachusetts Institute of Technology.

— (2001b). “Expectation propagation for approximate Bayesian inference”. In: *Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 362–369.

Minka, T. P. and J. Lafferty (2002). “Expectation-propagation for the generative aspect model”. In: *Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 352–359.

Mitzenmacher, M. (2004). “A brief history of generative models for power law and lognormal distributions”. In: *Internet mathematics* 1.2, pp. 226–251.

Neal, R. M. (2000). “Markov chain sampling methods for Dirichlet process mixture models”. In: *Journal of Computational and Graphical Statistics* 9.2, pp. 249–265.

— (2003). “Slice sampling”. In: *Annals of Statistics*, pp. 705–741.

Newman, M. E. J. (2005). “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary physics* 46.5, pp. 323–351.

Niu, F., B. Recht, C. Ré, and S. J. Wright (2011). “Hogwild!: A lock-free approach to parallelizing stochastic gradient descent”. In: *Neural Information Processing Systems*.

Opper, M. (1998). “A Bayesian approach to on-line learning”. In: ed. by D. Saad, pp. 363–378.

Paisley, J. W., D. M. Blei, and M. I. Jordan (2012). “Stick-breaking beta processes and the Poisson process”. In: *International Conference on Artificial Intelligence and Statistics*, pp. 850–858.

Paisley, J. W., L. Carin, and D. M. Blei (2011). “Variational inference for stick-breaking beta process priors”. In: *ICML*, pp. 889–896.

Paisley, J. W., A. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin (2010). “A stick-breaking construction of the beta process”. In: *International Conference on Machine Learning*. Haifa, Israel.

Papaspiliopoulos, O. (2008). *A note on posterior sampling from Dirichlet mixture models*. Tech. rep. 8. University of Warwick, Centre for Research in Statistical Methodology.

Papaspiliopoulos, O. and G. O. Roberts (2008). “Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models”. In: *Biometrika* 95.1, pp. 169–186.

Patil, G. P. and C. Taillie (1977). “Diversity as a concept and its implications for random communities”. In: *Proceedings of the 41st Session of the International Statistical Institute*. New Delhi, pp. 497–515.

Peña, J. M., J. A. Lozano, and P. Larrañaga (1999). “An empirical comparison of four initialization methods for the K-Means algorithm”. In: *Pattern Recognition Letters* 20.10, pp. 1027–1040.

Pitman, J. (1995). “Exchangeable and partially exchangeable random partitions”. In: *Probability Theory and Related Fields* 102.2, pp. 145–158.

— (1996). “Some developments of the Blackwell-MacQueen urn scheme”. In: *Lecture Notes-Monograph Series*, pp. 245–267.

Pitman, J. (2003). “Poisson-Kingman partitions”. In: *Lecture Notes-Monograph Series* 40, pp. 1–34. ISSN: 0749-2170.

— (2006). *Combinatorial Stochastic Processes.* Vol. 1875. Lecture Notes in Mathematics. Berlin: Springer-Verlag.

Pitman, J. and M. Yor (1997). “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator”. In: *Annals of Probability* 25, pp. 855–900.

Pólya, G. (1930). “Sur quelques points de la théorie des probabilités”. In: *Annales de l’I.H.P.* 1.2, pp. 117–161.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). “Inference of population structure using multilocus genotype data”. In: *Genetics* 155.2, pp. 945–959.

Qi, Feng and Lászlò Losonczi (2010). “Bounds for the ratio of two gamma functions”. In: *Journal of Inequalities and Applications* 2010, p. 204.

Ranganath, R., C. Wang, D. M. Blei, and E. P. Xing (2013). “An adaptive learning rate for stochastic variational inference”. In: *International Conference on Machine Learning.*

Rogers, L. C. G. and D. Williams (2000). *Diffusions, Markov Processes, and Martingales. Vol. 1.* Cambridge: Cambridge University Press.

Roweis, S. (2007). *MNIST handwritten digits.* URL: `http://www.cs.nyu.edu/~roweis/data.html`.

Russell, Bryan C., William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman (2006). “Using multiple segmentations to discover objects and their extent in image collections”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1605–1614.

Seeger, M. (2005). *Expectation propagation for exponential families.* Tech. rep. University of California at Berkeley.

Sethuraman, J. (1994). “A constructive definition of Dirichlet priors”. In: *Statistica Sinica* 4.2, pp. 639–650.

Sivic, J., B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman (2005). *Discovering object categories in image collections.* Tech. rep. AIM-2005-005. MIT.

Steinley, D. (2006). “K-means clustering: A half-century synthesis”. In: *British Journal of Mathematical and Statistical Psychology* 59.1, pp. 1–34.

Sung, K. and T. Poggio (1998). “Example-based learning for view-based human face detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1, pp. 39–51.

Teh, Y. W. (2006). “A hierarchical Bayesian language model based on Pitman-Yor processes”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, pp. 985–992.

Teh, Y. W. and D. Görür (2009). “Indian buffet processes with power-law behavior”. In: *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Teh, Y. W., D. Görür, and Z. Ghahramani (2007). “Stick-breaking construction for the Indian buffet process”. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics.* Vol. 11.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). "Hierarchical Dirichlet processes". In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581. ISSN: 0162-1459.

Teh, Y. W., D. Newman, and M. Welling (2006). "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation". In: *Neural Information Processing Systems*.

Thibaux, R. J. (2008). "Nonparametric Bayesian Models for Machine Learning". PhD thesis. UC Berkeley.

Thibaux, R. J. and M. I. Jordan (2007). "Hierarchical beta processes and the Indian buffet process". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico.

Thomaz, C. E. and G. A. Giraldi (June 2010). "A new ranking method for principal components analysis and its application to face image analysis". In: *Image and Vision Computing* 28.6. We use files `http://fei.edu.br/~cet/frontalimages_spatiallynormalized_partX.zip` with `X`=`1,2`., pp. 902–913.

Titsias, M. K. (2008). "The infinite gamma-Poisson feature model". In: *NIPS*, pp. 1513–1520.

Tricomi, F. G. and A. Erdélyi (1951). "The asymptotic expansion of a ratio of gamma functions." In: *Pacific Journal of Mathematics* 1.1, pp. 133–142.

Van De Weijer, J. and C. Schmid (2006). "Coloring local feature extraction". In: *Proceedings of the European Conference on Computer Vision*, pp. 334–348.

Verbeek, Jakob J. and Bill Triggs (2007). "Region classification with Markov field aspect models". In: *Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Wainwright, M. J. and M. I. Jordan (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1.1-2, pp. 1–305.

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices". In: *Communications in Statistics—Simulation and Computation* 36.1, pp. 45–54.

Wang, C. and D. M. Blei (2013). "Variational inference in nonconjugate models". In: *The Journal of Machine Learning Research* 14.1, pp. 1005–1031.

Wang, C., J. W. Paisley, and D. M. Blei (2011). "Online variational inference for the hierarchical Dirichlet process". In: *Artificial Intelligence and Statistics*.

Wang, Y. H. (1993). "On the number of successes in independent trials". In: *Statistica Sinica* 3.2, pp. 295–312.

Watterson, G. A. (1974). "The sampling theory of selectively neutral alleles". In: *Advances in Applied Probability*, pp. 463–488.

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Tech. rep. 92-A03. Institute of Statistics and Decision Sciences Discussion Paper.

Wolpert, R. L. and K. Ickstadt (2004). "Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes". In: *Inverse Problems* 20, p. 1759.

Wood, F., C. Archambeau, J. Gasthaus, L. James, and Y. W. Teh (2009). "A stochastic memoizer for sequence data". In: *International Conference on Machine Learning*. ACM, pp. 1129–1136.

Zhou, M., L. Hannah, D. Dunson, and L. Carin (2012). “Beta-negative binomial process and Poisson factor analysis”. In: *International Conference on Artificial Intelligence and Statistics*.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.