# Handbook of Cluster Analysis (provisional top level file)

C. Hennig, M. Meila, F. Murtagh, R. Rocci (eds.)

March 13, 2015

# Contents

# Chapter 1

# Dirichlet process mixtures and nonparametric Bayesian approaches to clustering

Vinayak Rao

Department of Statistics,

Purdue University

## Abstract

Interest in nonparametric Bayesian methods has grown rapidly as statisticians and machine learning researchers work with increasingly complex datasets, and as computation grows faster and cheaper. Central to developments in this field has been the Dirichlet process. In this chapter, we introduce the Dirichlet process, and review its properties and its different representations. We describe the Dirichlet process mixture model that is fundamental to clustering problems and review some of its applications. We discuss various approaches to posterior inference, and conclude with a short review of extensions beyond the Dirichlet process.

## 1.1 Introduction

The Bayesian framework provides an elegant model-based approach to bringing prior information to statistical problems, as well as a simple calculus for inference. Prior beliefs are represented by probabilistic models of data, often structured by incorporating latent variables. Given a prior distribution and observations, the laws of probability (in particular, Bayes' rule) can be used to calculate posterior distributions over unobserved quantities, and thus to update beliefs about future observations. This provides a conceptually simple approach to handling uncertainty, dealing with missing data, and combining information from different sources. A well known issue is that complex models usually result in posterior distributions that are intractable to exact analysis; however the availability of cheap computation, coupled with the development of sophisticated Markov chain Monte Carlo sampling methods and deterministic approximation algorithms has resulted in a wide application of Bayesian methods. The field of clustering is no exception, see for example [13, 39], and the references therein.

There still remain a number of areas of active research, and in this chapter, we consider the problem of model selection. For clustering, model selection in its simplest form boils down to choosing the number of clusters underlying a dataset. Conceptually at least, the Bayesian approach can handle this easily by mixing over different models with different numbers of clusters. An additional latent variable now identifies which of the smaller submodels *actually* generated the data, and a prior belief on model complexity is specified by a probability distribution over this variable. Given observations, Bayes' rule can be used to update the posterior distribution over models.

However, with the increasing application of Bayesian models to complex datasets from fields like biostatistics, computer vision and natural language processing, modelling data as a realization of one of a set of finite-complexity parametric models is inadequate. Such datasets raise the need for prior distributions that capture the notion of a world of unbounded complexity, so that *a priori*, one expects larger datasets to require more complicated explanations (e.g. larger numbers of clusters). A mixture of finite cluster models does not really capture this idea. A realization of such a model would first sample the number of clusters, and this would remain fixed no matter how many observations are produced subsequently. Under a truly nonparametric solution, the number of clusters would be infinite, with any finite dataset 'uncovering' only a finite number of active clusters. As more observations are generated, one would expect more and more clusters to become active. The Bayesian approach of maintaining full posterior distributions (rather than fitting point estimates) allows the use of such an approach without concerns about overfitting. Such a solution is also more elegant and principled than the empirical Bayes approach of adjusting the prior over the number of clusters *after* seeing the data [8].

The Dirichlet process (DP), introduced by Ferguson in [12], is the most popular example of a nonparametric prior for clustering. Ferguson introduced the DP in a slightly different context, as a probability measure on the space of probability distributions. He noted two desiderata for such a nonparametric prior: the need for the prior to have large support on the space of probability distributions, and the need for the resulting posterior distribution to be tractable. Ferguson described these two properties as antagonistic, and the Dirichlet process reconciles them at the cost of a peculiar property: a probability measure sampled from a DP is discrete almost surely (see Figure 1.1). While there has since been considerable work on constructing priors without this limitation, the discreteness of the DP was seized upon as ideal for clustering appications, providing a prior over mixing proportions in models with an infinite number of components.

The next section defines the DP, and introduces various representations useful in applications and for studying its properties. We introduce the DP mixture model that is typically used in clustering applications, and in Section 1.3, describe some applications. We follow that with a description of various approaches to posterior inference (Section 1.4), and end by outlining some approaches to extending the DP.

## 1.2 The Dirichlet Process

Consider a positive real number $\alpha$, as well as a probability measure $G_0$ on some space $\Theta$. For instance, $\Theta$ could be the real line, and $G_0$ could be the normal distribution with mean 0 and variance $\sigma^2$. A Dirichlet process, parametrized by $G_0$ and $\alpha$, is written as $\mathrm{DP}(\alpha, G_0)$, and is a stochastic process whose realizations (which we call $G$) are probability measures on $\Theta$. $G_0$ is called the base measure, while $\alpha$ is called the concentration parameter. Often, the DP is parametrized by a single finite measure $\alpha(\cdot)$, in which case, $\alpha = \alpha(\Theta)$ and $G_0(\cdot) = \alpha(\cdot)/\alpha(\Theta)$. We will adopt the former convention.

Observe that for any probability measure $G$, and for any finite partition $(A_1, \ldots, A_n)$ of $\Theta$, the vector $(G(A_1), \ldots G(A_n))$ is a probability vector, i.e., it is a point on the $(n-1)$-simplex, with nonnegative components that add up to one. We call this vector the *projection* of $G$ onto the partition. Then the Dirichlet process is defined as follows: for any finite partition of $\Theta$, the projected probability vector has a Dirichlet distribution as specified below:

$$(G(A_1), \ldots G(A_n)) \sim \mathrm{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_n)) \tag{1.1}$$

In words, the marginal distribution induced by projecting the Dirichlet process onto a finite partition is an appropriately parametrized Dirichlet distribution (in particular, the parameter vector of the Dirichlet distribution is the projection of the base measure $G_0$ onto the partition, multiplied by $\alpha$). A technical question now arises: does there exist a stochastic process whose projections simultaneously satisfy this condition for all finite partitions of $\Theta$? To answer this, note that equation (1.1) and the properties of the Dirichlet distribution imply

$$\begin{aligned}(G(A_1), \ldots, G(A_i)+G(A_{i+1}), \ldots G(A_n)) \sim& \\ \mathrm{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_i) + \alpha G_0(A_{i+1}), \ldots, \alpha G_0(A_n))&\end{aligned} \tag{1.2}$$

This distribution of the projection on the coarsened partition $(A_1, \ldots, A_i \cup A_{i+1}, \ldots A_n)$, implied indirectly via the distribution on the finer partition $(A_1, \ldots, A_i, A_{i+1}, \ldots A_n)$ agrees with the distribution that follows directly from the definition of the DP. This consistency is sufficient to imply the existence of the DP via Kolmogorov's consistency theorem[1][22].

We shall see a more constructive definition of the DP in Section 1.2.2, however there are a number of properties that follow from this definition of the DP. First, note that from the

[1] While Kolmogorov's consistency theorem guarantees the existence of a stochastic process with specified finite marginals, a subtle issue concerns whether realizations of this stochastic process are probability measures. This actually requires mild additional assumptions on the space $\Theta$, see e.g. [14] or [33] for more detailed accounts of this.

properties of the Dirichlet distribution, for any set $A$, $E[G(A)] = G_0(A)$. This being true for all sets $A$, we have that the mean of a draw from a DP is the base probability measure $G_0$. On the other hand, $\mathrm{var}[G(A)] = \frac{G_0(A)(1-G_0(A))}{1+\alpha}$, so that $\alpha$ controls how concentrated probability mass is around the mean $G_0$. As $\alpha \to \infty$, $G(A) \to G_0(A)$ for all $A$, so that a draw from the DP equals the base measure $G_0$. We will see in Section 1.2.2 that as $\alpha$ tends to 0, $G$ approaches a single Dirac measure, whose location is drawn from $G_0$.

To appreciate why the DP has large support on the space of probability distributions, observe that any probability measure can be approximated arbitrarily well as a piecewise-constant probability measure on a sufficiently fine, finite partition of $\Theta$. Assuming the partition has $m$ components, and calling the probability vector $\mathbf{p} \equiv (p_1, \ldots, p_m)$, note that $\mathbf{p}$ lies in the support of an $m$-dimensional Dirichlet distribution $\mathrm{Dirichlet}(\gamma_1, \ldots, \gamma_m)$ only if for all $i$ such that $p_i > 0$, we have $\gamma_i > 0$. When $\mathrm{Dirichlet}(\gamma_1, \ldots, \gamma_m)$ is the projection of DP with base measure $G_0$, this is always true if the support of $G_0$ includes all of $\Theta$ (e.g. when $\Theta$ is $\mathbb{R}^n$, and $G_0$ is a multivariate normal on $\Theta$: here, for any open set $A_i$, $\gamma_i = \int_{A_i} G_0(\mathrm{d}x) > 0$). In this case, the support[2] of the DP includes *all* probability meaures on $\Theta$. More generally, the weak support of $\mathrm{DP}(\alpha, G_0)$ includes all probability measures whose support is included in the support of $G_0$. Somewhat confusingly, in spite of this large support, we will see that any probability measure sampled from a DP is discrete almost surely.

We next characterize the DP posterior by considering the following hierarchical model:

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{1.3}$$

$$\theta_i \sim G \qquad \text{for } i \text{ in } 1 \text{ to } N \tag{1.4}$$

Thus, we sample $N$ observations independently from the random probability measure $G$ (itself drawn from a DP). Let $N_A$ represent the number of elements of the sequence $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_N)$ lying in a set $A$, i.e. $N_A = \sum_{i=1}^{N} \delta_A(\theta_i)$, where $\delta_A(\cdot)$ is the indicator function for the set $A$. For some partition $(A_1, \ldots, A_m)$, the vector of counts $(N_{A_1}, \ldots, N_{A_m})$ is multinomially distributed with a Dirichlet prior on the probability vector (equation (1.1)).

[2] weak support, to be precise.

From the conjugacy of the Dirichlet distribution, we have the posterior distribution

$$(G(A_1), \dots G(A_m))|(\theta_1, \dots, \theta_N) \sim \text{Dirichlet}(\alpha G_0(A_1) + N_{A_1}, \dots, \alpha G_0(A_m) + N_{A_m}) \quad (1.5)$$

This must be true for any partition of $\Theta$, so that the posterior is a stochastic process, all of whose marginals are Dirichlet distributed. It follows that the posterior is again a DP, now with concentration parameter $\alpha + N$, and base measure $\left(\frac{\alpha}{\alpha+N} G_0 + \frac{1}{\alpha+N} \sum_{i=1}^{N} \delta_{\theta_i}\right)$:

$$G|(\theta_1, \dots, \theta_N) \sim \text{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}\right) \quad (1.6)$$

This is the conjugacy property of the DP. The fact that the posterior is again a DP allows one to integrate out the infinite-dimensional random measure $G$, obtaining a remarkable characterization of the marginal distribution over observations. We discuss this next.

### 1.2.1 The Pólya urn scheme and the Chinese restaurant process

Consider a single observation $\theta$ drawn from the DP-distributed probability measure $G$. The definition of the DP in the previous section shows that the probability that $\theta$ lies in a set $A$, marginalizing out $G(A)$, is just $E[G(A)] = G_0(A)$. Since this is true for all $A$, it follows that $\theta \sim G_0$. Since the posterior given $N$ observations is also DP-distributed (equation (1.6)), the predictive distribution of observation $N + 1$ is just the posterior base measure:

$$\theta_{N+1}|\theta_1, \dots, \theta_N \sim \frac{1}{\alpha + N}\left(\alpha G_0(\cdot) + \sum_{i=1}^{N} \delta_{\theta_i}(\cdot)\right) \quad (1.7)$$

We see that the predictive distribution of a new observation is a mixture of the DP base measure $G_0$ (weighted by the concentration parameter), and the empirical distribution of the previous observations (weighted by the number of observations). The result above allows us to sequentially generate observations from a DP-distributed random probability measure $G$, without having to explicitly represent the the infinite-dimensional variable $G$. This corresponds to a sequential process knows as a Pólya urn scheme [3]. Here, at stage $N$, we have an urn containing $N$ balls, the $i^{th}$ 'coloured' with the associated value $\theta_i$. With probability $\alpha/(\alpha+N)$, the $(N+1)^{st}$ ball is assigned a colour $\theta_{N+1}$ drawn independently from the base measure $G_0$, after which it is added to the urn. Otherwise, we uniformly pick a ball from the urn, and set $\theta_{N+1}$ equal to its colour, and return *both* balls to the urn. Observe that there is non-zero probability that multiple balls have the same colour, and that the more balls share a particular colour, the more likely a new ball will be assigned that colour. This rich-get-richer scheme is key to the clustering properties of the DP.

Let $\boldsymbol{\theta^*} \equiv (\theta^*_1, \ldots, \theta^*_{K_N})$ be the sequence of unique values in $\boldsymbol{\theta}$, $K_N$ being the number of such values. Let $\pi^N_i$ index the elements of $\boldsymbol{\theta}$ that equal $\theta^*_i$, and let $\pi^N = \{\pi^N_1, \ldots, \pi^N_{K_N}\}$. Clearly, $(\boldsymbol{\theta^*}, \pi^N)$ is an equivalent representation of $\boldsymbol{\theta}$. $\pi^N$ is a partition of the integers 1 to $N$, while $\boldsymbol{\theta^*}$ represents the parameter assigned to each element of the partition. Define $n_i$ as the size of the $i^{th}$ cluster, so that $n_i = |\pi^N_i|$. Equation (1.7) can now be rewritten as

$$\theta_{N+1} \sim \frac{1}{\alpha+N}\left(\alpha G_0 + \sum_{c=1}^{K_N} n_c \delta_{\theta^*_c}\right) \tag{1.8}$$

The preceding equation is characterized by a different metaphor called the Chinese restaurant process (CRP) [35], one that is slightly more relevant to clustering applications. Here, a partition of $N$ observations is represented by the seating arrangement of $N$ 'customers' in a 'restaurant'. All customers seated at a table form a cluster, and the dish served at that table corresponds to its associated parameter, $\theta^*$. When a new customer (observation $N+1$) enters the restaurant, with probability proportional to $\alpha$, she decides to sit by herself at a new table, ordering a dish drawn from $G_0$. Otherwise, she joins one of the existing $K_N$ tables with probability proportional to the number of customers seated there. We can write down the marginal distribution over the observations $(\theta_1, \ldots, \theta_n)$:

$$P(\pi, \boldsymbol{\theta^*}) = \left(\frac{\alpha^{K_N-1}}{[\alpha+1]^N_1}\prod_{c=1}^{K_N}(n_c-1)!\right)\left(\prod_{c=1}^{K_N} G_0(\theta^*_c)\right) \tag{1.9}$$

Here, $[x]^n_a = \prod_{i=0}^{n-1}(x+ia)$ is the rising factorial. The equation above makes it clear that the partitioning of observations into clusters, and the assignment of parameters to each cluster are independent processes. The former is controlled by the concentration parameter $\alpha$, while all clusters are assigned parameters drawn independently from the base measure. As $\alpha$ tends to infinity, each customer is assigned to her own table, and has her own parameter (agreeing with the idea that the $\theta$'s are drawn i.i.d. from a smooth probability measure). When $\alpha$ equals 0, all customers are assigned to a single cluster, showing that the random measure $G$ they were drawn from was a distribution degenerate at the cluster parameter.

Marginalizing out the cluster parameters $\boldsymbol{\theta^*}$, consider the distribution over partitions specified by the CRP:

$$P(\pi^N) = \frac{\alpha^{K_N - 1}}{[\alpha+1]_1^N} \prod_{c=1}^{K_N} (n_c - 1)! \tag{1.10}$$

This can be viewed as a distribution over partitions of the integers 1 to $N$, and is called the Ewens' sampling formula [11]. Ewens' sampling formula characterizes the clustering structure induced by the CRP (and thus the DP); for a number of properties of this distribution over partitions, see [35]. We discuss a few below.

First, observe that the probability of a partition depends only on the number of the blocks of the partition and their sizes, and is independent of the identity of the elements that constitute the partition. In other words, the probability of a partition of the integers 1 to $N$ is invariant to permutations of the numbers 1 to $N$. Thus, despite its sequential construction, the CRP defines an *exchangeable* distribution over partitions. Exchangeability has important consequences which we discuss later in Section 1.5.1.

Next, under the CRP, the probability that customer $i$ creates a new table is $\alpha/(\alpha+i-1)$. Thus $K_N$, the number of clusters that $N$ observations will be partitioned into, is distributed as the sum of $N$ independent Bernoulli variables, the $i^{th}$ having probability $\alpha/(\alpha+i-1)$. Letting $\mathcal{N}(0,1)$ be the standard normal distribution, and letting $\overset{d}{\to}$ indicate convergence in distribution, one can show that as $N \to \infty$,

$$K_N/\log(N) \to \alpha, \qquad (K_N - \alpha\log(N))/\sqrt{\alpha\log(N)} \overset{d}{\to} \mathcal{N}(0,1) \tag{1.11}$$

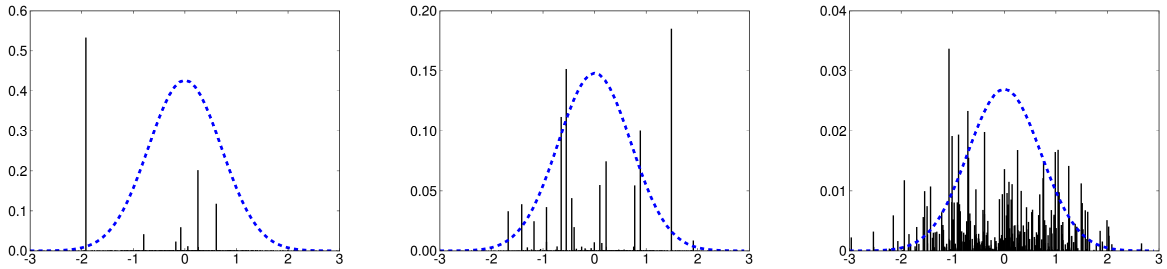

Figure 1.1: Samples from a DP whose base measure $G_0$ (the dashed blue curve) is the standard normal $\mathcal{N}(0,1)$. From left to right, the concentration parameter $\alpha$ equals 0.1, 1 and 10. $G_0$ is not normalized in the figure.

Finally, let $C_k^{(N)}$ represent the number of clusters with $k$ customers. As $N \to \infty$,

$$\left(C_1^{(N)}, C_2^{(N)}, C_3^{(N)}, \ldots\right) \overset{d}{\to} (Z_1, Z_2, Z_3, \ldots) \tag{1.12}$$

where the $Z_i, i = 1, 2, \ldots$ are independent Poisson distributed random variables with $E[Z_i] = \alpha/i$. Thus, even as the number of observations $N$ tends to infinity, one expects that under the CRP, the number of clusters with, say, just 1 observation remains $O(1)$. [35] also includes distributions over these quantities for finite $N$, but they are more complicated. Results like these are useful to understand the modelling assumptions involved in using a Dirichlet process for clustering. Despite its nonparametric nature, these assumptions can be quite strong, and an active area of research is the construction and study of more flexibile alternatives. We review a few in Section 1.5.

### 1.2.2 The stick-breaking construction

The previous section showed that observations drawn from a DP-distributed probability measure $G$ have non-zero probability of being identical. This suggests that $G$ has an atomic component. In fact $G$ is purely atomic [2], this follows from the fact *any* sample drawn from $G$ has non-zero probability of being repeated later on. Moreover, the previous section showed that the number of components is unbounded, growing as $\log(N)$, and implying that

$G$ has an infinite number of atoms. In fact, $G$ can be written as

$$G = \sum_{i=1}^{\infty} w_i \delta_{\theta_i} \tag{1.13}$$

As equation (1.9) suggests, under a DP, the sequence of weights $(w_i)$ and the sequence of locations $(\theta_i)$ are independent of each other, with the latter drawn i.i.d. from the base distribution $G_0$. The weights, which must sum to 1, are clearly not independent, however [40] provided a remarkably simple construction of these weights.

Consider a reordering the weights obtained by iteratively sampling without replacement from the set of weights $\{w_i\}$. At any stage, a weight $w_i$ selected from the infinite set of remaining weights with probability proportional to $w_i$. This constitutes a 'size-biased' reordering of the weights[3].

Consider a second sequence of weights, now obtained by repeatedly breaking a 'stick' of length 1. With an initial length equal to 1, repeatedly break off a Beta$(1, \alpha)$-distributed fraction of the remaining stick-length. Letting $V_i \sim \text{Beta}(1, \alpha)$, the sequence of weights are

$$(V_1, V_2(1 - V_1), V_3(1 - V_2)(1 - V_1), \ldots) \tag{1.14}$$

This countable sequence of weights (which adds up to 1) has what is called a GEM$(\alpha, 0)$ distribution (after Griffiths, Engen and McCloskey). Importantly, this sequence of weights has the same distribution as the size-biased reordering of the DP weights [35]. The sequence of weights, along with an infinite (and independent) sequence of $\theta_i$'s, defines a sample from a DP via equation (1.13).

The stick-breaking representation provides a simple constructive definition of the DP. Since the sequence of weights returned by the stick-breaking construction is stochastically decreasing, we can construct truncated approximations to the DP (e.g. by setting $V_k = 1$); such truncations are useful for posterior computation. The stick-breaking construction can also be generalized to construct other nonparametric priors [19].

[3] As a side note, when the weights are placed in the order their corresponding clusters were created under the CRP, one obtains a size-biased reordering.

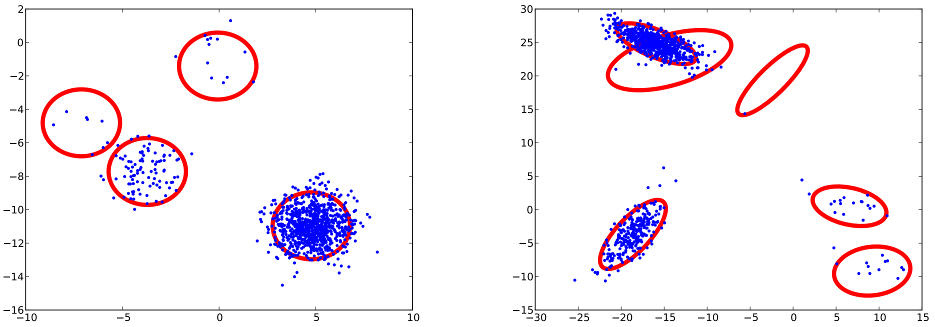

Figure 1.2: Realizations of two DP mixture models, with the DP determining cluster means (left), and means and covariance matrices (right). Ellipses are 2 standard deviations.

### 1.2.3 Dirichlet process mixtures and clustering

Since a sample $G$ drawn from a DP is discrete, it is common to smooth it by convolving with a kernel. The resulting model is a nonparametric prior over smooth probability densities and is called a Dirichlet process mixture model (DPMM) [29]. Let $f(x, \theta)$ be a nonnegative kernel on $\mathcal{X} \times \Theta$, with $\int_{\mathcal{X}} f(x, \theta)dx = 1$ for all $\theta$. The function $f$ thus defines a family of smooth probability densities on the space $\mathcal{X}$ (the observation space), indexed by elements of $\Theta$ (the parameter space). Observations then come from the following hierarchical model:

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{1.15}$$

$$\theta_i \sim G, \quad x_i \sim f(x, \theta_i) \tag{1.16}$$

In the context of clustering, the DPMM corresponds to an infinite mixture model [38]. The DP induces a clustering among observations, with observations in the $c^{th}$ cluster having parameter $\theta^*_c$. These cluster parameters are drawn i.i.d. from the base measure $G_0$, which characterizes the spread of clusters in parameter space. The cluster parameter determines properties like location, spread, and skewness of the smoothing kernel $f$.

Figure 1.2 shows the distribution of 1000 data points from two DPMMs. In both cases, the concentration parameter $\alpha$ was set to 1. Both models used Gaussian kernels (and so are DP mixtures of Gaussians), however for the first plot, all kernels had a fixed, isotropic covariance. Here, the parameter of each cluster was its mean, so that both $\Theta$ and $\mathcal{X}$ were the 2-dimensional Euclidean space $\mathbb{R}^2$. We set the base measure $G_0$ as a conjugate normal with mean zero and a standard deviation equal to 10. In the second case, the covariance of the Gaussian kernels also varied across clusters, so that $\Theta = \mathbb{R}^2 \times \mathbb{S}_2^+$ ($\mathbb{S}_+^2$ being the space of positive-definite, two-dimensional symmetric matrices). In this case, we set the base measure as the conjugate normal-inverse-Wishart distribution.

In the examples above, we set parameters to clearly demonstrate the clustering structure of DPMM. Real data is usually more ambigious with multiple explanations that are all plausible. It is in such situations that the Bayesian approach of maintaining a posterior distribution over all possible configurations is most useful. At the same time, it is also in such situations that inferences are most sensitive to aspects of the prior, and one needs to be careful about a cavalier use of conjugate priors. Often, the base-measure $G_0$ is chosen to have heavy tails, and typically, hyperpriors are placed on parameters of $G_0$. The clustering structure is sensitive to the parameter $\alpha$, and one often places priors on this as well.

The Chinese restaurant process of Section 1.2.1 to sample parameters from the DP is easily extended to sample observations from the DPMM. Now, after customer $N+1$ chooses her table (and thus her parameter $\theta_{N+1}$), she samples a value $x_{N+1}$ from the probability density $f(\cdot, \theta_{N+1})$. When the $G_0$ is conjugate to $f$, one need not even represent the table parameters $\boldsymbol{\theta^*}$, and can directly sample a new observation given the previous observations associated with that table.

Given observations from a DPMM, the posterior distribution over clusterings is much more complicated that with the DP. Why this is so is easy to see: given the set of parameters $\boldsymbol{\theta}$, we already have the clustering stucture, this is no longer the case given $\boldsymbol{X}$. Where previously the posterior over the random measure $G$ was also a Dirichlet process, it now is a *mixture* of Dirichlet processes (with a combinatorial number of components, one for each possible clustering of the observations). One thus has to resort to approximate inference techniques like MCMC or variational approximations. We describe a few such techniques later in Section 1.4, first, we concretize ideas by discussing some modelling applications.

## 1.3 Example applications

The Dirichlet process has found wide application as a model of clustered data. Often, applications involve more complicated nonparametric models extending the DP, some of which we discuss in Section 1.5. Applications include modelling text (clustering words into topics [45]), modelling images (clustering pixels into segments [41]), genetics (clustering haplotypes [50]), biostatistics (clustering functional response trajectories of patients [1]), neuroscience (clustering spikes [49]), as well as fields like cognitive science [16] and econometrics [15]. Even in density modelling applications, using a DPMM as a prior involves an implicit clustering of observations, this can be useful in interpreting results. Below, we look at three examples.

### 1.3.1 Clustering microarray expression data

A microarray experiment returns expression levels $y_{rgt}$ of a group of genes $g \in \{1, \ldots, G\}$ across treatment conditions $t \in \{1, \ldots, T\}$ over repetitions $r \in \{1, \ldots, R_t\}$. In [9], the expression levels over repetitions are modelled as i.i.d. draws from a Gaussian with mean and variance determined by the gene and treatment: $y_{rgt} \sim \mathcal{N}(\mu_g + \tau_{gt}, \lambda_g^{-1})$. Here $(\mu_g, \lambda_g)$ are gene-specific mean and precision parameters, while $\tau_{gt}$ represent gene-specific treatment effects. To account for the highly correlated nature of this data, genes are clustered as co-regulated genes, with elements in the same cluster having the same parameters $(\mu^*, \tau^*, \lambda^*)$. In [9], this clustering is modelled using a Dirichlet process, with a convenient conjugate base distribution $G_0(\mu^*, \tau^*, \lambda^*)$ governing the distribution of parameters across clusters. Besides allowing for uncertainty in the number of clusters and the cluster assignments, [9] show how such a model-based approach allows one to deal with nuisance parameters: in their situation, they were not interested in the effects of the gene-specific means $\mu_g$. The model (including all hyperparameters) was fitted using an MCMC algorithm, and the authors were able to demonstrate superiority over a number of other clustering methods.

### 1.3.2 Bayesian Haplotype inference

Most differences in the genomes of two individuals are variations in single nucleotides at specific sites in the DNA sequence. These variations are called single nucleotide polymorphisms (SNPs), and in most cases, individuals in the population have one of two base-pairs at any SNP site. This results in one of two *alleles* at each SNP site, labelled either '0' or '1'. A sequence of contiguous alleles in a local region of a single chromosome of an individual is called a haplotype. Animals like human beings are diploid with two chromosomes, and the *genotype* is made up of two haplotypes. Importantly, with present day assaying technology, the genotype obtained is usually *unphased*, and does not indicate which haplotype each of the two base pairs at each site belongs to. Thus, the genotype is a string of '00's,'11's and '01's, with the last pair ambiguious about the underlying haplotypes. Given a collection of length $M$ genotypes from a population, [50] consider the problem of disentangling the two binary strings (haplotypes) that compose each unphased genotype. To allow statistical sharing across individuals, they treat each haplotype as belonging to one of an infinite number of clusters, with weights assigned via a Dirichlet process. The parameter of each cluster is drawn from a product of $M$ independent Bernoulli random variables (this is the base measure $G_0$), and can be thought of as the cluster prototype. The individual haplotypes in the cluster differ from the cluster prototype at any location with some small probability of mutation (this forms the kernel $f$). The two haplotypes of each individual are two independent samples from the resulting DP mixture model. Given the observed set of genotypes, the authors infer the latent variables in the model by running an MCMC sampling algorithm.

### 1.3.3 Spike sorting

Given a sequence of action potentials or spikes recorded by an electrode implanted in an animal, a neuroscientist has to contend with the fact that a single recording includes activity from multiple neurons. Spike sorting is essentially the process of clustering these action potentials, assigning each spike in a spike train to an appropriate neuron. The number of neurons is unknown, and [49] assume an infinite number using a Dirichlet process. Each neuron has its own stereotypical spike shape, which can be modelled using a Gaussian process as the base measure ([49], as is typical, model lower-dimensional PCA projections of the spike shape as multivariate Gaussians). Adding a Gaussian smoothing kernel $f$ to model measurement noise, the resulting model is a DP mixture of Gaussians. Again, given observations, one can perform posterior inference using techniques we describe next.

## 1.4 Posterior inference

Given observations $\boldsymbol{X}$ from a Dirichlet process mixture model, inference involves characterizing the posterior distribution over cluster assignments of data points, as well as the cluster parameters. Implicit in this distribution are quantities like the distribution over the number of clusters underlying the observed dataset, the probability that two observations belong to the same cluster and so on. A straightforward expression for this distribution follows from the CRP representation (recall that $\pi^N$ is a partition of the integers 1 to $N$):

$$p(\pi^N, \{\theta_1^*, \ldots, \theta_{|\pi^N|}^*\}|\boldsymbol{X}) \propto \left( \alpha^{|\pi^N|-1} \prod_{i=1}^{|\pi^N|} (|\pi_i^N| - 1)! \prod_{i=1}^{|\pi^N|} G_0(\theta_i^*) \right) \prod_{i=1}^{N} F(x_i|\theta_{c_i}^*) \tag{1.17}$$

Unfortunately, hidden in the expression about is an intractible normalization constant which involves summing over the combinatorial number of partitions of $N$ data points. This makes calculating exact posterior expectations computationally intractable even when all distribution are chosen to be conjugate, and one has to resort to approximate inference techniques. Below, we look at two dominant approaches, sampling and deterministic approximations.

### 1.4.1 Markov chain Monte Carlo

By far the most widespread approach to posterior inference for DP mixture models is Markov chain Monte Carlo (MCMC). The idea here is to set up a Markov chain whose state at any iteration instantiates all latent variables of interest. At each iteration, the state of the chain is updated via a Markov transition kernel whose stationary distribution is the desired posterior distribution over the latent variables. By running the chain for a large number of iterations, we obtain a sequence of samples of the unobserved variables whose empirical distribution converges to the posterior distribution. After discarding initial 'burn-in' samples corrupted by the arbitrary initialization of the chain, the remaining samples can be used to calculate posterior expectations and to make predictions. Typical quantities of interest include the probability two observations are co-clustered (often represented by a co-occurence matrix), the posterior over the number and sizes of clusters, and the posterior over cluster parameters. Note that each MCMC sample describes a hard clustering of the observations, and very often a single sample is used to obtain a 'typical' clustering of observations.

The different representations of the DP outlined in the earlier sections can be exploited to construct different samplers with different properties. The simplest class of samplers are *marginal* samplers that integrate out the infinite dimensional probability measure $G$, and directly represent the partition structure of the data [10]. Both the Chinese restaurant process and the Pólya urn scheme provide such representations, moreover, they provide straightforward cluster assignment rules for the last observation. Exploiting the exchangeability of these processes, one cycles through all observations, and treating each as the last, assigns it to a cluster given the cluster assignments of all other observations, and the cluster parameters.

Algorithm 1.4.1 outlines the steps involved, the most important being step 4. In words, the probability an observation is assigned to an existing cluster is proportional to the number of the remaining observations assigned to that cluster, and to how well the cluster parameter explains the observation value. The observation is assigned to a new cluster with probability proportional to the product of $\alpha$ and its marginal probability integrating $\theta$ out. When the base measure $G_0$ is conjugate, the latter integration is easy. The nonconjugate case needs some care, and we refer the reader to [32] for an authoritative account of marginal Gibbs sampling methods for the DP.

```
Algorithm  An iteration of MCMC using the CRP representation
Input:   The observations $(x_1, \ldots, x_N)$
         A partition $\pi$ of the observations, and the cluster parameters $\boldsymbol{\theta}^*$
Output:  A new partition $\tilde{\pi}$, and new cluster parameters $\tilde{\boldsymbol{\theta}}^*$
1. for $i$ from 1 to $N$ do:
2.     Discard cluster assignment $c_i$ of observation $i$, and call the updated partition $\tilde{\pi}^{\setminus i}$.
3.     If $i$ belonged to its own cluster, discard $\theta_{c_i}$ from $\boldsymbol{\theta}^*$.
4.     Update $\tilde{\pi}$ from $\tilde{\pi}^{\setminus i}$ by assigning $i$ to a cluster with probability

       $$p(c_i = k|\tilde{\pi}^{\setminus i}, \boldsymbol{\theta}^*) \propto \begin{cases} |\tilde{\pi}_k^{\setminus i}| f(x_i, \theta_k^*) & k \leq |\tilde{\pi}^{\setminus i}| \\ \alpha \int_\Theta f(x_i, \theta) G_0(\mathrm{d}\theta) & k = |\tilde{\pi}^{\setminus i}| + 1 \end{cases}$$

5. If we assign $i$ to a new cluster, sample a cluster parameter from

       $$p(\tilde{\theta}_{c_i}^*|\tilde{\pi}, x_1, \ldots, x_N) \propto G_0(\tilde{\theta}_{c_i}^*) f(x_i, \tilde{\theta}_{c_i}^*)$$

6. end for
7. Resample new cluster parameters $\tilde{\theta}_c^*$ (with $c \in \{1, \ldots, \tilde{\pi}|\}$) from the posterior

       $$p(\tilde{\theta}_c^*|\tilde{\pi}, x_1, \ldots, x_N) \propto G_0(\tilde{\theta}_c^*) \prod_{i \text{ s.t. } c_i = c} f(x_i, \tilde{\theta}_c^*)$$
```

While the Gibbs sampler described above is intuitive and easy to implement, it makes very local moves, making it difficult for the chain to explore multiple posterior modes. Consider a fairly common situation where two clusterings are equally plausible under the posterior, one where a set of observations are all allocated to a single cluster, and one where they are split into two nearby clusters. One would hope that the MCMC sampler spends an equal amount of time in both configurations, moving from one to the other. However splitting a cluster into two requires the Gibbs sampler to sequentially detach observations from a large cluster, and assign them to a new cluster. The rich-get-richer property of the CRP, which encourages parsimony by penalizing fragmented clusterings, makes these intermediate states unlikely, and results in a low probability valley separating these two states. This can lead to the sampler mixing poorly. To overcome this, one needs to interleave the Gibbs updates with more complex Metropolis-Hastings proposals that attempt to split or merge clusters. Marginal samplers that attempt more global moves include [20] and [27].

A second class of samplers are called *blocked* or *conditional* Gibbs samplers, and explicitly represent the latent mixing probability measure $G$. Since the observations are drawn i.i.d. from $G$, conditioned on it, the assignment of observations to the components of $G$ can be jointly updated. Thus, unlike the marginal Gibbs sampler, conditioned on $G$, the new partition structure of the observations is independent of the old. A complication is that the measure $G$ is infinite-dimensional, and cannot be represented exactly in a computer simulation. A common approach is to follow [19] and maintain an approximation to $G$ by truncating the the stick-breaking construction to a finite number of components.

Since the DP weights are stochastically ordered under the stick-breaking construction, one would expect only a small error for a sufficiently large truncation. In fact, [19] show that if the stick-breaking process is truncated after $K$ steps, then the error decreases exponentially with $K$. Letting $N$ be the number of observations, and $\|G - \tilde{G}_K\|_1 = \int_\Theta |G(\mathrm{d}\theta) - \tilde{G}_K(\mathrm{d}\theta)|$ be the $L_1$-distance between $G$ and its truncated version $\tilde{G}_K$, we have

$$\|G - \tilde{G}_K\|_1 \sim 4N \exp\left(-(K-1)\alpha\right) \tag{1.18}$$

As [19] point out, for $N = 150$ and $\alpha = 5$, a truncation level of $K = 150$ results in an error bound of $4.57 \times 10^{-8}$, making it effectively indistinguishable from the true model. Algorithm 1.4.1 outlines a conditional sampler with truncation.

```
Algorithm  AN ITERATION OF MCMC USING THE STICK-BREAKING REPRESENTATION
  Input:   The observations $(x_1, \dots, x_N)$, and a truncation level $K$
           Stick-breaking proportions $\boldsymbol{V} = (V_1, \dots, V_{K-1})$ (with $V_K = 1$)
           Component parameters $\boldsymbol{\theta^*} = (\theta_1, \dots, \theta_K)$
           Component indicator variables $\boldsymbol{Z} = (z_1, \dots, z_N)$
  Output:  New values of $\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{\theta}}$, and $\tilde{\boldsymbol{Z}}$
  1. Sample new component assignments $\tilde{z}_i$, (with $i \in \{1, \dots, N\}$) with
       $$p(\tilde{z}_i = k) \propto V_k \prod_{j=1}^{k-1} (1 - V_j) f(x_i, \theta_k), \quad k \leq K$$
  2. Resample new component parameters $\tilde{\theta}_k^*$ (with $c \in \{1, \dots, K\}$) from the posterior
       $$p(\tilde{\theta}_k^* | \tilde{\pi}, x_1, \dots, x_N) \propto G_0(\tilde{\theta}_k^*) \prod_{i \text{ s.t. } \tilde{z}_i = k} f(x_i, \tilde{\theta}_k^*)$$
  3. Resample new stick-breaking proportions $\tilde{V}_k$ (with $k \in \{1, \dots, K-1\}$) with
       $$\tilde{V}_k \sim \text{Beta}(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j), \quad \text{with } m_k = \sum_{j=1}^{N} \delta_k(\tilde{z}_j)$$
```

Often, there are situations when one does not wish to introduce a truncation error. This is particularly true when the infinite-dimensional $G$ has some prior other than the Dirichlet process, leading to error bounds that decay much more slowly. A solution is to have a random truncation level, where the data is allowed to guide the truncation level. See [46] or [34] for descriptions of such samplers; these remain asymptotically unbiased despite working with finite truncations of infinite-dimensional objects.

### 1.4.2 Variational inference

The idea behind variational inference is to approximate the intractable posterior with a simpler distribution, and use this to approximate quantities like cluster parameters and assignment probabilities. Variational methods have the added advantage of providing approximations to the marginal likelihood of the observed data. For concreteness, we assume the smoothing kernel in the DPMM belongs to an exponential family distribution parametrized by $\theta^*$: $f(x, \theta^*) = h(x) \exp\left(x^T \theta^* - a(\theta^*)\right)$ ($a(\theta^*)$ is the log normalization constant). We also assume the base-measure $G_0$ belongs to the conjugate exponential family with sufficient statistics $(\theta^*, -a(\theta^*))$ and natural parameters $(\lambda_1, \lambda_2)$: $G_0(\theta^*) \propto \exp((\theta^*)^T \lambda_1 - a(\theta^*)\lambda_2)$.

Recall the DPMM posterior is a complicated joint distribution over variables like a probability measure with infinite atoms as well as indicator variables assigning observations to atoms. These variables are dependent, for instance the distribution over atom locations depends on the assigned observations, and the weights depend on the locations of the atoms and assigned observations. In the mean-field approximation of [5], these dependencies are discarded, and the posterior is approximated as a product of two independent distributions, one over probability measures, and the other over cluster assignments. The distribution over probability measures is restricted to measures with $K$ atoms ($K$ is a parameter of the algorithm) of the form

$$\tilde{G} = \sum_{i=1}^{K} w_i \delta_{\theta_i^*} \tag{1.19}$$

Further under the posterior approximation, the locations and the weights are assumed independent. The posterior over the $i^{th}$ location $\theta_i^*$ is approximated as a member of the same exponential family distribution as the base-measure, with natural parameters $(\tau_{i1}, \tau_{i2})$. Write it as $q(\cdot|\tau_i)$. The set of weights $(w_1, \ldots, w_K)$ are distributed according a generalized truncated stick-breaking construction, with the $i^{th}$ stick-breaking proportion drawn from a $\text{Beta}(\gamma_{i1}, \gamma_{i2})$ distribution. In equations,

$$w_i = V_i \prod_{j<i} (1 - V_j); \quad V_K = 1, \quad V_i \sim \text{Beta}(\gamma_{i1}, \gamma_{i2}), \quad i < K; \qquad \theta_i^* \sim q(\cdot|\tau_i) \tag{1.20}$$

Finally, the posterior assignment probability of the $i^{th}$ observation is independent of the other quantities, and is specified by a $K$-dimensional probability vector $\boldsymbol{\phi}_i$. The set of parameters $\tau_i, \gamma_{i1}, \gamma_{i2}, \phi_{ij}$ specify the approximate posterior distribution, and one optimizes these to minimize the Kullback-Leibler divergence from the true posterior.

Algorithm An iteration of the variational Bayes algorithm of [5]

Input: The observations $(x_1, \dots, x_N)$, and a truncation level $K$

Stick-breaking proportions $\boldsymbol{V} = (V_1, \dots, V_{K-1})$ (with $V_K = 1$)

Component parameters $\boldsymbol{\theta^*} = (\theta_1, \dots, \theta_K)$

Cluster assignment probabilities $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N)$

Output: New values of $\boldsymbol{V}, \boldsymbol{\theta}$, and $\boldsymbol{\Phi}$

1. Update the Beta parameters for the stick-breaking proportions $\boldsymbol{V}$ as

$$\gamma_{k1} = 1 + \sum_{i=1}^{N} \phi_{ik}, \quad \gamma_{k2} = 1 + \sum_{i=1}^{N} \sum_{j=k+1}^{K} \phi_{ij}, \quad k \in \{1, \dots, N\}$$

2. Update $\boldsymbol{\tau}$

$$\tau_{k1} = \lambda_1 + \sum_{i=1}^{N} \phi_{ik} x_i, \quad \tau_{k1} = \lambda_2 + \sum_{i=1}^{N} \phi_{ik} \quad \{k = 1, \dots, K\}$$

3. Update the component assignment probability vectors $\boldsymbol{\phi}_i$ as

$$\phi_{ik} \propto \exp(S_i), \quad S_i = E_q[\log V_i] + \sum_{j=1}^{i-1} E_q[\log(1 - V_i)] + E_q[(\theta_i^*)^T X_i] - E_q[a(\theta_t^*)]$$

Algorithm 1.4.2 from [5] describes an iteration of a coordinate-descent algorithm reducing the KL-divergence to a local minimum. Variational algorithms are fast and relatively simpler to debug than the stochastic MCMC algorithms. They however introduce bias that is hard to quantify, and are not as modular as MCMC which is quite easily extended to more complex hierarchical models. Other deterministic inference schemes include [26, 25, 31, 47].

### 1.4.3 Comments on posterior computation

In general, beyond a little additional bookkeeping, these algorithms for posterior inference are not much less efficient than those for the corresponding finite mixture models. For MCMC sampling algorithms (especially without truncation), the number of clusters will vary from iteration to iteration, and a balance must be struck between reallocating/deallocating memory for the various data structures, and maintaining large and fragmented data structures.

As far as MCMC sampling is concerned, marginal samplers are generally acknowledged to have better mixing properties, though they usually require some form of split-merge to get out of local optima. Conditional samplers can suffer from correlations between cluster assignments and Dirichlet process weights, though the fact that observations can be assigned to clusters independently offers promise for parallelized algorithms. Samplers are typically run for 5000 to 10000 iterations, and mixing is assessed using standard MCMC diagnostics [7] on statistics like the number of clusters, the probability two observations are co-clustered or the size of the cluster an observation is assigned to.

While variational algorithms offer a number of advantages, these often get trapped in local optima and might require multiple reruns. Additionally, the mean-field nature of these algorithms results in updates that are no longer sparse: unlike an MCMC update which conditions on the cluster assignments of a set of observations, a variational update typically depends on the probabilities of each observation being assigned to all clusters. Thus, even if a variational algorithm requires fewer iterations than an MCMC sampler, each update can require more computation that an MCMC update.

## 1.5 Extensions

### 1.5.1 Exchangeability and consistency

Our concern in this chapter was the study of Bayesian nonparametric approaches to clustering. Abstractly, this can viewed as the study of flexible probability distributions over partitions of integers. Via the Dirichlet process, we arrived at the Ewens' sampling formula (equation (1.10)) giving probabilities of partitions of the integers 1 to $N$. Though constructed sequentially via the Chinese restaurant process, we saw that the resulting probability is *exchangeable*: it is independent of the order in which the customers arrive, and is thus invariant to permutations of the integers 1 to $N$. Exchangeability is important in many clustering applications where we do not want the order of observations to affect inferences.

The Ewens' sampling formula also has a consistency (or projectivity) property: the probability of $\pi^N$ is equal to the sum of the probabilities of all partitions $\pi^{N+1}$ of 1 to $N+1$ that are consistent with $\pi^N$. Consistency follows directly from the sequential construction of the CRP and is an important property as well: we do not want observations that were not seen to affect our inferences, these should remain irrelevant. A sequence of consistent partitions for all natural numbers implies a distribution over partitions of the natural numbers $\mathbb{N}$ [35], and when each finite distribution is exchangeable as well, the resulting distribution is called an *infinitely exchangeable partition function* (EPPF).

There is another way to see why Ewens' sampling formula is infinitely exchangeable: this is a consequence of the fact that the observations are drawn i.i.d. from the DP-distributed probability measure $G$. Conditioned on $G$, the order of the observations is irrelevant, and will remain so with $G$ integrated out. Consistency is an easy consequence of the i.i.d. construction as well. According to de Finetti's theorem [22], for *any* sequence of infinitely exchangeable observations, there exists a latent variable (possible infinite dimensional), conditioned on which, the observations are i.i.d. For the CRP, this latent variable is the DP-distributed probability $G$. Drawing observations i.i.d. from $G$ induces a partition of $\mathbb{N}$ with observations with the same value in the same cluster. Treating these values as colours drawn i.i.d. from a 'paintbox' $G$, Kingman's paintbox construction [24] specializes de Finetti's theorem to infinitely exchangeable partitions: any infinite EPPF has a mixture of paintboxes respresentation, and can be induced by sampling from a random probability measure $G$.

Even if one is only interested in the clustering of observations, a consistent, exchangeable distribution over clusterings corresponds to an underlying random measure $G$. This viewpoint can facilitate the study of asymptotics of clustering processes, the development of

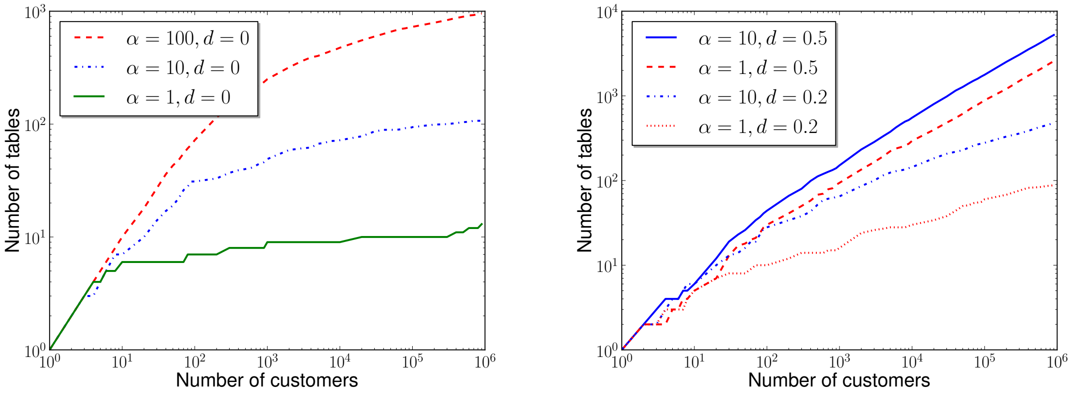

Figure 1.3: Number of clusters vs number of observations for the DP (left) and the Pitman-Yor process (right)

efficient inference techniques, the construction of new clustering models as well as extensions to more structured data where exchangeability is relaxed. We consider two extensions below.

### 1.5.2 Pitman-Yor processes

Under the CRP, the number of clusters grows logarithmically with the number of observations. Conditioned on the the number of clusters, the distribution over partitions is independent of the concentration parameter, suggesting a somewhat limited control in the prior specification. The Pitman-Yor process [36] (also called the two-parameter Poisson-Dirichlet process) is a popular extension of the DP that remedies some of these limitations. Like the DP, this too is a prior over discrete probability measures, and is parametrized as $\mathcal{PY}(\alpha, d, G_0)$. The extra parameter $d$, called the discount parameter, takes values in $[0, 1)$, while $\alpha$ (the concentration parameter) satisfies $\alpha > -d$. The Pitman-Yor process also has a stick-breaking construction; in this case, rather than all stick-breaking proportions being i.i.d. distributed, the $i^{th}$ breaking proportion $V_i$ has a $\text{Beta}(1-d, \alpha+id)$ distribution. Again, the random measure can be integrated out, resulting in a sequential clustering process that generalizes the Chinese restaurant process (now called the two-parameter CRP). Here, when the $(N+1)^{st}$ customer enters the restaurant, she joins a table with $n_c$ customers with probability proportional to $n_c - d$. On the other hand, she creates a new table with probability proportional to $\alpha + K_N d$, where as before, $K_N$ is number of tables. When the discount parameter equals 0, the Pitman-Yor process reduces to the Dirichlet process. Setting $d$ greater than 0 allows the probability of a new cluster to increase with the number of existing clusters, and results in a power-law behaviour:

$$\frac{K_N}{N^d} \to S_d \tag{1.21}$$

Here $S_d$ is a strictly positive random variable (having the so-called polynomially-tilted Mittag-Leffler distribution [35]). Power law behaviour has been observed in models of text and images, and applications of the $\mathcal{PY}$ process in these domains [42, 41] have been shown to perform significantly better than the DP (or simpler parametric models).

The two-parameter CRP representation of the $\mathcal{PY}$-process allows us to write down an EPPF that generalizes Ewens' sampling formula:

$$P(\pi^N) = \frac{[\alpha + d]_d^{K_N - 1}}{[\alpha + 1]_1^N} \prod_{c=1}^{K_N} [1 - d]_1^{n_c} \tag{1.22}$$

The EPPF above belongs to what is known as a *Gibbs-type prior*: there exist two sequences of nonnegative reals $\boldsymbol{v} = (v_0, v_1, \ldots)$ and $\boldsymbol{w} = (w_1, w_2, \ldots)$ such that the probability of a partition $\pi^N$ is

$$P(\pi^N) \propto w_{|\pi^N|} \prod_{c=1}^{|\pi^N|} v_{|\pi_c^N|} \tag{1.23}$$

This results in a simple sequential clustering process: given a partition $\pi^N$, the probability that customer $N+1$ joins cluster $c \leq |\pi^N|$ is proportional to $v_{|\pi_c^N|+1}/v_{|\pi_c^N|}$, while the probability of creating a new cluster is $v_1 w_{|\pi^N|+1}/w_{|\pi^N|}$. [35] shows that any exchangeable and consistent Gibbs partition must result from a Pitman-Yor process or an $m$-dimensional Dirichlet distribution (or various limits of these). The CRP for the finite Dirichlet distribution has $\alpha = -\kappa < 0$ and $d = m\kappa$ for some $m = 1, 2, \ldots$. Any other consistent and exchangeable distribution over partitions will result in a more complicated EPPF (and thus, for example, a more complicated Gibbs sampler). For more details, see [4]

### 1.5.3 Dependent random measures

The assumption of exchangeability is sometimes a simplification that disregards structure in data. Observations might come labelled with ordinates like time, position or category, and lumping all data points into one homogeneous collection can be undesirable. The other extreme of assigning each group an independent clustering is also a simplification, it is important to share statistical information across groups. For instance, a large cluster in one region of space in one group might *a priori* suggest a similar cluster in other groups as well. A popular extension of the DP that achieves this is the *hierarchical Dirichlet process* (HDP)[45]. Here, given a set of groups $\mathcal{T}$, any group $t \in \mathcal{T}$ has its own DP-distributed random measure $G_t$ (so that observations within each group are exchangeable). The random measures are coupled via a shared, random base measure $G$ which itself is distributed as a DP. The resulting hierarchical model corresponds to the following generative process:

$$G \sim \mathrm{DP}(\alpha_0, G_0) \tag{1.24}$$

$$G_t \sim \mathrm{DP}(\alpha, G) \qquad \text{for all } t \in \mathcal{T} \tag{1.25}$$

$$\theta_{it} \sim G_t \qquad \text{for } i \text{ in 1 to } N_t \tag{1.26}$$

In our discussion so far, we have implicitly assumed the DP base measure to be smooth. For the group-specific probability measures $G_t$ of the HDP, this is not the case; now, the base measure $G$ (which itself is DP-distributed) is purely atomic. A consequence is that the parameters of clusters across groups (which are drawn from $G$) now have nonzero probability of being identical. In fact, these parameters themselves are clustered according to a CRP, and [45] show how all random measures can be marginalized out to give a sequential clustering process they call the Chinese restaurant franchise. The clustering of parameters means that the more often a parameter is present, the more likely it is to appear in a new group. Additionally, a large cluster in one group implies (on average) large clusters (with the same parameter) in other groups.

A very popular application of the HDP is the construction of infinite topic models for document modelling. Given a fixed vocabulary of words, a *topic* is a multinomial distribution over words. A document, on the other hand, is characterized by a distribution over topics,

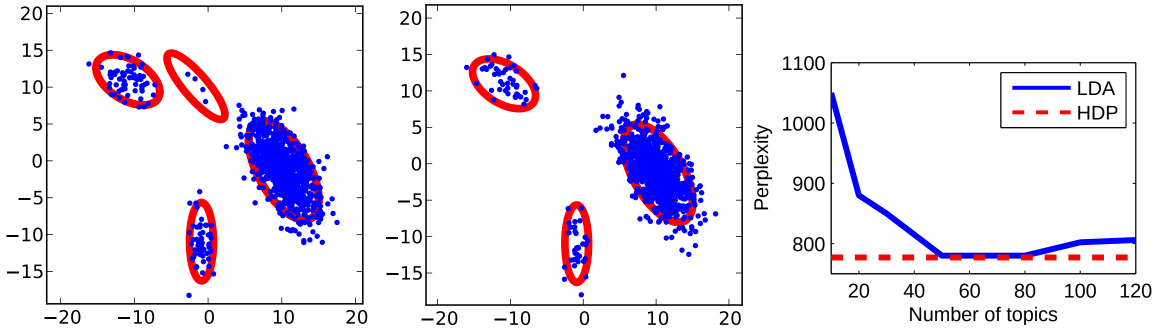

Figure 1.4: (left) and (middle) 1000 samples from an HDP with two groups. Both groups have three clusters in common, with an additional cluster in the first. (right) Perplexity on test documents for HDP and LDA with different number of topics

and the clustering problem is to infer a set of topics, as well as an assignment of each word of each document to a topic. A nonparametric approach assumes an infinite number of topics, with a random DP-distributed measure $G$ that characterizes the average distribution over topics across *all* documents. Any particular document has its own distribution over topics, centered around $G$ by drawing it from a DP with base measure $G$. The discreteness of $G$ ensures the same infinite set of topics is shared across all documents, and allows inferences from one document to propagate to other documents. Figure 1.4, adapted from [45], shows that this nonparametric model automatically achieves a performance comparable to choosing the optimal number of topics in the parametric version of the model. Here the performance measure is 'perplexity', corresponding to how surprised different models are upon seeing a held-out test document.

The HDP assumes that the groups themselves are exchangeable, again, one might wish to refine this modelling assumption. An obvious approach is to hierarchically organize the groups themselves. [42] consider $n$-level hierarchies, while [48] allow these hierarchies to be infinitely deep. Alternatively, the groups could be indexed by elements of a space $\mathcal{T}$ with some topological structure ($\mathcal{T}$ could be space or time), and one might wish to construct a *measure-valued stochastic process* $G_t$ for all $t \in \mathcal{T}$. Here, the measure $G_t$ at any point has an appropriate distribution (say, the Dirichlet process), and the similarity between measures varies gradually with $\Delta t$. Note that under the HDP, the group-specific measures $G_t$ are DP-distributed *conditioned* on $G$, this no longer is true when $G$ is marginalized out. Consequently, the clustering assumptions one makes at the group-level are different from those specified under a DP. An active area of research is the construction of dependent random probability measures with specified marginal distributions over probability measures, as well as flexible correlation structures [30].

## 1.6 Discussion

In this chapter, we introduced the Dirichlet process, and described various theoretical and practical aspects of its use in clustering applications. There is a vast literature on this topic, and a number of excellent tutorials. Among these we especially recommend [43, 44, 14]. In spite of its nonparametric nature, the modelling assumptions underlying the DP can be quite strong. While we briefly discussed a few extensions, there is much we have had to leave uncovered (see [28] for a nice overview). Many generalizations build on the stick-breaking representation or the CRP representation of the DP. Another approach extends the DP's construction as a normalized *Gamma process* [12] to construct random probability measures by normalizing *completely random measures* [23]; these are atomic measures whose weights $w_i$ are essentially all independent. The resulting class of normalized random measures [21] form a very flexible class of nonparametric priors that can be used to construct different extensions of the DP (see for example, [37]).

A different class of nonparametric priors are based on the Beta process [18], and are used to construct infinite feature models [17]. Here, rather than assigning each observation to one of a infinite number of clusters, each observation can have a finite subset of a infinite number of features. Such a distributed representation allows a more refined representation for sharing statistical information. Infinite feature models also come with notions of exchangeability and consistency. We recommend [6] for a nice overview of these ideas, and their relation to ideas discussed here.

A major challenge facing the more widespread use of Bayesian nonparametric methods is the development of techniques for efficient inference. While we described a few MCMC and variational approaches, there are many more. More recent areas of research include the development of algorithms for online inference and parallel inference.

Finally, there is scope for a more widespread application of nonparametric methods to practical problems. As our understanding of the properties of these models develops, it is important that they are applied thoughtfully. At the end, these priors represent rich and sophisticated modelling assumptions, and should be treated as such, rather than as convenient ways of bypassing the question “how many clusters?”.

## 1.7 Acknowledgements

We thank Marina Meila and Ricardo Silva for reading and suggesting improvements to earlier versions of this chapter. We also thank Jan Gasthaus and Yee Whye Teh for helpful discussions.

## References

[1] J. L. Bigelow and D. B. Dunson. Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association*, 104(485):26–36, 2009.

[2] D. Blackwell. Discreteness of Ferguson selections. *The Annals of Statistics*, 1(2):356–358, 1973.

[3] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[4] P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prunster, and M. Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? DEM Working Papers Series 054, University of Pavia, Department of Economics and Management, October 2013.

[5] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[6] T. Broderick, M. I. Jordan, and J. Pitman. Clusters and features from combinatorial stochastic processes. *pre-print*, June 2012. 1206.5862.

[7] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[8] G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):pp. 83–87, 1985.

[9] D. B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. In Kim-Anh Do, Peter Müller, and Marina Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 2006.

[10] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[11] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112, 1972.

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

[13] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? - Answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

[14] S. Ghosal. The Dirichlet process, related priors, and posterior asymptotics. In Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[15] J. E. Griffin. Inference in infinite superpositions of non-Gaussian Ornstein–Uhlenbeck Processes using Bayesian nonparametic methods. *Journal of Financial Econometrics*, 9(3):519–549, 2011.

[16] T. L. Griffiths, K. R. Canini, A. N. Sanborn, and D. J. Navarro. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 29, 2007.

[17] T. L. Griffiths, Z. Ghahramani, and P. Sollich. Bayesian nonparametric latent feature models (with discussion and rejoinder). In *Bayesian Statistics*, volume 8, 2007.

[18] N. L. Hjort. Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

[19] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[20] S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the

Dirichlet process mixture model. Technical report, Department of Statistics, University of Toronto, 2004.

[21] L. F. James, A. Lijoi, and I. Pruenster. Bayesian inference via classes of normalized random measures. ICER Working Papers - Applied Mathematics Series 5-2005, ICER - International Centre for Economic Research, April 2005.

[22] O. Kallenberg. *Foundations of Modern Probability.* Probability and its Applications. Springer-Verlag, New York, Second edition, 2002.

[23] J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[24] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society*, 37:1–22, 1975.

[25] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 20, 2007.

[26] K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational DP mixture models. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[27] P. Liang, M. I. Jordan, and B. Taskar. A permutation-augmented sampler for Dirichlet process mixture models. In *Proceedings of the International Conference on Machine Learning*, 2007.

[28] A. Lijoi and I. Pruenster. Models beyond the Dirichlet process. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice.* Cambridge University Press, 2010.

[29] A.Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

[30] S. MacEachern. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, 1999.

[31] T. P. Minka and Z. Ghahramani. Expectation propagation for infinite mixtures. Presented at NIPS2003 Workshop on Nonparametric Bayesian Methods and Infinite Models, 2003.

[32] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[33] P. Orbanz. Projective limit random probabilities on Polish spaces. *Electron. J. Stat.*, 5:1354–1373, 2011.

[34] O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

[35] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley, 2002. Lecture notes for St. Flour Summer School.

[36] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

[37] V. Rao and Y. W. Teh. Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems*, 2009.

[38] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

[39] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society*, 59(4):731–792, 1997.

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

[41] E. B. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *Advances in Neural Information Processing Systems*, pages 1585–1592, 2008.

[42] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of the 21st International Conference on Comp. Linguistics and 44th Annual Meeting of the Association for Comp. Linguistics*, pages 985–992, 2006.

[43] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[44] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

[45] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[46] S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45, 2007.

[47] L. Wang and D. B. Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2010.

[48] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh. The sequence memoizer. *Communications of the Association for Computing Machines*, 54(2):91–98, 2011.

[49] F. Wood, S. Goldwater, and M. J. Black. A non-parametric Bayesian approach to spike sorting. In *Proceedings of the IEEE Conference on Engineering in Medicine and Biologicial Systems*, volume 28, 2006.

[50] E. P. Xing, M. I. Jordan, and R. Sharan. Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14:267–284, 2007.