Dirichlet Processes and the Dirichlet Process Mixture

Brian Azizi

Cavendish Laboratory, Department of Physics, J J Thomson Avenue, Cambridge. CB3 0HE

Abstract

In many clustering problems the actual number of clusters K exhibited in the underlying population is ambiguous and indeed, for many applications, we would expect the number of observed clusters to increase as we increase the size of our data set. Traditional parametric approaches to clustering, such as finite mixture models, are too restrictive to accurately capture this phenomenon. In this paper, we will discuss the Dirichlet process mixture model, a Bayesian nonparametric approach to clustering based on the Dirichlet process.

1. Introduction

In this paper, we discuss the *Dirichlet Process Mixture (DPM)* as a *non-parametric Bayesian* model for clustering. This paper is intended to be largely self-contained and can therefore be used as a tutorial. We have implemented the Dirichlet process mixture as part of this project and will discuss the details of our implementation.

We start off with a discussion on the Bayesian analysis of discrete random variables in section 2, and introduce some key concepts that we will repeatedly make use of throughout the paper. We then extend this analysis in section 3 to the nonparametric case and formally define the Dirichlet process. Section 4 discusses various representations of the Dirichlet process that give us a better understanding of its properties. Following this, in section 5 we show how Dirichlet processes can be used for clustering and density estimation via the Dirichlet process mixture model. Here, we will also derive a general inference algorithm based on Gibbs sampling. In section 6, we derive the Dirichlet process mixture for a Gaussian likelihood model using a conjugate prior and describe important details of our implementations. Finally, we briefly discuss extensions the Dirichlet process and conclude in section 7.

2. Conjugate Bayesian Analysis of Categorical Variables

This section develops a parametric Bayesian for discrete random variables. It is intended to provide background for the nonparametric models studied in later sections. We introduce a number of important concepts that will be repeatedly used throughout this paper such as the Dirichlet and categorical distributions, delta functions and the notion of conjugacy within Bayesian inference.

2.1. Discrete Random Variables and the Categorical Distribution

Consider a discrete random variable X that can take one of K possible categorical values, $X \in \{1, \dots, K\}$. The distribution of X is then fully specified by the probabilities $\pi_k = P(X = k)$ and its *probability mass function* (pmf) is given by

$$p(x \mid \pi_1, \dots, \pi_K) = \prod_{k=1}^{K} \pi_k^{\delta_{x,k}} \tag{1}$$

where we made use of the Kronecker delta defined by

$$\delta_{i,j} := \mathbb{1}\{i = j\} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

We say that X follows the *categorical distribution* [1] , denoted $X \sim \mathrm{Cat}(\pi_1, \dots, \pi_K)$. This can be viewed as a generalization of the Bernoulli distribution to more than two outcomes.

Alternatively, we can specify the distribution of X through a *(generalized) probability density function*

[1] This distribution is sometimes also referred to as the *Discrete*, the *Multinoulli* or (somewhat inaccurately) the *Multinomial* distribution.

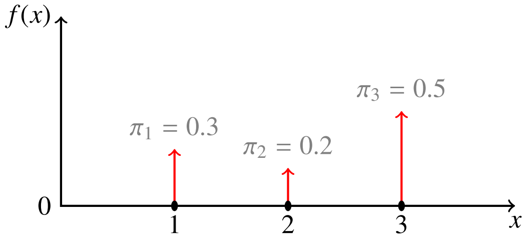

Figure 1: The probability density function of a discrete random variable that can occupy 3 possible states, $f(x) = \text{Cat}(x \,|\, \pi_1, \pi_2, \pi_3)$.

(pdf) by making use of the *Dirac delta function.* [2]

$$\text{Cat}(x \,|\, \pi_1, \ldots, \pi_K) = \sum_{k=1}^{K} \pi_k \delta(x - k) \tag{3}$$

Generalized probability density functions greatly unify the analysis of discrete and continuous random variables (as well as variables that are discrete on some parts of the probabiliy space and continuous on others). For instance, the usual definitions of expectation, variance, etc for continuous variables can be directly carried over to the case of discrete variables. In this context, points in probability space with non-zeros probability mass are referred to as *atoms* (thus, the pdf of categorical variables consists entirely of atoms). Figure 1 illustrates the pdf of X.

In order for the pmf and the pdf of the categorical distribution to be well-defined, we require that $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ since the π_k represent probabilities. In other words, we require $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ to live on the *K-dimensional probability simplex* [3] defined as

$$\Delta_K = \left\{ (\pi_1, \ldots, \pi_K) \in \mathbb{R}^K : \pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1 \right\}. \tag{4}$$

Figure 2 depicts Δ_K when $K = 3$.

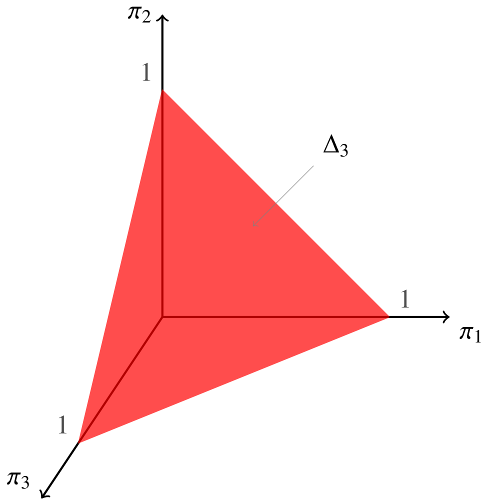

Figure 2: The 3-dimensional probability simplex Δ_3.

Now suppose we have a dataset $S = \{x^{(1)}, \ldots, x^{(N)}\}$ consisting of N independent observations of X. The corresponding likelihood function takes the form

$$\begin{aligned} p(S \,|\, \boldsymbol{\pi}) &= \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{\delta_{x^{(i)},k}} \\ &= \prod_{k=1}^{K} \pi_k^{\left(\sum_{i=1}^{N} \delta_{x^{(i)},k}\right)} = \prod_{k=1}^{K} \pi_k^{m_k} \end{aligned} \tag{5}$$

where $m_k = \sum_{i=1}^{N} \delta_{x^{(i)},k}$ represents the number of observation in which X takes the value k. The quantities m_k are the sufficient statistics of the categorical distribution.

2.2. Maximum Likelihood Inference of Categorical Variables

In the frequentist setting, we typically infer the parameter $\boldsymbol{\pi}$ through *maximum likelihood estimation*. Given dataset S, we choose the parameter $\boldsymbol{\pi}$ that maximises the log of the likelihood function $p(S \,|\, \boldsymbol{\pi})$ taking into account the constraint that $\boldsymbol{\pi} \in \Delta_K$.

To perform the constrained optimization, we use the method of Lagrange multipliers. We form the Lagrangian [4] using (5).

$$\begin{aligned} L(\boldsymbol{\pi}, \lambda) &= \log p(S \,|\, \boldsymbol{\pi}) + \lambda \left(1 - \sum_{k=1}^{K} \pi_k\right) \\ &= \sum_{k=1}^{K} m_k \log \pi_k + \lambda \left(1 - \sum_{k=1}^{K} \pi_k\right) \end{aligned} \tag{6}$$

[2] The Dirac delta can be loosely thought of as a function on $\mathbb{R}$ which is zero everywhere except at the origin where it is infinite,

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

and whose integral over the real line is unity

$$\int_{-\infty}^{+\infty} \delta(x) dx = 1.$$

Note however, that this is not a formal definition. δ can be rigorously defined using measure theory or the theory of generalized functions.

[3] Note however that, due to the sum-to-one constraint, Δ_K is in fact only a $(K-1)$-dimensional manifold embedded in K-dimensional Euclidean space. For this reason, some authors refer to Δ_K as the $(K-1)$-dimensional probability simplex (and use the notation Δ_{K-1} instead).

[4] Generally speaking, we should include the inequality constraints $\pi_k \geq 0$ in the Lagrangian and perform the optimization via the Karush-Kuhn-Tucker conditions. In our case we may safely ignore these constraints since our solution (equation 10) happens to satisfy them regardless.

Setting the derivate of L with respect to the Lagrange multiplier λ to zero, we get our sum-to-one constraint

$$\frac{\partial L}{\partial \lambda} = 1 - \sum\nolimits_{k=1}^{K} = 0. \tag{7}$$

Differentiating with respect to π_k gives

$$\frac{\partial L}{\partial \pi_k} = \frac{m_k}{\pi_k} - \lambda \tag{8}$$

and setting the derivative to zero yields

$$\begin{aligned} \lambda \pi_k &= m_k \\ \Rightarrow \lambda \sum\nolimits_k \pi_k &= \sum\nolimits_k m_k \\ \Rightarrow \lambda &= N \end{aligned} \tag{9}$$

Therefore, the maximum likelihood solution is

$$\pi_k = \frac{m_k}{N} \tag{10}$$

which is simply the fraction of observations in which $x = k$.

If our dataset is small (e.g. $N < K$), the maximum likelihood approach might (incorrectly) estimate many of the parameters to be zero. Based on this analysis, we would conclude that the probability of a new sample $x^{(N+1)}$ taking on certain values to be zero. This is sometimes referred to as the *zero count problem*.

2.3. Bayesian Inference and Conjugacy

In this section, we give a quick discussion on Bayesian inference for our categorical variable X. Along the way, we will introduce the Dirichlet distribution and demonstrate how the Bayesian framework is less prone to the zero count problem.

In a Bayesian framework, we need to specify a prior distribution $p(\boldsymbol{\pi})$ for our model parameters $\boldsymbol{\pi}$. We then use our dataset S to update the prior and obtain a posterior distribution, $p(\boldsymbol{\pi} \,|\, S)$, for $\boldsymbol{\pi}$. The update is done in accordance with *Bayes rule*:

$$p(\boldsymbol{\pi} \,|\, S) = \frac{p(S \,|\, \boldsymbol{\pi}) p(\boldsymbol{\pi})}{\int_{\pi} p(S \,|\, \boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi}}. \tag{11}$$

In general, the integral in the denominator is analytically intractable and numerical approximations can be very computationally intensive. However, for some likelihood models, we can find a parametric family of distributions such that the prior and the posterior are both members of that family. The prior is then said to be *conjugate* to the likelihood function and the parameters of the conjugate family are referred to as *hyper-parameters*. If we have conjugacy, computing the posterior is simply a matter of updating the hyper-parameters.

2.4. The Dirichlet Distribution

Looking at (5), we see that the conjugate prior to a categorical likelihood function needs to have support over the probability simplex Δ_K and have a pdf of the form

$$p(\boldsymbol{\pi}) \propto \prod\nolimits_{k=1}^{K} \pi_k^{\hat{\alpha}_k}. \tag{12}$$

The Dirichlet distribution satisfies both criteria. It has support over the probability simplex Δ_K and its pdf [5] is defined as

$$\mathrm{Dir}(\boldsymbol{\pi} \,|\, \boldsymbol{\alpha}) = \mathbb{1}\{\boldsymbol{\pi} \in \Delta_K\} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod\nolimits_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{13}$$

The quantities $(\alpha_1, \ldots, \alpha_K)$ are the *concentration parameters* of the distribution and we require them to be strictly positive real numbers.

Figure 3 shows heat plots of the dirichlet pdf over Δ_3 for different parameter values $\boldsymbol{\alpha}$. When $\alpha_k = 1$ for all values of k we get the uniform distribution. If $\alpha_k > 1$ we get a unimodal distribution where the position of the peak depends on the relative sizes of the α_k and the size of the peak depends on the magnitudes of the α_k. For $\alpha_k < 0$, the distribution is unbounded, multimodal and the probability mass concentrates on the edge of the simplex.

Define $\alpha_0 = \sum_{k=1}^{K} \alpha_k$. To work out the mean of a Dirichlet distributed variable, we first find the mean of the first component π_1,

$$\begin{aligned} \mathbb{E}\left[\pi_1\right] &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int_{S_K} \pi_1 \, \mathrm{Dir}(\boldsymbol{\pi} \,|\, \boldsymbol{\alpha}) d\boldsymbol{\pi} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma \alpha_k} \Bigg[\int_0^1 \int_0^{1-\pi_1} \cdots \int_0^{1-\sum_{k=1}^{K-1}} \\ &\pi_1^{\alpha_1} \pi_2^{\alpha_2 - 1} \ldots \pi_K^{\alpha_K - 1} d\alpha_K \ldots d\alpha_1 \Bigg] \\ &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\Gamma(\alpha_1 + 1) \prod_{k=2}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0 + 1)} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\alpha_1 \prod_{k=1}^{K} \Gamma(\alpha_k)}{\alpha_0 \Gamma(\alpha_0)} \\ &= \frac{\alpha_1}{\alpha_0} \end{aligned} \tag{14}$$

where we used the fact that the integral is the normalization constant of the $\mathrm{Dir}(\alpha_1 + 1, \alpha_2, \ldots, \alpha_K)$ pdf in the

[5] $\Gamma(t)$ is the *Gamma function* defined as

$$\Gamma(t) = \int_{-\infty}^{\infty} x^{t-1} e^{-x} dx$$

for $t \in \mathbb{R}_+$. It has the important property that $\Gamma(t+1) = t\,\Gamma(t)$ and can be regarded as a generalization of the factorial function to the positive real line since for $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

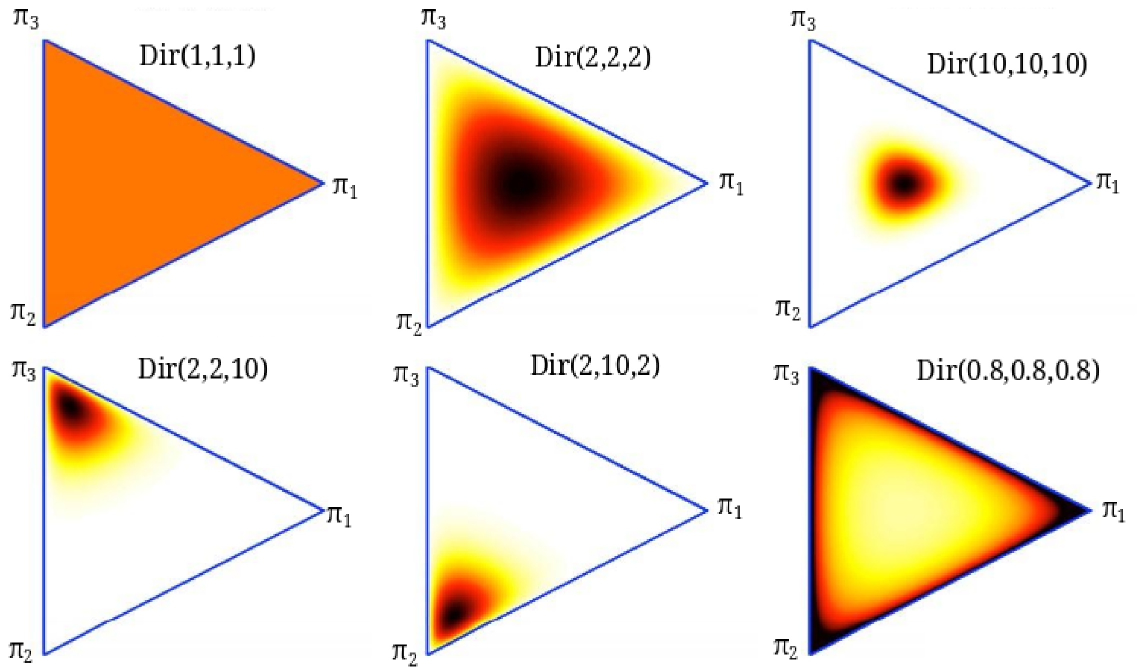

Figure 3: Examples of Dirichlet distributions on Δ_3. (Figure from [1])

third line and the property of the Gamma function that $\Gamma(t+1) = t\,\Gamma(t)$ in the fourth line. The derivation is symmetric for the remaining components of $\boldsymbol{\pi}$, thus

$$\mathbb{E}\left[\boldsymbol{\pi}\right] = \frac{\boldsymbol{\alpha}}{\alpha_0} \tag{15}$$

Often, a symmetric Dirichlet distribution of the form $\alpha_k = \alpha\ \forall k$ is used as prior. In this case, the mean of the kth component of π_k is $\mathbb{E}\left[\pi_k\right] = 1/K$. It can be shown that the variance of π_k is given by

$$\mathbb{V}\left[\pi_k\right] = \frac{K-1}{K^2(\alpha K+1)}. \tag{16}$$

In this case, α corresponds to the inverse variance of the distribution.

Previously, we had $x\,|\,\boldsymbol{\pi} \sim \mathrm{Cat}(\pi_1, \ldots, \pi_K)$. Suppose we give $\boldsymbol{\pi}$ a Dirichlet prior: $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.

Once we have a set of observations $S = \{x^{(i)}\}_{i=1}^{N}$, we are interested in how this affects our believe about the parameters $\boldsymbol{\pi}$, i.e. what is the posterior? From (11), we know that the posterior pdf, $p(\boldsymbol{\pi}\,|\,S)$ of $\boldsymbol{\pi}$ is proportional to the likelihood function $p(S\,|\,\boldsymbol{\pi})$ times the prior pdf $p(\boldsymbol{\pi})$. Thus, from (5) and (13), we get

$$\begin{aligned} p(\boldsymbol{\pi}\,|\,S) &\propto \prod\nolimits_{k=1}^{K} \pi_k^{m_k} \prod\nolimits_{k=1}^{K} \pi_k^{\alpha_k - 1} \\ &= \prod\nolimits_{k=1}^{K} \pi_k^{\alpha_k + m_k - 1} \end{aligned} \tag{17}$$

and we recognize this as the pdf of a Dirichlet distribution with parameter vector $(\alpha_1 + m_1, \ldots, \alpha_K + m_K) = \boldsymbol{\alpha} + \boldsymbol{m}$.

The *posterior predictive distribution* of our model (that is, the distribution that a new sample $x^{(N+1)}$ would have, conditional on the observed data S) is obtained by marginalising out the model parameters $\boldsymbol{\pi}$ over the posterior distribution:

$$\begin{aligned} p(X=k\,|\,S,\boldsymbol{\alpha}) &= \int p(X=k\,|\,\boldsymbol{\pi})p(\boldsymbol{\pi}\,|\,S)d\boldsymbol{\pi} \\ &= \int p(X=k\,|\,\pi_k)\left[\int p(\boldsymbol{\pi}_{-k}, \pi_k\,|\,S)d\boldsymbol{\pi}_{-k}\right]d\pi_k \\ &= \int \pi_k p(\pi_k\,|\,S)d\pi_k \\ &= \mathbb{E}\left[\pi_k\,|\,S\right] \\ &= \frac{\alpha_k + m_k}{\sum_j(\alpha_j + m_j)} = \frac{\alpha_k + m_k}{\alpha_0 + N} \end{aligned} \tag{18}$$

where we use $\boldsymbol{\pi}_{-k}$ to denote all components of $\boldsymbol{\pi}$ except π_k.

Note that this method of prediction circumvents the zero count problem. We also see that we may interpret the hyperparameters as *pseudo data*. α_0 is the effective size of the pseudo data set and α_k is the effective number of times that x takes on value k in the pseudo data. The posterior distribution of $\boldsymbol{\pi}$ takes into account both our prior believe about x (via the pseudo data) and the actual observed data.

For future reference, we also state the *prior predictive distribution* of our model. This is the distribution of a sample X before any data is observed and is obtained by marginalising out $\boldsymbol{\pi}$ over its prior:

$$p(x = k \,|\, \boldsymbol{\alpha}) = \frac{\alpha_k}{\alpha_0} \tag{19}$$

which is just the prior expected value of π_k. Thus,

$$X \sim \text{Cat}(\frac{\alpha_1}{\alpha_0}, \ldots, \frac{\alpha_K}{\alpha_0}) \tag{20}$$

For more information on the Bayesian and frequentist analysis of categorical data and the Dirichlet distribution, see [2, 3].

3. Definition of the Dirichlet Process

Having discussed the Dirichlet distribution and its role in the parametric Bayesian analysis of discrete random variables, we are now ready to explore the Dirichlet process and its role in Bayesian nonparametrics. In this section, we will first give an informal description of the Dirichlet process. We then give a formal mathematical definition of the DP and discribe its parameters.

3.1. From the Dirichlet Distribution to the Dirichlet Process

We saw that the Dirichlet distribution can be used as a prior for categorical variables when the number of possible states K is finite and known. Draws from a Dirichlet distribution give us a probability vector $\boldsymbol{\pi}$ of length K. For this reason, we can think of the Dirichlet distribution as a distribution over distributions.

The Dirichlet Process is the non-parametric extension of the Dirichlet distribution. Instead of a random vector $\boldsymbol{\pi}$, a draw from a Dirichlet process is a *random function* G with certain properties. And analogously to $\boldsymbol{\pi}$ being a valid probability distribution (since $\boldsymbol{\pi}$ lives on the probability simplex Δ_K), G is a valid probability function, a so-called *probability measure*. A probability measure is a (set) function that maps subsets from some probability space Ω onto the unit interval $[0, 1]$ and that satisfies certain properties. The most notable properties are that $G(\emptyset) = 0$, $G(\Omega) = 1$ and if $A_1, A_2, \ldots$ are pairwise disjoint subsets of Ω (meaning $A_i \cap A_j = \emptyset$ if $i \neq j$), then $G(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} G(A_i)$. [6]

In the parametric case, our categorical variable x could take on an integer between 1 and K. Thus, our probability space in that case was $\Omega = \{1, \ldots, K\}$. In (3) we defined a probability density for x. It is also possible to define a probability measure μ on $A \subset \Omega$ using the *Dirac measure* defined by [7]

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

The measure μ can then be expressed as

$$\mu(A) = \sum_{k=1}^{K} \pi_k \delta_x(A) \tag{22}$$

and is known as the *discrete measure*.

In the case of the Dirichlet process, the Ω can be any general valid probability space (not necessarily finite or even countable) and G can be any probability measure on Ω. If we have a random variable θ that takes values in Ω, then the measure G induces a probability distribution for θ. If A is measurable subset of Ω, we can define $Pr(\theta \in A) = G(A)$.

3.2. A mathematical definition of the Dirichlet process

Let $A_1, \ldots, A_K$ be a finite partition of Ω. This means $A_1 \cup \cdots \cup A_K = \Omega$ and A_k are pairwise disjoint. If G is a measure on Ω, then (A_k) is some number between 0 and 1 and $\sum_{k=1}^{K} G(A_k) = 1$. In other words, the vector $G(A_1), \ldots, G(A_K)$ lives on the probabilty simplex Δ_K. Suppose now that G is randomly drawn from a Dirichlet process, $G \sim \text{DP}$. Then $(G(A_1), \ldots, G(A_K))$ is a random vector and we can specify a distribution for it.

The Dirichlet process is implicitly defined by the requirement that $(G(A_1), \ldots, G(A_K))$ has a joint Dirichlet distribution. We write

$$G \sim \text{DP}(\alpha, H)$$

[6] There are also conditions on the domain of a measure. The collection of subsets of Ω on which a measure is defined is called a *σ-algebra* of Ω. A σ-algebra must contain the empty subset, be closed under complement and be closed under union or intersection of countably many subsets.

[7] This is the third δ-type object we have introduced in this document. Both the Kronecker delta and the Dirac delta function can be thought of as special cases of the Dirac measure.

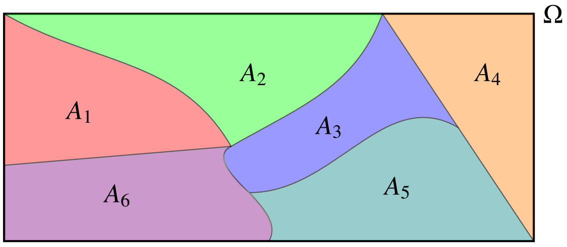

Figure 4: A finite partition of Ω.

if

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K)) \quad (23)$$

for any finite partition $A_1, \ldots, A_K$ of Ω.

A Dirichlet process is specified by a *concentration parameter* α which is a positive real number, and a *base measure* H which is a probability measure on the probability space Ω. Two describe the expectation and variance of Dirichlet process distributed measure G, consider any measurable subset A of Ω. The expectation is given by

$$\mathbb{E}[G(A)] = H(A) \quad (24)$$

and the variance is given by

$$\mathbb{V}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}. \quad (25)$$

We see that we may think of the base measure H as the mean of the Dirichlet process, whereas α plays the role of the inverse-variance.

4. Representations of the Dirichlet Process

So far, the discussion on the Dirichlet process has been very abstract. In the previous section, we gave a formal definition of Dirichlet processes, first formalized by Thomas Ferguson [4]. It is a non-constructive definition and somewhat limited for practical applications. Since then, there have been several representations of the Dirichlet process exposing its various properties. In this section, we will briefly review some of these representations and show off its conjugacy, clustering and discreteness properties. This section leans heavily on [5].

4.1. Conjugacy and the Blackwell-MacQueen Urn Scheme

Suppose G is Dirichlet process distributed, $G \sim \mathrm{DP}(\alpha, H)$. Then G is a random probability measure over Ω and we may treat it as a probability distribution over Ω (with a slight omission of the measure theoretic details). Let $\theta \in \Omega$ be a sample drawn from G: $\theta \sim G$. What is the prior predictive distribution of θ, $p(\theta)$? Furthermore, given θ, what can we say about the posterior distribution of G?

Let $(A_1, \ldots, A_K)$ be a measurable partition of Ω. We know that $(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K))$ and that $P(\theta \in A_k \,|\, G) = G(A_k)$. Using our analysis of the Dirichlet-Categorical model in section 2.4, we can express the prior predictive distribution of θ as

$$\begin{aligned} p(\theta \in A_k) &= \frac{\alpha H(A_k)}{\alpha \sum_{j=1}^{K} H(A_j)} \\ &= \frac{\alpha H(A_k)}{\alpha H(\Omega)} \\ &= H(A_k). \end{aligned} \quad (26)$$

Since this is true for any finite partition of Ω, it follows that, a priori, θ is distributed according to the base distribution H.

We also know that, due to conjugacy, the posterior of $(G(A_1), \ldots, G(A_K))$ is the Dirichlet distribution,

$$\begin{aligned} &(G(A_1), \ldots, G(A_K)) \,|\, \theta \\ &\sim \mathrm{Dir}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_K) + \delta_\theta(A_K)). \end{aligned} \quad (27)$$

Again, since this holds for any measurable partition $(A_1, \ldots, A_K)$ of Ω, it follows that the posterior distribution of G is also a Dirichlet process:

$$G \,|\, \theta \sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right). \quad (28)$$

Thus, the Dirichlet process forms a conjugate family of priors over distributions that is closed under posterior updates given observations. Furthermore, the posterior predictive distribution is the base distribution of the posterior DP.

We can extend this analysis by considering a set of independent draws $\theta^{(1)}, \ldots, \theta^{(N)} \sim G$. We saw that, if $\theta^{(1)} \,|\, G \sim G$ and $G \sim \mathrm{DP}(\alpha, H)$, then $\theta^{(1)} \sim H$ and $G \,|\, \theta^{(1)} \sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha H + \delta_{\theta^{(1)}}}{\alpha + 1}\right)$. Similarly, for the second sample, we have $\theta^{(2)} \,|\, \theta^{(1)}, G \sim G$ and $G \,|\, \theta^{(1)} \sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha H + \delta_{\theta^{(1)}}}{\alpha + 1}\right)$. Therefore, the predictive distribution of $\theta^{(2)}$ conditional on the first sample $\theta^{(1)}$ is $\frac{\alpha H + \delta_{\theta^{(1)}}}{\alpha + 1}$ and the posterior of G given both $\theta^{(1)}$ and $\theta^{(2)}$ is $G \,|\, \theta^{(1)}, \theta^{(2)} \sim \mathrm{DP}\left(\alpha + 2, \frac{\alpha H + \delta_{\theta^{(1)}} + \delta_{\theta^{(2)}}}{\alpha + 2}\right)$. After seeing the entire data set, we have the following posterior for G

$$G \,|\, \theta^{(1)}, \ldots, \theta^{(N)} \sim \mathrm{DP}\left(\alpha + N, \frac{\alpha H + \sum_{i=1}^{N} \delta_{\theta^{(i)}}}{\alpha + N}\right) \quad (29)$$

and the posterior predictive distribution of a new sample $\theta^{(N+1)}$ is

$$\theta^{(N+1)} \,|\, \theta^{(1)}, \dots, \theta^{(N)} \sim \frac{\alpha H + \sum_{i=1}^{N} \delta_{\theta^{(i)}}}{\alpha + N}. \tag{30}$$

There are several things to note about this predictive distribution. First of all, we can express it as $\frac{\alpha}{\alpha+N} H + \frac{N}{\alpha+N} \frac{\sum_{i=1}^{N} \delta_{\theta^{(i)}}}{N}$ which is a weighted average of the prior base distribution H and the *empirical distribution* $\frac{\sum_{i=1}^{N} \delta_{\theta^{(i)}}}{N}$. The weight of H is proportional to the concentration parameter α while the weight of the empirical distribution is proportional to the size of the data set N. Therefore, α can be interpreted as the strength of our prior believe. As the size N of our data set increases, the predictive distribution is dominated by the empirical distribution. For large N, the empirical distribution is a good approximation to the true underlying distribution of the data. Thus, we the Dirichlet process has a consistency property in the sense that the posterior DP converges to the true distribution of the data.

Secondly, there is a useful interpretation of the predictive distribution as a *Polya urn scheme*, due to Blackwell and MacQueen [6]. Suppose each value $\theta \in \Omega$ represents a unique colour. Starting with an empty urn, we draw a colour at random from H, $\theta^{(i)} \sim H$, paint a ball with that colour and drop it into the urn. In the $(N + 1)$th step we have N balls in our urn and we want to pick a colour for the $(N + 1)th$ ball. We either a draw a new colour from H, $\theta^{(N+1)} \sim H$ with probability $\frac{\alpha}{\alpha+N}$, or, with probability $\frac{N}{\alpha+N}$ we reach into the urn to pick a ball at random, let $\theta^{(N+1)}$ be that ball's colour and drop the ball back into the urn (i.e. $\theta^{(N+1)} \sim (\sum_{i=1}^{N} \delta_{\theta^{(i)}}/N)$). Therefore, the distribution of $\theta^{(N+1)} \,|\, \theta^{(1)}, \dots, \theta^{(N)}$ is the same as the posterior predictive distribution of the Dirichlet process.

We derived the Blackwell-MacQueen urn scheme starting from the Dirichlet process. It is also possible to start at the urn scheme with conditional probabilities defined as in (30) and prove the existence of the Dirichlet process using de Finetti's theorem (see [6]).

The Blackwell-MacQueen urn scheme gives us intuition about the conjugacy property of the Dirichlet process. It also gives us a first insight into the discreteness and clustering properties.

From the posterior predictive distribution of $\theta^{(N+1)}$, we see that

$$Pr(\theta^{(N+1)} = \theta^{(k)} \,|\, \theta^{(1)}, \dots, \theta^{(N)}) = \frac{\sum_{i=1}^{N} \delta_{\theta^{(i)},\theta^{(k)}}}{N} = \frac{N_k}{N} \tag{31}$$

where $N_k \geq 1$ is the number of times we observed $\theta^{(k)}$. Thus, there is a strictly positive probability that we observe the same value multiple times, regardless of whether H is continuous or not. We also said that $\theta^{(1)}, \dots, \theta^{(N)} \,|\, G \sim G$. It follows that the distribution G must have atoms (point-masses). Furthermore, we saw that the posterior predictive distribution will eventually be dominated by the empirical distribution. Thus, in the limit, new samples of G will almost surely take on a previously seen value. Hence, G must be entirely made up of a sum of point-masses. In other words, G is a discrete measure. In the next section, we will see even more explcitly that samples G from a Dirichlet process are discrete.

4.2. The Stick-breaking construction

We now give a constructive definition of the Dirichlet process known as the *stick-breaking construction*. This will solidify our intuition that samples from DP(α, H) are discrete measures.

Let $\boldsymbol{\pi} = \pi_1, \pi_2, \dots$ be an infinite sequence constructing in the following way: For $k = 1, 2, \dots,$ first draw a random number β_k from a beta distribution [8] with shape parameters 1 and α,

$$\beta_k \sim \text{Beta}(1, \alpha). \tag{32}$$

Then, let

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) = \beta_k (1 - \sum_{j=1}^{k-1} \pi_j). \tag{33}$$

This construction is often denoted $\boldsymbol{\pi} \sim \text{Stick}(\alpha)$ or $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$. [9] To understand the stick metaphor, consider a stick of unit length. We randomly break off a portion β_1 and set π_1 to the length of that pieve. From the remaining piece, we break off another portion β_2 and assign its length to π_2. We recursively break off pieces from the remaining stick to generate the sequence π_k (see Figure 5). The number of generated pieces increases with α, but it is possible to show that this sequence will terminate with probability one.

Next, generate a sequence of samples $\theta^{(k)} \sim H$ and let

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta^{(k)}} \tag{34}$$

[8] The Beta(u, v) distribution is a continuous distribution over the unit interval $[0, 1]$. It is defined for $u, v > 0$ and its probability density function is given by

$$\text{Beta}(x \,|\, u, v) = \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} x^{u-1} (1 - x)^{v-1}.$$

The Beta distribution is equivalent to the Dirichlet distribution when $K = 2$.

[9] This notation was introduced by [7]. GEM stands for Griffiths, Engen and McClosky.

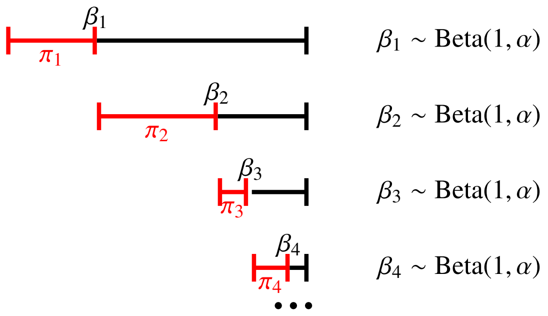
β_1
π_1
$\beta_1 \sim \text{Beta}(1,\alpha)$
β_2
π_2
$\beta_2 \sim \text{Beta}(1,\alpha)$
β_3
π_3
$\beta_3 \sim \text{Beta}(1,\alpha)$
β_4
π_4
$\beta_4 \sim \text{Beta}(1,\alpha)$
• • •

Figure 5: Stick-breaking construction of $\boldsymbol{\pi}$.

Sethuraman [8] proved that $G \sim \text{DP}(\alpha, H)$ under very general conditions (thereby also giving a more straightforward and general proof of the existence of Dirichlet processes than seen before). The stick-breaking construction shows in a very straightforward way that samples from a Dirichlet distribution are discrete measures.

4.3. Clustering property of the Dirichlet process and the Chinese Restaurant Process

In this section, we explore the clustering property of Dirichlet processes. We will show that the DP induces a distribution over partitions of integers known as the *Chinese restaurant process* (CRP). The CRP teases out the clustering property from the DP, just like the stick-breaking construction teased out the discreteness-property.

Given a set of samples $\theta^{(1)}, \dots, \theta^{(N)}$ generated via the Blackwell-MacQueen urn scheme, we used (31) to argue that there is a positive probability several samples take on the same value. Since we have repeated values, let us denote the distinct values by $\bar{\theta}_1, \dots, \bar{\theta}_K$, where $K \leq N$. We may think of these disctinct values as different clusters.

The clustering induced by the DP exhibits a so-called rich-gets-richer behaviour, in that large clusters grow faster than small clusters. To see that, consider again (31) and note that it implies that

$$Pr(\theta^{(N+1)} = \bar{\theta}_k) \propto N_k = \sum_{i=1}^{N} \mathbb{1}\{\theta^{(i)} = \bar{\theta}_k\}. \tag{35}$$

Thus, the larger the cluster size N_k, the greater the probability that a new sample joins cluster k.

Consider again our samples $\theta^{(1)}, \dots, \theta^{(N)}$ and its unique values $\bar{\theta}_1, \dots, \bar{\theta}_K$. This sample induces a partitioning of the set of integers $[N] = 1, \dots, N$ into K clusters such that i is in cluster k if $\theta^{(i)} = \bar{\theta}_k$. Since our values $\theta^{(i)}$ are randomly, the induced partitioning is also random and its distribution is known as the Chinese restaurant process.

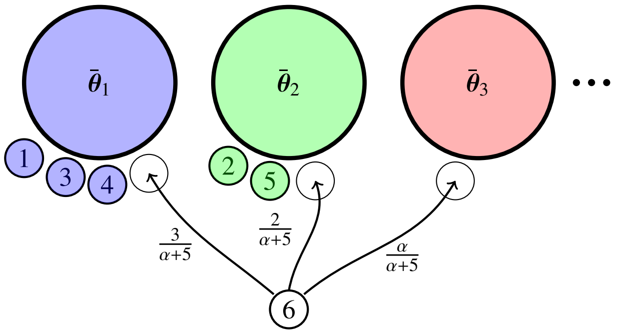
$\bar{\boldsymbol{\theta}}_1$
$\bar{\boldsymbol{\theta}}_2$
$\bar{\boldsymbol{\theta}}_3$
• • •
1
3
4
2
5
$\frac{3}{\alpha+5}$
$\frac{2}{\alpha+5}$
$\frac{\alpha}{\alpha+5}$
6

Figure 6: The Chinese restaurant process. 5 customers are currently seated across 2 tables ($K = 2$). The 6th customer is waiting to be seated. The quantities on the arrows are the probabilities with which customer 6 will be seated at the corresponding table. Currently, the induced partition over the integers $\{1, 2, 3, 4, 5\}$ is $\{\{1, 3, 4\}, \{2, 5\}\}$.

The name is due to a metaphor in which we have a Chinese restaurant with an infinite number of tables and each table can seat an infinite number of customers. At first, the restaurant is empty and the first customer is seated at the first table. The $(i+1)$th customer enters the restaurant is either seated at an already occupied table k or at a new (empty) table $K + 1$, with the following probabilities

$$\begin{aligned} \Pr(i \text{ sits at table } k) &= \frac{N_k}{i+\alpha} \\ \Pr(i \text{ sits at a new table } K+1) &= \frac{\alpha}{i+\alpha} \end{aligned} \tag{36}$$

where N_k is the current number of people sitting at table k. In the metaphor, integers correspond to customers and tables correspond to clusters. After the Nth customer is seated, the tables define a partition of $[N]$ and its distribution is the same as the one induced by the DP.

Note that the probability of opening up a new table, i.e. of creating a new cluster is proportional to the concentration parameter α. If α is large, we should expect to see many distinct values $\bar{\theta}_1, \dots, \bar{\theta}_K$. In other words, K increases with α on average and it is possible to show that the number of occupied tables, K, approaches $\alpha \log(N)$ as $N \to \infty$ in probability.

It is possible to start with CRP and get back to the Blackwell-MacQueen urn scheme (and thus the Dirichlet process). For each table k, we draw a value $\bar{\theta}_k$ from the base distribution, $\bar{\theta}_k \sim H$. We then set all samples whose index is in cluster k (i.e. who "sit" at table k) equal to that value

$$\theta^{(i)} = \bar{\theta}_{z_i} \tag{37}$$

where z_i is the table at which customer i is sitting at.

We will return to the Chinese restaurant process in the next section when we discuss Dirichlet mixtures as it builds the basis for our inference method.

5. The Dirichlet Process Mixture

So far, the discussion has been about the properties of the Dirichlet process as a stochastic object. In this section, we show how Dirichlet processes can be used for clustering and to model data via *Dirchlet process mixtures (DPM)*. We first describe the model in general terms. Next we explain how to fit the parameters of the model via Gibbs sampling and give a brief overview of alternative inference methods.

5.1. Density Estimation and Clustering using Dirichlet Processes

Suppose we have a dataset $S = \{x^{(i)}\}_{i=1}^{N}$ that are independently drawn from some unknown distribution and suppose we would like to estimate the density of this distribution. In the Bayesian nonparametric literature, the standard approach is to put a nonparametric prior distribution over the space of distributions. The aim is to have a very flexible model (since we are not limiting ourselves to some parametric family but instead considering all distributions) while also avoiding overfitting (by using Bayesian inference). The Dirichlet process provides a good choice for our prior as it has broad cover over the space of distributions while also being mathematically convenient (through conjugacy). However, we saw that draws from a DP are discrete distributions and therefore not particularly useful as a direct model for real data. To get around this problem, we convolve the DP with a smooth distribution to create a nonparametric density for our data.

Let $G \sim \mathrm{DP}(\alpha, H)$ and let $f(x|\,\theta)$ be a parametric family of smooth densities with parameters θ. For example, we may use $f(x|\,\theta) = \mathcal{N}(x|\,\mu, \Sigma)$ in which case $\theta = (\mu, \Sigma)$. The density of x is then modelled as

$$p(x) = \int f(x|\,\theta)\, G(\theta) d\theta. \tag{38}$$

This model is a mixture of distributions in which G is the mixing distribution over θ and $F(\theta)$ are the base distributions with density functions $f(x|\,\theta)$. More precisely, we model out data set S using a set of *latent parameters* $\{\theta^{(i)}\}$. The parameters are drawn independently from G and the distribution of each $x^{(i)}$ is given by $F(\theta^{(i)})$:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H) \\ \theta^{(i)}|\, G &\sim G \\ x^{(i)}|\, \theta^{(i)} &\sim F(\theta^{(i)}) \end{aligned} \tag{39}$$

We assume that, given the $\theta^{(i)}$, the observations $x^{(i)}$ are independent of each other and of G.

We already saw that multiple $\theta^{(i)}$ can take on the same value (due to the discreteness of G). Thus, as well as density estimation, we can also use this model for clustering where observations $x^{(i)}$ whose parameter $\theta^{(i)}$ take on the same value belong to the same cluster.

To make the connection to mixture models more explicit, we introduce a latent categorical variable $z^{(i)} \in \{1, 2, \ldots\}$ that tells us the cluster assignment for each observation $x^{(i)}$. The DPM model can then be expressed in the following way:

$$\begin{aligned} x^{(i)}|\, z^{(i)}, \{\bar{\theta}_1, \bar{\theta}_2, \ldots\} &\sim F(\bar{\theta}_{z^{(i)}}) \\ z^{(i)}|\, \{\pi_1, \pi_2, \ldots\} &\sim \mathrm{Cat}(\pi_1, \pi_2, \ldots) \\ \bar{\theta}_1, \bar{\theta}_2, \ldots &\sim H \\ \pi_1, \pi_2, \ldots &\sim \mathrm{GEM}(\alpha) \end{aligned} \tag{40}$$

so that

$$p(x) = \sum_{k=1}^{\infty} \pi_k f(x|\,\bar{\theta}_k) \tag{41}$$

In this formulation we see very clearly that Dirichlet process mixtures can be viewed as (countably) infinite mixture models. However, in any finite data set we can observe at most $K = N$ of the mixture components. In practice, $K << N$ and as we mentioned earlier, K tends to grow logarithmically with the sample size N.

The advantage over using a finite mixture model is that we do not need to specify the number of components K, but instead infer it from the data. However, note that an underlying assumption of the DPM is that the true number of clusters is infinite and thus, the larger our dataset the more clusters we will observe. For many applications, this assumption is realistic. But, if we know that the number of clusters in the underlying population is finite, the DPM is a misspecified model. In this case, using a finite mixture model along with a model selection tool is a more appropriate approach.

5.2. Inference in Dirichlet Process mixtures

The most popular approaches to inference in Dirichlet process mixture models are based on Markov chain Monte Carlo (MCMC) methods. The 2000 paper by

Radford Neal [9] provides an excellent review of inference methods based on Gibbs sampling using the Chinese restaurant process representation of Dirichlet processes. In particular, "algorithm 8" in his paper is still considered to be one of the best MCMC based inference methods for handling non-conjugate priors.

Since then, there have been proposals for better MCMC inference algorithm such as a blocked Gibbs sampler based on the stick-breaking representation [10], Metropolis-Hastings with larger moves [11] and sequential Monte Carlo [12].

Besides MCMC based algorithms, there have also been other approaches to inference in DP mixtures. An algorithm based on expectation propagation was derived by [13]. The first variational Bayes approximation was developed by [14].

5.3. Gibbs Sampling

Although better methods for inference in DPMs have been developed, we will focus on Gibbs sampling. The reason for this is two-fold. First, it is generaly considered to be the standard sampler for Dirichlet Process mixtures and very popular due to its simplicity. Second, it is the inference method used in the current version of our implementation.

Recall that a Gibbs sampler simulates random draws $\mathbf{\Phi} = (\Phi_1, \ldots, \Phi_D)$ from some multivariate distribution Q on Ω (with $\dim \Omega = D$) by looping over the dimensions $d \in \{1, \ldots, D\}$ and sampling Φ_d from its conditional distribution given the remaining dimensions (the so-called *full conditional* distribution of Φ_d). When sampling Φ_d from its full conditional $Q(\Phi_d | \mathbf{\Phi}_{-d} = \boldsymbol{\phi}_{-d})$, the Gibbs sampler always conditions on the most recently generated values of $\mathbf{\Phi}_{-d}$. [10] More explicitly, in the jth iteration of the Gibbs sampler, the algorithm samples as follows:

$$\begin{aligned}
\Phi_1^{(j)} &\sim Q(\Phi_1 | \Phi_2 = \phi_2^{(j-1)}, \ldots, \Phi_D = \phi_D^{(j-1)}) \\
&\vdots \\
\Phi_d^{(j)} &\sim Q(\Phi_d | \Phi_1 = \phi_1^{(j)}, \ldots, \Phi_{d-1} = \phi_{d-1}^{(j)}, \\
&\qquad \Phi_{d+1} = \phi_{d+1}^{(j-1)}, \ldots, \Phi_D = \phi_D^{(j-1)}) \\
&\vdots \\
\Phi_D^{(j)} &\sim Q(\Phi_D | \Phi_1 = \phi_1^{(j)}, \ldots, \Phi_{D-1} = \phi_{D-1}^{(j)})
\end{aligned} \tag{42}$$

In practice, it is often more efficient to randomly permute the dimensions at the start of each iteration, so that the components are sampled in a random order.

5.4. A naïve Gibbs sampler for Dirichlet Process Mixtures

Recall that in the DPM, observations are assumed to be generated according to (39) and that two sample $x^{(i)}$ and $x^{(j)}$ are in the same cluster if $\theta^{(i)} = \theta^{(j)}$. While the conditional joint distribution of $(\theta^{(1)}, \ldots, \theta^{(N)})$ given the data is complicated, the derivation of the full conditional $p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}, S)$, where $S = \{x^{(1)}, \ldots, x^{(N)}\}$, is relatively straight forward. [11] From the Blackwell-MacQueen urn scheme (30), we can derive the following conditional distribution

$$p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}) = \frac{\alpha H(\theta^{(i)})}{\alpha + N - 1} + \frac{\sum_{j \neq i} \delta_{\theta^{(j)}}(\theta^{(i)})}{\alpha + N - 1} \tag{43}$$

To see why, imagine that observation I is the last of the N observations. We are allowed to do so because we assumed that the $\theta^{(i)}$ are conditionally independent given G and therefore satisfy the condition of *exchangeability*, meaning that any order of a finite number of samples is equally likely.

Finally, we also need to condition on the observed data. For that, we use the assumption that, for $j \neq i$, $\theta^{(i)}$ is conditionally independent of observation $x^{(j)}$ given $\theta^{(j)}$. This means that $p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}, X) = p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}, x^{(i)})$. To compute that last quantity we treat $p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)})$ as a prior for $\theta^{(i)}$ and compute its posterior under the single observation $x^{(i)}$ using Bayes' rule:

$$\begin{aligned}
p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}, x^{(i)}) &= \frac{p(\theta^{(i)} | \boldsymbol{\theta}^{(-i)}) f(x^{(i)} | \theta^{(i)})}{\int_\Omega p(\theta | \boldsymbol{\theta}^{(-i)}) f(x^{(i)} | \theta) d\theta} \qquad (44) \\
&= \frac{\alpha H(\theta^{(i)}) f(x^{(i)} | \theta^{(i)}) + \sum_{j \neq i} f(x^{(i)} | \theta^{(i)}) \delta_{\theta^{(j)}}(\theta^{(i)})}{\left(\alpha \int_\Omega H(\theta) f(x^{(i)} | \theta) d\theta + \sum_{j \neq i} f(x^{(i)} | \theta^{(j)})\right)}
\end{aligned}$$

In order for this algorithm to be feasable, we need to be able to compute the integral in (44). This is generally the case when H is conjugate to f.

The resulting inference method is summarized in algorithm 1.

The algorithm was developed by [15] and used in [16]. It is a valid sampler corresponding to "algorithm 1" in [9]. However, convergence to the true posterior distribution is generally very slow (we say it has slow mixing behaviour). The reason is that often, multiple samples x will be associated with the same parameter value θ. However, the algorithm is unable to change the parameter value for more than one x simultaneously.

[10] Recall that we used the "$-d$" subscript to denote "all except d" so that $\boldsymbol{\phi}_{-d}$ is a short-hand for $(\phi_1, \ldots, \phi_{d-1}, \phi_{d+1}, \ldots, \phi_D)$.

[11] Note that, in accordance with the general description on Gibbs sampling above, we view the variables $\theta^{(1)}, \ldots, \theta^{(N)}$ as N coordinate variables.

Algorithm 1 Naïve Gibbs sampler for DP mixtures

```
Initialize $\theta^{(1)}, \ldots, \theta^{(N)}$
Set the total number of Gibbs iterations $L$
for $j = 1, \ldots, L$ do
    for $i = 1, \ldots, N$ in random order do
        Sample $\theta^{(i)} | \boldsymbol{\theta}^{(-i)}, x^{(i)}$ according to (44)
    end for
end for
return $\{\theta^{(1)}, \ldots, \theta^{(N)}\}$
```

So in order for the algorithm to change the parameter value for all x within a cluster, it has to do so individually for each observation, thereby passing through a low-probability intermediate state in which the cluster is split.

5.5. The standard Gibbs Sampler for Dirichlet Process Mixtures

We can improve our algorithm by splitting up inference for the cluster assignments and inference for the cluster parameters. To do this, we again introduce our cluster assignment variable $z^{(i)}$ defined such that $z^{(i)} = k$ if and only if $x^{(i)}$ is in cluster k. We also shift focus to the distinct cluster parameters $\bar{\theta}_1, \ldots, \bar{\theta}_K$ present in our data.

A Gibbs iteration now consists of two distinct loops. First, for each i, we remove $x^{(i)}$ from its current cluster and sample a new cluster assignment $z^{(i)}$ from the full conditional. Next, we sample a parameter $\bar{\theta}_k$ from the posterior for each currently present cluster k.

We can derive the full conditionals for the $z^{(i)}$ from (44) in the following way (where we use $\bar{\boldsymbol{\theta}}_{-k}$ to denote the collection of all currently present cluster parameters, $\bar{\boldsymbol{\theta}} = \{\bar{\theta}_1, \ldots, \bar{\theta}_K\}$):

$$
\begin{aligned}
&\Pr(z^{(i)} = k | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) = \Pr(\theta^{(i)} \in \{\bar{\theta}_k\} | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) \\
&= \int_{\{\bar{\theta}_k\}} p(\theta | \, \boldsymbol{\theta}^{(-i)}, x^{(i)}) d\theta \\
&= B \left[\alpha \int_{\{\bar{\theta}_k\}} H(\theta) f(x^{(i)} | \, \theta) d\theta + \sum_{j \neq i} \int_{\{\bar{\theta}_k\}} f(x^{(i)} | \, \theta) \delta_{\theta^{(j)}}(\theta) d\theta \right]
\end{aligned}
\tag{45}
$$

where we used B to denote the normalization constant (i.e. the inverse of the denominator in (44)). Note that, if H and F are smooth distributions (i.e. do not contain point-masses) then integrating them over a single point results in zero. Thus, we are left with

$$
\begin{aligned}
\Pr(z^{(i)} = k | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) &= B f(x^{(i)} | \, \bar{\theta}_k) \sum_{j \neq i} \mathbb{1} \left\{ \theta^{(j)} = \bar{\theta}_k \right\} \\
&= B f(x^{(i)} | \, \bar{\theta}_k) N_k^{(-i)} \\
&= \phi_k
\end{aligned}
\tag{46}
$$

where $N_k^{(-i)}$ is the current size of cluster k without counting sample $x^{(i)}$.

Of course, it is also possible that $x^{(i)}$ is in none of the current clusters. We may represent this by letting $z^{(i)} = K + 1$, where K is the current number of clusters (without counting sample $x^{(i)}$). The conditional probability is given by

$$
\begin{aligned}
&\Pr(z^{(i)} = K + 1 | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) = \Pr(\theta^{(i)} \in \Omega / \bar{\boldsymbol{\theta}} | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) \\
&= B \left[\alpha \int_{\Omega / \bar{\boldsymbol{\theta}}} H(\theta) f(x^{(i)} | \, \theta) d\theta + \sum_{j \neq i} \int_{\Omega / \bar{\boldsymbol{\theta}}} f(x^{(i)} | \, \theta) \delta_{\theta^{(j)}}(\theta) d\theta \right]
\end{aligned}
\tag{47}
$$

This time, the integral over the delta function terms vanishes since $\theta^{(j)} \in \bar{\boldsymbol{\theta}}$ for all $j \neq i$ and we are excluding those points from the domain of integration. Moreover, the integrand in the first term is continuous and so, taking its integral over $\Omega / \bar{\boldsymbol{\theta}}$ yields the same result as taking its integral over the entire space Ω since we are only excluding countably many points. Thus

$$
\begin{aligned}
\Pr(z^{(i)} = K + 1 | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)}) &= B \, \alpha \int_{\Omega} H(\theta) f(x^{(i)} | \, \theta) d\theta \\
&= \phi_{K+1}
\end{aligned}
\tag{48}
$$

Having worked out the full conditionals, we can sample

$$
z^{(i)} | \, \boldsymbol{z}^{(-i)}, \bar{\boldsymbol{\theta}}, x^{(i)} \sim \mathrm{Cat}(\phi_1, \ldots, \phi_K, \phi_{K+1}) \tag{49}
$$

If $z^{(i)} = K + 1$, $x^{(i)}$ is joining an empty cluster and we therefore need to sample a new parameter $\bar{\theta}_{K+1}$. The new parameter is sampled from the posterior distribution for θ based on the prior H and the single observation $x^{(i)}$:

$$
\bar{\theta}_{K+1} | \, x^{(i)} \sim \frac{H(\theta) f(x^{(i)} | \, \theta)}{\int_{\Omega} H(\theta') f(x^{(i)} | \, \theta') d\theta'} \tag{50}
$$

We also need to increment K by 1 in this case. This completes the first part of the Gibbs sampler.

In the second part of the algorithm, we loop over each currently present cluster k and sample a new parameter

Algorithm 2 Gibbs sampling inference for DP mixtures

```
Set the initial number of occupied clusters $K$
Initialize $\bar{\theta}_1, \dots, \bar{\theta}_K$ and $z^{(1)}, \dots, z^{(N)}$
Set the total number of Gibbs iterations $L$
for $j = 1, \dots, L$ do
    for $i = 1, \dots, N$ in random order do
        Set $N_{z^{(i)}} = N_{z^{(i)}} - 1$
        if $N_{z^{(i)}} = 0$ then
            Remove cluster $k$: Delete $\bar{\theta}_k$
            Set $K = K - 1$
            Rearrange indeces $k$ such that clusters
              $k = 1, \dots, K$ are non-empty clusters
        end if
        Sample $z^{(i)}$ according to (49)
        if $z^{(i)} = K + 1$ then
            Sample $\bar{\theta}_{K+1}$ according to (50)
            Set $K = K + 1$
        end if
    end for
    for $k = 1, \dots, K$ do
        Sample $\bar{\theta}_k$ according to (51)
    end for
end for
return $\{z^{(1)}, \dots, z^{(N)}\}$, $\{\bar{\theta}_1, \dots, \bar{\theta}_K\}$
```

$\bar{\theta}_k$ for that cluster. The sample is drawn from the posterior distribution of θ based on the prior H and all data points that are currently in cluster k:

$$\bar{\theta}_k | \{x^{(i)} : z^{(i)} = k\} \sim \frac{\left(\prod_{i:z^{(i)}=k} f(x^{(i)} | \theta)\right) H(\theta)}{\int_\Omega \left(\prod_{i:z^{(i)}=k} f(x^{(i)} | \theta')\right) H(\theta') d\theta'} \tag{51}$$

Note that we have dealt with the problem that we faced in the naïve Gibbs sampler. We now draw a new parameter value $\bar{\theta}_k$ for all datapoints in cluster k simultaneously.

In the Chinese restaurant metaphor, the first part of the algorithm corresponds to individually asking each customer i to get up from her current table and join a new table table at random. The probability for joining table k depends on what "dish" $\bar{\theta}_k$ is currently being served at that table. It is also possible for the customer to join a currently empty table, in which case a new dish will be served according to her taste (thinking of $x^{(i)}$ as "culinary preferences"). Once we have made each customer move, we clean all the tables and serve a new dish at each table (taking into account the preferences of the customers sitting at that table).

We have summarized this inference method in algorithm 2. This is the standard sample for Dirichlet process mixtures. It is due to [17] and corresponds to "algorithm 2" in Neal's paper [9].

As was the case for the previous Gibbs sampler, this method is only feasable if we can efficiently compute the integrals in (48), (50) and (51). This is generally the case if H is the conjugate prior to the likelihood function f.

Lastly, we briefly return to the problem of density estimation. After running our Gibbs sampler, we have estimates for the cluster assignments and the cluster parameters. Using these, we are able to formulate the posterior predictive distribution for a new sample $x^{(N+1)}$ as follows

$$\begin{aligned} p(x^{(N+1)} | x^{(1)}, \dots, x^{(N)}) = &\sum_{k=1}^{K} \frac{N_k}{\alpha + N} f(x^{(N+1)} | \bar{\theta}_k) \\ &+ \frac{\alpha}{\alpha + N} \int_\Omega f(x^{(N+1)} | \theta) H(\theta) d\theta \end{aligned} \tag{52}$$

Comparing this with the mixture density in (41), we can set the mixing coefficients $\pi_1, \dots, \pi_K$ as follows

$$\pi_k = \frac{N_k}{\alpha + N} \tag{53}$$

We are only able to find explicit formulas for the clusters that we have estimated to be present in our data set. The second term in the posterior predictive corresponds to all clusters that are currently empty (of which there are an infinite number). We can interpret $\int_\Omega f(x^{(N+1)} | \theta) H(\theta) d\theta$ as a mixture component representing the possibility that $x^{(N+1)}$ does not belong to any of the other clusters.

For more details on Gibbs sampling in DP mixtures, see [9, 18].

6. Implementation of the Normal-Inverse-Wishart Dirichlet Process Mixture model

In this section, we will describe our implementation of the Dirichlet process mixture model. We assumed that our data lives in D-dimensional Euclidean space, $\boldsymbol{x} \in \mathbb{R}^D$, and therefore chose a Gaussian likelihood model with a conjugate prior.

6.1. Conjugate Bayesian analysis of the multivariate Gaussian

Before descibing the details of the implementation, we shall briefly review Bayesian inference for the Gaussian distribution.

Suppose we have a set of D-dimensional variables $X^{(1)}, \dots, X^{(N)}$ that are independent and identically distributed according to a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$. The pdf of the Gaussian is given by

$$\mathcal{N}(\boldsymbol{x}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}D} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big((\boldsymbol{\mu} - \boldsymbol{x})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{x})\Big). \tag{54}$$

If we have a dataset of observations $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}\}$ so that $X^{(i)} = \boldsymbol{x}^{(i)}$, we can formulate the likelihood of the data as follows

$$p(\mathcal{D}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}^{(i)}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{55}$$

For simplicity, we will put a conjugate prior over the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which, in our case, is the *Normal-inverse-Wishart* (NIW) distribution. Its density function is defined as follows

$$\begin{aligned} \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \boldsymbol{m}_0, \kappa_0, \boldsymbol{S}_0, \nu_0) &= Z_{NIW} |\boldsymbol{\Sigma}|^{-\frac{\nu_0 + D + 2}{2}} \times \\ \exp\left(-\frac{\kappa_0}{2} (\boldsymbol{\mu} - \boldsymbol{m}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{m}_0) - \frac{1}{2} \mathrm{Tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}_0)\right) \end{aligned} \tag{56}$$

where Z_{NIW} is the normalization constant defined by

$$\frac{1}{Z_{NIW}} = 2^{\frac{\nu_0 D}{2}} \left(\frac{2\pi}{\kappa_0}\right)^{\frac{D}{2}} |\boldsymbol{S}_0|^{-\frac{\nu_0}{2}} \Gamma_D\left(\frac{\nu_0}{2}\right) \tag{57}$$

and Γ_D is the multivariate Gamma function defined as

$$\Gamma_D(t) = \pi^{\frac{D(D-1)}{4}} \prod_{j=1}^{D} \Gamma\left(t + \frac{1 - j}{2}\right) \tag{58}$$

We can interpret the hyper-parameters

$$\lambda_0 = \{\boldsymbol{m}_0, \kappa_0, \boldsymbol{S}_0, \nu_0\}$$

as follows: $\boldsymbol{m}_0 \in \mathbb{R}^D$ is our prior belief for $\boldsymbol{\mu}$ and $\kappa_0 > 0$ is the strength of that belief. The so-called scale matrix $\boldsymbol{S}_0$ is a positive definite $D \times D$ matrix and represents our prior belief for $\boldsymbol{\Sigma}$ (it is proportional to the prior mean of $\boldsymbol{\Sigma}$). The final hyper-parameter ν_0 is called the "degrees of freedom". It must satisfy $\nu_0 \geq D$ and measures the strength of our belief in the prior for Σ.

Because of conjugacy, we know that the posterior distribution of our parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is also the Normal-inverse-Wishart:

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \mathcal{D}) &= \frac{p(\mathcal{D}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \lambda_0)}{\iint p(\mathcal{D}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \lambda_0) d(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \\ &= \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \boldsymbol{m}_{\mathcal{D}}, \kappa_{\mathcal{D}}, \boldsymbol{S}_{\mathcal{D}}, \nu_{\mathcal{D}}) \end{aligned} \tag{59}$$

and in order to evaluate it we only need to know how to update the hyperparameters $\lambda_{\mathcal{D}} = \{\boldsymbol{m}_{\mathcal{D}}, \kappa_{\mathcal{D}}, \boldsymbol{S}_{\mathcal{D}}, \nu_{\mathcal{D}}\}$. The update formulas are

$$\begin{aligned} \boldsymbol{m}_{\mathcal{D}} &= \frac{\kappa_0 \boldsymbol{m}_0 + N \overline{\boldsymbol{x}}}{\kappa_0 + N} \\ \kappa_{\mathcal{D}} &= \kappa_0 + N \\ \boldsymbol{S}_{\mathcal{D}} &= \boldsymbol{S}_0 + \boldsymbol{S} + \kappa_0 \boldsymbol{m}_0 \boldsymbol{m}_0^T - \kappa_{\mathcal{D}} \boldsymbol{m}_{\mathcal{D}} \boldsymbol{m}_{\mathcal{D}}^T \\ \nu_{\mathcal{D}} &= \nu_0 + N \end{aligned} \tag{60}$$

where we have defined $\overline{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}^{(i)}$ as the sample mean and $\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T}$ as the uncentered sum-of-squares matrix.

For a derivation and discussion of these results (as well as a deeper look into the conjugate Bayesian analysis of the Gaussian) consult [19].

6.2. Implementation details

We have implemented the Dirichlet process mixture model with a Gaussian likelihood and conjugate prior using algorithm 2 for inference. The implementation is written in C++11 and we used the open-source linear algebra library "Armadillo" [20].

Most steps in the pseudo-code of algorithm 2 are independent of our choice for the likelihood model and the prior. The only steps that require some elaboration are sampling the cluster assignment variable $z^{(i)}$ (line 13 in the pseudo-code) and sampling the cluster parameters $\overline{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (lines 15 and 20).

6.3. Sampling from the Categorical distribution

We need to sample $z^{(i)}$ from the categorical distribution $\mathrm{Cat}(\phi_1, \dots, \phi_K, \phi_{K+1})$ where the quantities ϕ_k are defined in equations (46) and (48). Note that we do not need worry about the normalization constant B since we can always retrieve it from the sum-to-one constraint $\sum_{k=1}^{K+1} \phi_k = 1$. Furthermore, we will work with the logarithm of the unnormalized probabilities $\hat{\phi}_k = \log(\phi_k / B)$. We do this in order to counteract numerical instability which may occur since the unnormalized probabilites are typically very small in magnitude.

For $k = 1, \dots, K$, we have

$$\begin{aligned} \hat{\phi}_k &= \log\Big(N_k^{(-i)} f(x^{(i)}|\,\bar{\theta}_k)\Big) \\ &= \log\Big(N_k^{(-i)} \mathcal{N}(\boldsymbol{x}^{(i)}|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\Big) \\ &= \log N_k^{(-i)} - \frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_k| \\ &\quad + (\boldsymbol{\mu}_k - \boldsymbol{x}^{(i)})^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\mu}_k - \boldsymbol{x}^{(i)}) \end{aligned} \tag{61}$$

Algorithm 3 Smith & Hocking procedure [21] for generating $\mathcal{IW}(\boldsymbol{S}, \nu)$ distributed random matrices

```
Decompose $\boldsymbol{S}$ into $\boldsymbol{S} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is a lower-triangular $D \times D$ matrix. This can be done with a Cholesky-decomposition.
Initialize $\boldsymbol{A}$ to be the $D \times D$ null matrix: $\boldsymbol{A} = \boldsymbol{0}_D$

▹ Fill diagonal with square root of Gamma samples
for $i = 1, \dots, D$ do
    Sample $x \sim \text{Gamma}\left(\frac{\nu+1-i}{2}, 2\right)$
    Set $\boldsymbol{A}_{i,i} = \sqrt{x}$
end for

    ▹ Fill upper triangular part with $\mathcal{N}(0, 1)$ samples
for $i = 1, \dots, D$ do
    for $j = i + 1, \dots, D$ do
        Sample $z \sim \mathcal{N}(0, 1)$
        Set $\boldsymbol{A}_{i,j} = z$
    end for
end for

return $\boldsymbol{\Sigma} = \boldsymbol{L}\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{L}^T$
```

whereas for $k = K + 1$ we have

$$\begin{aligned}\hat{\phi}_{K+1} &= \log\left(\alpha \int_{\Omega} f(x^{(i)}|\,\theta)H(\theta)d\theta\right) \\ &= \log\alpha + \log J\end{aligned} \tag{62}$$

where

$$J = \iint \mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\lambda_0)\mathcal{N}(x^{(i)}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma})d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{63}$$

The integral J is the same as the integral in the evaluation of the posterior distribution for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (59) if we only consider the single observation $x^{(i)}$, so that $\mathcal{D} = \{x^{(i)}\}$. This implies that

$$\begin{aligned}J &= \frac{\mathcal{N}(\boldsymbol{x}^{(i)}|\,\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\,\lambda_0)}{\mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\,\lambda_{\mathcal{D}})} \\ &= \frac{Z_{NIW}(\lambda_0)(2\pi)^{-D/2}}{Z_{NIW}(\lambda_{\mathcal{D}})}\end{aligned} \tag{64}$$

where we have made the dependence of the normalization constant of the NIW distribution, Z_{NIW}, on the hyperparameters, λ, explicit.

With a little algebra (and making repeated use of the $\Gamma(t + 1) = t\,\Gamma(t)$ property of the Gamma function) we can show that

$$\begin{aligned}\hat{\phi}_{K+1} = &\log\alpha + \frac{1}{2}D\left(\log\kappa_0 - \log\kappa^{(i)} - \log\pi\right) \\ &+ \frac{1}{2}\left(\nu_0 \log|\boldsymbol{S}_0| - \nu^{(i)}\log|\boldsymbol{S}^{(i)}|\right) \\ &+ \log\Gamma\left(\frac{1}{2}(\nu_0 + 1)\right) - \log\Gamma\left(\frac{1}{2}(\nu_0 + 1 - D)\right)\end{aligned} \tag{65}$$

where we used the notation $\boldsymbol{S}^{(i)} := \boldsymbol{S}_{\mathcal{D}}$ and $\kappa^{(i)} := \kappa_{\mathcal{D}}$ for $\mathcal{D} = \{x^{(i)}\}$.

Once we have computed $\hat{\phi}_k$, we are ready to sample $z^{(i)}$ from $\text{Cat}(\phi_1, \dots, \phi_{K+1})$. First, we recover the unnormalized probabibilities

$$\phi_k = \frac{\exp\hat{\phi}}{\sum_{j=1}^{K+1}\exp\hat{\phi}_j} \tag{66}$$

Next, we sample a uniformly distributed variable

$$u \sim \text{Uniform}[0, 1] \tag{67}$$

Finally, we set

$$z^{(i)} = \min\{1, \dots, K + 1\} \qquad \text{such that} \qquad u \leq \sum_{j=1}^{z^{(i)}} \phi_j \tag{68}$$

We can easily verify that

$$\begin{aligned}\Pr\left(z^{(i)} = k\right) &= \Pr\left(\sum_{j=1}^{k-1}\phi_j < u \leq \sum_{j=1}^{k}\phi_j\right) \\ &= \sum_{j=1}^{k}\phi_j - \sum_{j=1}^{k-1}\phi_j \\ &= \phi_k\end{aligned} \tag{69}$$

and we are thus sampling from the correct distribution.

6.4. Sampling from the Normal-inverse-Wishart distribution

Both, in lines 15 and 20 of the pseudo-code for our Gibbs sampler (algorithm 2), we need to generate a sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from $\mathcal{NIW}(\lambda_{\mathcal{D}})$. In line 15, we have $\mathcal{D} = \{x^{(i)}\}$ and in line 20 we have $\mathcal{D} = \{x^{(j)} : z^{(j)} = k\}$.

We are able to factorise the pdf of the NIW as follows

$$\mathcal{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\,\boldsymbol{m}, \kappa, \boldsymbol{S}, \nu) = \mathcal{N}(\boldsymbol{\mu}|\,\boldsymbol{m}, \frac{1}{\kappa}\boldsymbol{\Sigma})\mathcal{IW}(\boldsymbol{\Sigma}|\,\boldsymbol{S}, \nu) \tag{70}$$

Algorithm 4 Procedure for sampling $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$

```
Decompose $\boldsymbol{S}$ into $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}^T$ using, for example, the Cholesky factorization.

▷ Generate $D$ independent standard normal samples
for $i = 1, \ldots, D$ do
    Sample $z_i \sim \mathcal{N}(0, 1)$
end for
Set $z = (z_1, \ldots, z_D)^T$

return $\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{A}z$
```

where $\mathcal{IW}(\boldsymbol{\Sigma}|\, \boldsymbol{S}, \nu)$ is the pdf of the *inverse Wishart* distribution. [12]

Thus, we can sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathcal{NIW}(\boldsymbol{m}, \kappa, \boldsymbol{S}, \nu)$ in the following way:

1. Sample $\boldsymbol{\Sigma}$ from an inverse Wishart distribution: $\boldsymbol{\Sigma} \sim \mathcal{IW}(\boldsymbol{S}, \nu)$ using algorithm 3
2. Having generated $\boldsymbol{\Sigma}$, we can sample $\boldsymbol{\mu}$ from a multivariate Gaussian distribution: $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}, \frac{1}{\kappa}\boldsymbol{\Sigma})$ using algorithm 4

This assumes that we have access to standard normal samples ($\mathcal{N}(0, 1)$) and also to samples following a Gamma(a, b) distribution. [13] C++11 has built-in functions that allow for the generation of both types of random numbers. If we only have access to Uniform[0, 1] samples, we can generate $\mathcal{N}(0, 1)$ samples using the Box-Muller transform [22]. Once we have access to both Uniform[0, 1] samples and $\mathcal{N}(0, 1)$ samples, we can generate Gamma(a, b) samples using the method proposed in [23].

6.5. Demonstration

In this secion, we briefly demonstrate the DP mixture model. For all examples shown, we set the concentration parameter to 1 and used a weakly informative data-dependent prior, as suggested in [3]:

$$\begin{aligned}
\alpha &= 1 \\
\boldsymbol{m}_0 &= \bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}^{(i)} \\
\kappa_0 &= 0.01 \\
\boldsymbol{S}_0 &= \frac{1}{N}\text{diag}\left(\sum_{i=1}^{N} \left(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}}\right)\right) \\
\nu_0 &= D + 2
\end{aligned} \tag{71}$$

Figures 7 and 8 illustrate how the number of clusters K discovered by the algorithm tends to grow with the size of the data. The model parameters are shown as ellipses in Figure 7. Each ellipse corresponds to a cluster, with the location of the ellipse corresponding to the mean parameter of that cluster and its shape corresponding to the covariance parameter. Each panel in Figure 8 depicts the estimated posterior distribution of K for the corresponding panel in Figure 7.

Figure 9 shows an example of how the Gibbs sampler evolves in the course of the algorithm. We use the same data as in Figure 7 (d) and initialize the algorithm with $K = 100$. At first, the algorithm creates a few redundant cluster (so that $K > 100$ after the first iteration). This allows the method to get away from the poor initial parameters. It is this behaviour that generally allows the algorithm to escape local optima and usually find the globul optimum. After 50 iterations, the number of clusters has already shrunk to $K = 20$ and after 100 iterations $K = 10$. We saw in Figure 8 (d) that the map estimate after 1200 iterations was $K = 8$.

Finally, we give an example of applying Dirichlet process mixtures to image segmentation in Figure 10. Colour-based image segmentation and image compression is often discussed as an application of finite mixture models and the K-Means algorithm. The main advantage of the DPM over finite mixture models in this application is that we do not need to specify the number of clusters a priori.

7. Extensions and Further Work

The Bayesian nonparametrics approach to machine learning is an active area of research. The Dirichlet process plays an important role in the Bayesian nonparametric framework as a prior distribution. It is the canonical distribution over probability measures due its wide support and tractable inference.

As such, many generalizations and extensions have been explored in the literature. A very popular general-

[12] The inverse Wishart distribution is a distribution over symmetric and positive definite $D \times D$ matrices. It is the conjugate prior distribution for the covariance parameter $\boldsymbol{\Sigma}$ of a Gaussian likelihood if we assume that the mean parameter $\boldsymbol{\mu}$ is known.

[13] The Gamma distribution with shape parameter a and rate parameter b, denoted Gamma(a, b), is a continuous distribution that has support over the positive real line $\mathbb{R}_+ = (0, +\infty)$. Its pdf is given by

$$\text{Gamma}(x|\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\left(-bx\right)$$

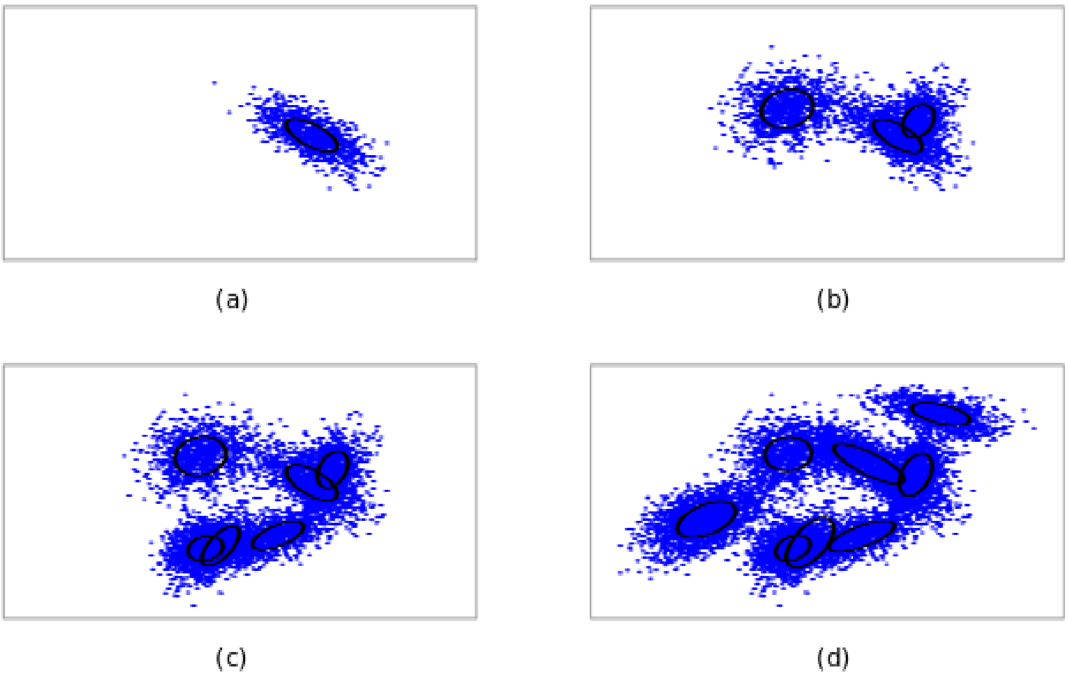

Figure 7: Clustering of N data points using the Dirichlet process mixture. (a) $N = 1000$ (b) $N = 3000$ (c) $N = 6000$ (d) $N = 10000$. Each data set contains the previous one as a subset.

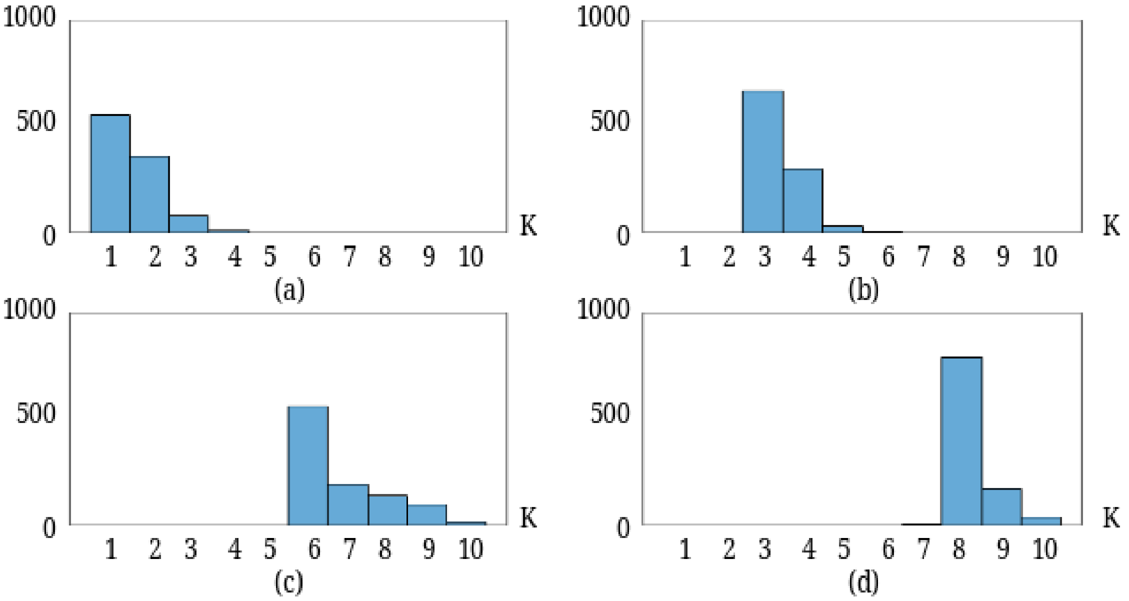

Figure 8: Posterior distribution of the number of clusters K for the data sets in figure 7. The Gibbs sampler was run for 1200 iterations and we discarded the first 200 as burn-in.

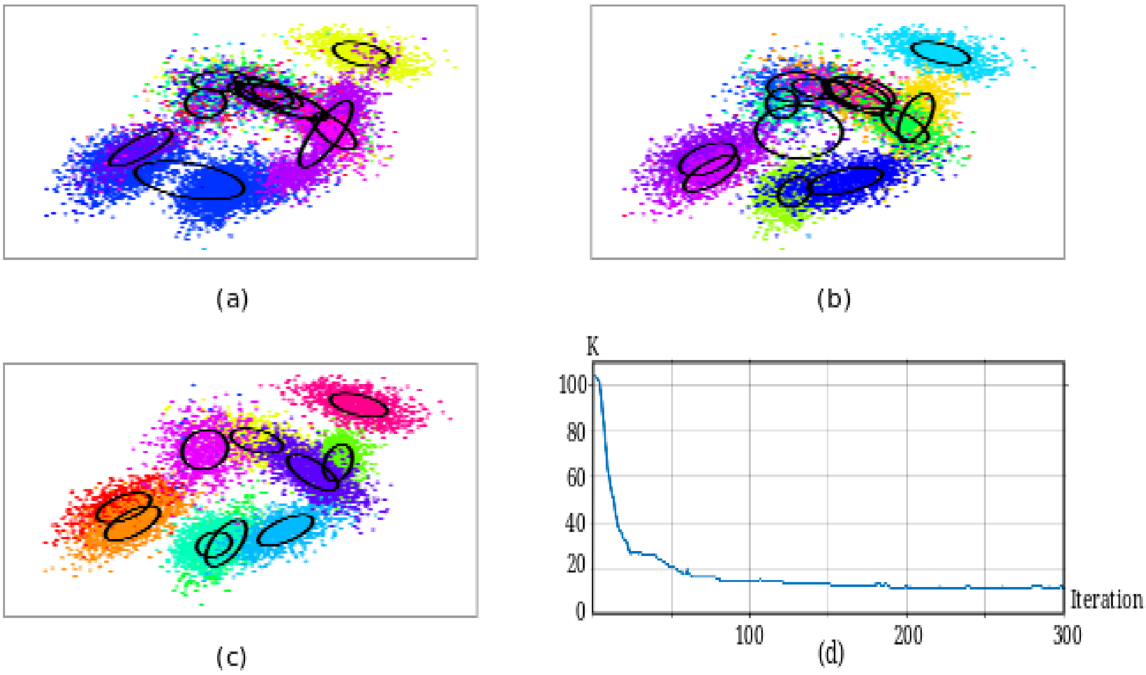

Figure 9: Clustering of the data set in figure 7 (d) using the Dirichlet process mixture model. The inital number of clusters is set to $K = 100$ and we show the output of the algorithm after j iterations. (a) $j = 10$ (b) $j = 100$ (c) $j = 300$. Panel (d) plots K against the number of iterations.

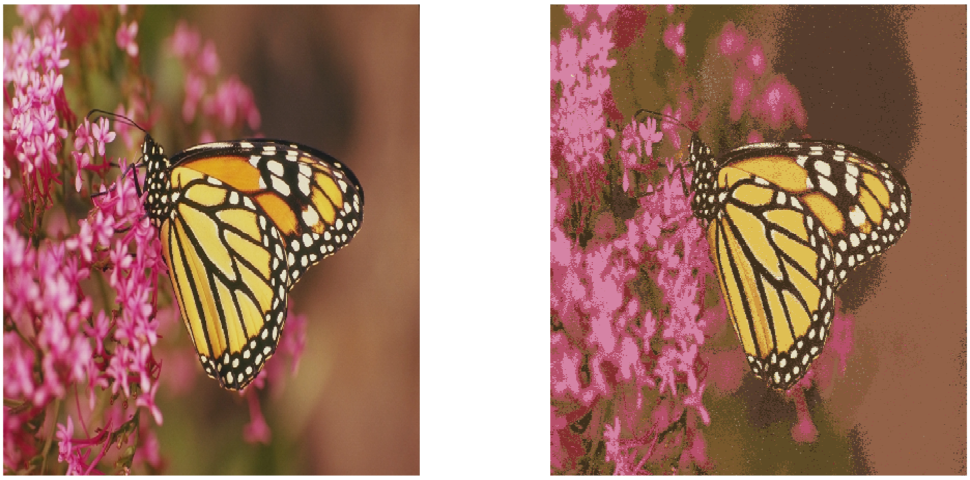

Figure 10: Example an application of the Diirichlet process mixture to image segmentation. The original image is on the left and the segmented image is on the right. The mode of the posterior distribution for the number of clusters K is $K = 19$.

ization is the *Pitman-Yor process* [24, 25]. It has an additional parameter $d \in [0, 1)$ and reduces to the Dirichlet process when $d = 0$. If d is close to 1, the clusters generated by the Pitman-Yor process exhibit a power-law behaviour in the sense that it will produce a few large clusters and many very small clusters. All of the representations of the Dirichlet process discussed in section 4 can be generalized to the Pitman-Yor process.

Other generalizations of the DP include Pólya trees, stick-breaking priors and Poisson-Kingman models. These can be derived by extending one of the representations of the DP. A different class of extensions uses the Dirichlet process as building blocks to develop more complex models. Two such models are the dependent Dirichlet processes [26] and hierarchical Dirichlet processes [27].

Beside extensions of the DP, recent research has also been focused on exploring more efficient inference methods in Dirichlet process models that go beyond our simple Gibbs sampler. Finally, there has also been considerable interest, both in establishing theoretical properties of DP models and in their practical applications.

Bayesian nonparametric methods provide a powerful set of tools for machine learning and show great potential for a wide range of applications. However, the complexity of these models and the expense in implementing them poses a significant hurdle to practitioners and researchers. We hope that, with this paper and our software package, we can make Bayesian nonparametric methods available to a wider audience and encourage growth in their application.

Bibliography

[1] Y. W. Teh, Dirichlet Processes: Tutorial and Practical Course, Machine Learning Summer School, 2007.

[2] C. M. Bishop, Pattern recognition and machine learning, Vol. 1, springer, 2006.

[3] K. P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

[4] T. S. Ferguson, A bayesian analysis of some nonparametric problems, Ann. Statist. 1 (2) (1973) 209–230.

[5] Y. W. Teh, Dirichlet processes, in: Encyclopedia of Machine Learning, Springer, 2010.

[6] D. Blackwell, J. B. MacQueen, Ferguson distributions via polya urn schemes, Ann. Statist. 1 (2) (1973) 353–355.

[7] W. J. Ewens, Mathematical and Statistical Developments of Evolutionary Theory, Springer Netherlands, Dordrecht, 1990, Ch. Population Genetics Theory - The Past and the Future, pp. 177–227.

[8] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica 4 (1994) 639–650.

[9] R. M. Neal, Markov chain sampling methods for dirichlet process mixture models, Journal of Computational and Graphical Statistics 9 (2) (2000) 249–265.

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

[11] S. Jain, R. M. Neal, A split-merge markov chain monte carlo procedure for the dirichlet process mixture model, Journal of Computational and Graphical Statistics 13 (1) (2004) 158–182.

[12] P. Fearnhead, Particle filters for mixture models with an unknown number of components, Statistics and Computing 14 (1) 11–21.

[13] T. Minka, Z. Ghahramani, Expectation propagation for infinite mixtures, NIPS Workshop on Nonparametric Bayesian Methods and Infinite Models 19.

[14] D. M. Blei, M. I. Jordan, Variational inference for dirichlet process mixtures, Bayesian Anal. 1 (1) (2006) 121–143.

[15] M. D. Escobar, Estimating normal means with a dirichlet process prior, Journal of the American Statistical Association 89 (425) (1994) 268–277.

[16] M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (430) (1995) 577–588.

[17] S. N. MacEachern, Estimating normal means with a conjugate style dirichlet process prior, Communications in Statistics - Simulation and Computation 23 (3) (1994) 727–741.

[18] P. Orbanz, Lecture notes on bayesian nonparametrics.

[19] K. P. Murphy, Conjugate bayesian analysis of the gaussian distribution, def 1 ($2\sigma 2$) (2007) 16.

[20] C. Sanderson, Armadillo: An open source c++ linear algebra library for fast prototyping and computationally intensive experiments, Tech. rep., NICTA (2010).

[21] W. B. S. Smith, R. R. Hocking, Algorithm as 53: Wishart variate generator, Journal of the Royal Statistical Society. Series C (Applied Statistics) 21 (3) (1972) 341–345.

[22] G. E. P. Box, M. E. Muller, A note on the generation of random normal deviates, Ann. Math. Statist. 29 (2) (1958) 610–611.

[23] G. Marsaglia, W. W. Tsang, A simple method for generating gamma variables, ACM Trans. Math. Softw. 26 (3) (2000) 363–372.

[24] J. Pitman, M. Yor, The two-parameter poisson-dirichlet distribution derived from a stable subordinator, Ann. Probab. 25 (2) (1997) 855–900.

[25] Y. W. Teh, A Bayesian interpretation of interpolated Kneser-Ney, Tech. Rep. TRA2/06, School of Computing, National University of Singapore (2006).

[26] S. N. MacEachern, Dependent nonparametric processes, in: ASA proceedings of the section on Bayesian statistical science, Alexandria, Virginia. Virginia: American Statistical Association; 1999, 1999, pp. 50–55.

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