Bayesian Non-parametrics Priors with Density Estimation

Andi Wang Kaspar Märtens Ho Chung Leon Law

University Of Oxford
Department Of Statistics

11/02/16

Abstract

In this report we motivate the definitions for the Dirichlet process and Pólya trees as non-parametric Bayesian priors. We discuss some technical conditions related to the definitions and their use in non-parametric density estimation.

1 Introduction

It is generally acknowledged in modelling that fewer assumptions are preferable to more assumptions. This is the broad motivation for non-parametric methods, where the model parameter is infinite-dimensional[1]. However, this is especially difficult in a Bayesian context, since this involves placing priors on infinite-dimensional objects and in this report, we discuss two such priors: the Dirichlet Process and Pólya Trees.

In section 2 we describe the Dirichlet process (DP), starting from the Dirichlet distribution and concluding with its extension to Dirichlet process mixture models. Section 3 goes on to discuss a generalisation of Dirichlet processes: Pólya trees, with some original discussion on various technical details. Finally in section 4 we apply both objects to the problem of non-parametric density estimation.

2 The Dirichlet Processes

2.1 Motivation and Definition

One straightforward (and sometimes only) way to analyse data is to categorise it. That is, we devise a set of k categories $\{i_1, \ldots, i_k\}$, so that given an observation X, it belongs to exactly one category. After we observe our data $X_1, \ldots, X_n$, we can then ask, what would be the probability that a new observation X_{n+1} is in say, category i? Now, thinking only of category labels and assuming independence, it is natural to model the data as being generated from a *multinomial distribution.*

At this point, the Bayesian way to proceed is to place a prior on the probabilities p_i that an observation will have label i, for each $i \in \{1, \ldots, k\}$. The probability vector $(p_1, \ldots, p_k)$ must of

[1] So technically our model is still parametric, but the parameter space is infinite-dimensional.

course live in

$$S_k = \{(x_1, \dots, x_k) \in \mathbb{R}^k : x_i \geq 0 \ \forall i, \sum_{i=1}^{k} x_i = 1\}.$$

But what prior to choose? The natural choice is the *Dirichlet distribution* with positive parameters $\alpha = (\alpha_1, \dots, \alpha_k)$ with probability density function on ${S_k}^2$ proportional to

$$f(x_1, \dots, x_k; \alpha) \propto \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

Now it is clear why this is a good idea, since by Bayes's Theorem the posterior probabilities are also distributed according a Dirichlet distribution with updated parameters $\alpha' = \{\alpha_1 + n_1, \dots, \alpha_k + n_k\}$ where n_i is the number of data points in category i.

The Dirichlet distribution also has a remarkable *aggregation property*, that is if $(p_1, \dots, p_k)$ has Dirichlet distribution with parameters α as above, the random vector $(p_1, \dots, p_i + p_j, \dots, p_k)$ where p_i and p_j are removed and replaced by their sum $p_i + p_j$ is distributed according to the Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_i + \alpha_j, \dots, \alpha_k)$, where α_i and α_j are correspondingly removed and replaced with their sum. This follows from an alternative characterisation of the Dirichlet distribution in terms of normalised independent Gamma random variables and using their summation property.

With this in mind, we hope to generalise the Dirichlet distribution to larger, non-discrete spaces. Let $(\Theta, \mathcal{B})$ be a measure space and G_0 a probability measure on this space. By the 'fractal' aggregation property of the Dirichlet distribution, we might ask: can we define a random probability measure which sprinkles mass on Θ according to a Dirichlet distribution? The answer is yes, and we have the *Dirichlet process* $\mathrm{DP}(\alpha_0, G_0)$ with parameters $\alpha_0 > 0, G_0$, which is the distribution of a random measure G such that for any finite measurable partition $(A_1, \dots, A_r)$ of Θ, the random vector $(G(A_1), \dots, G(A_r))$ is distributed according to the Dirichlet distribution, with parameters $(\alpha_0 G_0(A_1), \dots, \alpha_0 G_0(A_r))^3$. With this definition, the Dirichlet distribution is simply the Dirichlet process on $(\Theta, \mathcal{B}) = (\{1, \dots, k\}, \mathcal{P}(\{1, \dots, k\})$ and G_0 is the normalised α vector from before.

It can be shown that the Dirichlet process is discrete with probability one, and is the nonparametric conjugate prior for an unknown (discrete) distribution. Figure 1 shows some typical draws from the Dirichlet process defined on $\mathbb{R}$ with G_0 being a standard Gaussian for various values of α.

One way to represent the Dirichlet process is the *stick-breaking construction* [2]. Given the parameters α_0, G_0, generate $\pi'_k \overset{iid}{\sim} \mathrm{Beta}(1, \alpha_0)$ and $\phi_k \overset{iid}{\sim} G_0$, then define $\pi_k = \pi'_k \prod_{l=1}^{k-1}(1 - \pi'_l)$ and $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$. The resulting G is a random measure distributed according to $\mathrm{DP}(\alpha_0, G_0)$.

Another perspective on the Dirichlet process is provided by the *Chinese restaurant process*, which refers to draws from the Dirichlet process random measure; see [2] for details.

So the Dirichlet process is a random discrete measure, defined on any measurable space. We now move on to a popular application of Dirichlet processes, Dirichlet process mixture models.

[2] Note that S_k is a $(k-1)$-dimensional space (since $x_k = 1 - x_1 - \dots - x_{k-1}$), and this density is with respect to Lebesgue measure on $\mathbb{R}^{k-1}$.

[3] We have to now define the Dirichlet distribution when one or more parameters are 0 to cover the case of G_0-null sets. In this case, any such components are defined to be degenerate at 0, and we omit that component in the density. Since the random vector must lie in the simplex it may be the case that there is only one component remaining (i.e. if $G_0(A_i) = 1$), in which case that component is degenerate at 1.

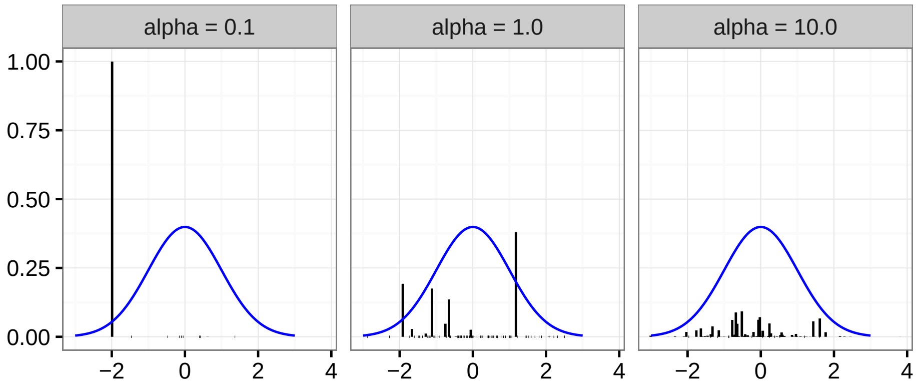

Figure 1: Draws from the Dirichlet process $\mathrm{DP}(\alpha_0, G_0)$ for $G_0 = \mathcal{N}(0,1)$ base measure (density shown in blue) and various values of $\alpha_0 \in \{0.1, 1.0, 10\}$ (in three panels) illustrated. Each of the draws is a discrete distribution, whereas the height of the bars denotes their weights. When α_0 is close to zero, the realizations become concentrated on a single value.

2.2 Dirichlet Process Mixture Models

Dirichlet processes offer a flexible non-parametric framework for modelling discrete data, but they are not directly applicable to the case when our data has a density with respect to Lebesgue measure. As an extension, we introduce the Dirichlet process mixture (DPM) models [2] [3].

First we need to introduce the family of parametric distributions F parametrised by ϕ, with density $p(x|\phi)$. Now, letting $G \sim \mathrm{DP}(\alpha_0, G_0)$, the DPM is obtained by smoothing out G using densities p, i.e. the DPM specifies the following model

$$p(x) = \int p(x|\phi) G(d\phi)\,.$$

That is, the DPM is defined as follows[4]

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha_0, G_0) \\ \phi_i \,|\, G &\sim G \\ x_i \,|\, \phi_i &\sim F(\phi_i) \end{aligned} \tag{1}$$

So, in the DPM, we specify the Dirichlet process as a prior on the parameters ϕ and use G as a mixing measure, hence the term *mixture model.* To give more intuition, recall that $G = \sum_k \pi_k \delta_{\phi_k}$ is a discrete measure, allowing to replace the integral with a countable sum

$$p(x) = \sum_{k=1}^{\infty} \pi_k p(x|\phi_k). \tag{2}$$

A common choice is to specify the family of distributions to be Gaussian. For example, $p(x|\phi_k)$ could be the density of $\mathcal{N}(\phi_k, \sigma^2)$ with known variance. Figure 2 provides an example of a DPM $p(x) = \sum_k \pi_k \mathcal{N}(\phi_k, 0.1)$ (on the right), where the mixing measure G is shown on the left.

Note that if the sum in (2) was finite, with K non-zero terms, it would correspond to the parametric mixture model. It turns out that the DPM is obtained as the limit $K \to \infty$.

[4] The notation $X \sim F$ means that X is distributed according to distribution F.

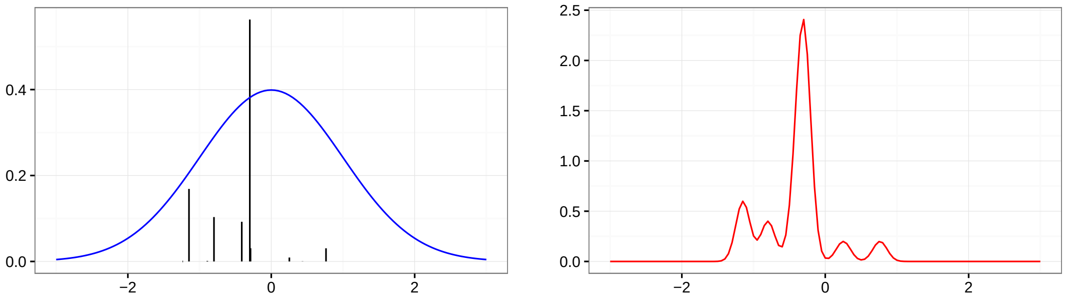

Figure 2: Dirichlet process mixture model: A draw G from $\mathrm{DP}(\alpha_0, G_0)$ on the left, $G = \sum_k \pi_k \delta_{\phi_k}$, and a corresponding Gaussian mixture model $\sum_k \pi_k \mathcal{N}(\phi_k, 0.1)$ on the right.

To obtain a more traditional formulation of the mixture model, we could introduce cluster assignment variables z_i, which take value k with probability π_k, and replace ϕ_i with distinct values ϕ_{z_i}.

Sampling from the posterior distribution of a DP mixture model can be carried out via Gibbs sampling; see [4] for details.

3 Pólya Trees

We now discuss a generalisation of the Dirichlet process, Pólya trees, which can place positive mass on continuous measures. This generalisation is much more flexible and allows us to work on a wide range of problems, and is especially popular in survival analysis, where censored data is common. The following discussions draws from [5], [6] and [7].

3.1 Motivation and Definition

We start by defining some notation. Let Ω be a separable measurable space and for each $m \in \mathbb{N}$ let E^m be the space of m-digit binary numbers $E^m = \{\epsilon = e_1 \dots e_m \colon e_j \in \{0,1\}, j = 1, \dots, m\}$. Define $\Pi_m = \{B_\epsilon,\ \epsilon \in E^m\}$ to be a partition of Ω into 2^m subsets as follows: let $\Pi_0 = \Omega$ and sequentially define Π_{m+1} by refining each $B_{e_1 \dots e_m}$ from Π_m into $B_{e_1 \dots e_m} = B_{e_1 \dots e_m 0} \cup B_{e_1 \dots e_m 1}$. Degenerate splits are permitted, i.e. $B_\epsilon \cup \emptyset$, but we require $\Pi_0, \Pi_1, \dots$ to be such that $\cup_0^\infty \Pi_m$ generates the measurable sets. Finally, let $E = \cup_0^\infty E^m$.

We wish to define a random probability measure P on this set of tree-like partitions, and we need to define it coherently through the partitions, with $P(B_\epsilon) = P(B_{\epsilon 0}) + P(B_{\epsilon 1})$. An intuitive way to do this is to define P through a sequence of conditional probabilities G with $B_\emptyset$ defined as Ω. For $\epsilon = e_1 \dots e_m \in E^m$ set

$$P(B_\epsilon) = \prod_{j=1}^{m} G(B_{e_1 \dots e_{j-1} e_j} | B_{e_1 \dots e_{j-1}}) \tag{3}$$

This motivates the definition of Pólya trees to be the following [5]:

A random probability measure P on Ω is said to have a Pólya tree distribution, or a Pólya tree prior, with parameters $(\Pi = \{\Pi_m; m = 0, 1, \dots\}, \mathcal{A})$, written $P \sim \mathrm{PT}(\Pi, \mathcal{A})$, if there exist non-negative numbers $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in E\}$ and random variables $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E\}$ such that the following holds:

1. All the random variables in $\mathcal{Y}$ are independent;
2. for every $\epsilon \in E, Y_\epsilon$ has a beta distribution[5] with parameters $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$;
3. for every $m = 1, 2, \ldots$ and every $e_1 \ldots e_m \in E^m$,

$$P(B_{e_1 \ldots e_m}) = \Big(\prod_{j=1;\, e_j=0}^{m} Y_{e_1 \ldots e_{j-1}} \Big) \Big(\prod_{j=1;\, e_j=1}^{m} (1 - Y_{e_1 \ldots e_{j-1}}) \Big) \tag{4}$$

A draw from this distribution can thought of as a particle cascading through the partitions, where at each level, given the previous partition, it falls into one of the two partitioning subsets at the next level with some random probability Y_ϵ. This randomness at each level is what allows us to construct a random probability measure, and we will see that this simple construction gives us very nice properties.

3.2 Properties of Pólya Tree and Prior Mean

One of the main attractions of the Pólya Tree is that it can assign positive mass to a set of continuous distributions[6], when the $\alpha_{e_1 \ldots e_m}$ increase sufficiently fast with m. A common choice for $\alpha_{e_1 \ldots e_m} = cm^2$, $c > 0$, although in general a prior can be placed on c. On the other hand, the Dirichlet process, which is discrete with probability 1, can also arise as a special case of the Pólya tree [1], when for every $\epsilon \in E, \alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$, for instance when $\alpha_{e_1 \ldots e_m} = c2^{-m}$.

Let us give some intuition for this phenomenon in the case when $\alpha_{e_1 \ldots e_m} = c\rho(m)$. By properties of the beta distribution, for $\epsilon \in E^m$ the Y_ϵ have mean $1/2$ and variance $1/4(2c\rho(m) + 1)$. So if ρ grows quickly with m, the variance of the random probabilities decreases to zero, so the random probabilities are more tightly distributed around 1/2 the further down the tree we travel. This means as the particle cascades down to the finer levels of the partition it tends to smooth out locally, since it goes 'left' and 'right' with roughly equal probability. This 'local smoothness' is exactly what means the measure has a density with respect to Lebesgue measure. On the other hand, when ρ *decreases* with m, as it does for the Dirichlet distribution, the random probabilities tend to be more extremal, taking values close to 0 or 1, so the particles become more and more concentrated, following the same paths at deeper levels of the tree, leading to discrete point masses.

Another particularly attractive feature of the Pólya tree is that you can center it around a given distibution G_0. In the case when $\Omega = \mathbb{R}$ with its Borel σ-algebra $\mathcal{B}(\mathbb{R})$, and $F : \mathbb{R} \to [0, 1]$ is the cumulative distribution function associated with G_0, one way to do this using the following simple lemma, which we have proved here. Let use write $\sigma(F)$ for the σ-algebra generated by the inverse images of F, the sets $\{F^{-1}(A) : A \in \mathcal{B}([0, 1])\}$.

Lemma 1 $\sigma(F) = \mathcal{B}(\mathbb{R}) \Leftrightarrow F$ *is strictly increasing.*

Proof ($\Rightarrow$): Suppose F is not strictly increasing. Then there exists $x < y$ such that $F(x) = F(y)$. Take a, b such that $x < a < b < y$. Then (a, b) is not in $\sigma(F)$, since any set containing (a, b) must also include x and y.

($\Leftarrow$): Since the sets $(-\infty, x]$ for each $x \in \mathbb{R}$ generate $\mathcal{B}(\mathbb{R})$ and $\sigma(F) \subset \mathcal{B}(\mathbb{R})$ (since F is measurable), it suffices to check that for any $x \in \mathbb{R}$, $(-\infty, x] \in \sigma(F)$. Fix $x \in \mathbb{R}$. Claim:

[5] Here we allow degenerate Beta distributions, i,.e. random variables which take value 0 with probability 1 or value 1 with probability 1.

[6] That is, the associated random variable has a density with respect to Lebesgue measure with positive probability. Note that this density itself is not necessarily continuous.

$F^{-1}([0, F(x)]) = (-\infty, x]$. Suppose $y \leq x$. Then since F is increasing, $F(y) \leq F(x)$. Conversely, suppose $F(y) \leq F(x)$. Then we must have $y \leq x$, since if not, $y > x$, and since F is strictly increasing we would have $F(y) > F(x)$, a contradiction. This proves the claim. □

Note that requiring F to be strictly increasing is a strong assumption; in particular if $X \sim G_0$ it means for any $a < b$ we have $\mathbb{P}(a < X \leq b) > 0$.

Suppose F is a strictly increasing cumulative distribution function. Then F has an inverse $g : (0, 1) \to \mathbb{R}$, and set $g(0) = -\infty$ and $g(1) = \infty$. One could fix the partitions Π_m to be the dyadic quantiles of G_0, $(g(k/2^m), g((k+1)/2^m)]$, where $k = 0, 1, \ldots, 2^m - 1$, and by the Lemma these would indeed generate $\mathcal{B}(\mathbb{R})$.

For $m = 1$, the subsets $\{B_0, B_1\}$ are simply the subsets around the median of G_0 and for $m = 2$, the subsets are the quartiles of the distribution, and so on. One can show that taking $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ gives $E(P(B)) = G_0(B)$, where the expectation is under the described Pólya tree. Intuitively this is because symmetry in the beta parameters means that the random probabilities are are symmetrically distributed around $1/2$, so a particle cascading down the tree at each level is on average equally likely to go left or right, and since the G_0-probabilities at each level are equal by construction, we see the expecion should be $G_0(B)$.

Note that if the the random measure generated by the Pólya tree does have a density with respect to Lebesgue measure, in general it will have discountinuities at the boundaries of the partitions. One way to smooth out the density at the boundaries is to allow our centering distribution be dependedent on some hyperparameter β, where $\beta \sim g(\beta)$, this is called a Pólya tree mixture.

3.3 Posterior Update and Computational Approach

Let us assume the following:

$$x|P \sim P \qquad P \sim \mathrm{PT}(\Pi, \mathcal{A}) \tag{5}$$

Then the posterior is given by a Pólya tree, $P|x \sim \mathrm{PT}(\Pi, \mathcal{A}^*)$ with:

$$\alpha^* = \begin{cases} \alpha_\epsilon + 1, & \text{if } x \in B_\epsilon \\ \alpha_\epsilon, & \text{otherwise} \end{cases} \tag{6}$$

This means that the Pólya tree is a nonparametric conjugate prior, and the α_ϵ parameters are incremented by one on each subset B_ϵ that contains x; recall that $Y_\epsilon \sim \mathrm{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$. In practice, since we cannot simulate infinitely many levels, we would have to use a finite tree with say M_{fin} levels (normally ≈ 7, which creates 2^7 sets in M_{fin}). Also, this type of posterior update is very useful in survival analysis, where censored data is very common; if we only know an observation x lie inside some interval, and if this interval happens to be equal to some B_ϵ, then we just update up to that partition.

In general, to simulate from $\mathrm{PT}(\Pi, \mathcal{A})$, we first simulate from all the Y_ϵ, for $\epsilon \in E^m$, $m = 1, \ldots, M_{\text{fin}}$, which then allows us to calculate $P(B_\epsilon)$ for $\epsilon \in \cup_{j=1}^{M_{\text{fin}}} E^j$. A problem with stopping at the M_{fin} level is that probability is not defined within the final M_{fin} partition. To deal with this, one choice is to use uniforms, but this is somewhat *ad hoc*. Instead, if the Pólya tree is centered around some distribution, then the mean measure G_0 can be used, appropriately scaled on each set.

One such example of this is drawing a new future observation x_{n+1} from the model defined in 5, to do this, first generate Y_0^* (a beta random variable with updated parameters as in 6), then simulate $e_1 = I(x_{n+1} \in B_0) \sim \text{Ber}(y_0)$, given $Y_0^* = y_0$, then simulate similarly $e_2 = I(x_{n+1} \in B_{e_1 0})$ according to $Y_{e_1 1}^*$ similarly and iterating this until arrive at level K where $\alpha_{e_1 \dots e_K} = \alpha_{e_1 \dots e_K}^*$. Then, we can simply generate x_{n+1} from the restricted prior mean G_0; $x_{n+1} \sim G_0$ conditional on $x_{n+1} \in B_{e_1 \dots e_K}$. Note that there is no difference if we did not stop this iteration, and proceed until the last level M_{fin}, since at from this level onwards it agrees with the prior.

We will give now give some intuition on the choice of α in terms of the predictive posterior distribution. Suppose the α_ϵ are large, then the distribution of $x_{n+1}|x_1, \dots x_n$ will be close to the prior mean. If the α_ϵ are large, the beta distributions are quite peaked (low variance) hence draws from the prior tend to be similar. For the posterior, incrementing the α_ϵ will not have a large effect since the α_ϵ are already large. On the other hand, if the α_ϵ are quite small, the increment of the parameters will change the distribution from a higher variance distribution to a more peaked distribution, so the distribution of $x_{n+1}|x_1, \dots x_n$ will be closer to that of the sampling distribution function.

Following on from this, we give some more motivation here for choosing an increasing function $\rho(m)$[7] (if we believe the underlying distribution is continuous). Note for small m, it is not necessary for $P(B_{\epsilon 0})$ to be similar to $P(B_{\epsilon 1})$, as the partitions are still of reasonable size, and we want the model to be adaptive to the data, hence a large amount of variability is desirable, i.e. large variance in the beta distributions. But for large m, we would expect $P(B_{\epsilon 0})$ to not change too much from $P(B_{\epsilon 1})$ if we believe that P should be continuous, so a large α_ϵ is more desirable, where the beta distribution has less variance.

4 Applications and experiments

Density estimation is a common problem in statistics, and we would prefer to assume as little about the data as possible. First, we will compare two non-parametric approaches for density estimation: Dirichlet process mixture models and Pólya trees. The latter can be directly used for density estimation, whereas for the DPM, we need to fix the parametric family of distributions. A natural choice is the family of Gaussian distributions. Afterwards, we will discuss the suitability of DPM for identifying the number of mixture components.

4.1 Density estimation

We generated data from a bimodal distribution $0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1)$ (a histogram of observed values is in Figure 3), and carried out density estimation with a DPM and Pólya trees.

We specified a DPM using Gaussian densities with mean ϕ_k and common precision λ, and used a normal-gamma base measure on those parameters. We implemented the Gibbs sampling scheme according to [4], and plotted the resulting density estimate in Figure 3. Each of the grey lines represents a draw from the posterior (i.e. we drew ϕ_k, λ from the posterior and constructed the respective mixture density), whereas the red line shows the averaged density estimate.

Next, we fitted a Pólya tree model to the same data, shown in Figure 4. As we have discussed earlier, there are discontinuities at the boundaries of the partitions. For the density on the right, the partitions (defined by the dyadic quantiles) are wider as the base measure has a greater variance.

[7] Recall $\alpha_{e_1 \dots e_m} = c\rho(m)$.

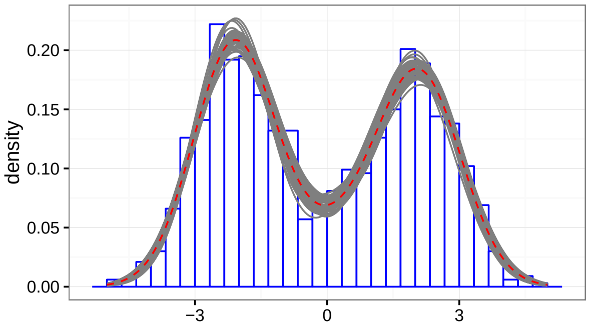

Figure 3: Density estimate obtained by DPM. We generated 1000 data points (histogram in blue) and fitted a DPM model. Samples from the posterior are shown in grey, and their average value with a red dashed line.

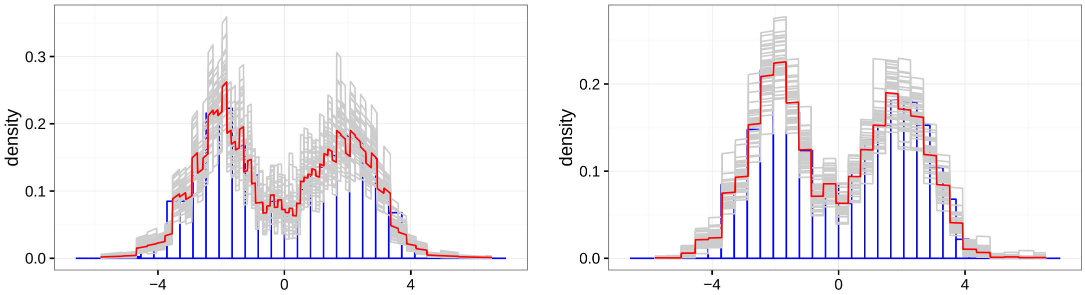

Figure 4: Density estimate obtained by Pólya trees with two base measures, $G_0 = \mathcal{N}(0, 3^2)$ (left) and $G_0 = \mathcal{N}(0, 10^2)$ (right) with $\alpha_\epsilon = \alpha m^2$, where $\alpha \sim \Gamma(1, 0.01)$. We generated 1000 data points (histogram in blue) and fitted a PT. Samples from the posterior are shown in grey, and their averaged density estimate in red.

Even though the approximation by Dirichlet process mixture model may seem to be more suitable in this case, it is important to note that our data generating mechanism coincided with the model used by DPM, both being a mixture of Gaussians. This is in contrast to the Pólya tree, which estimates the density directly, thus making less assumptions. This can simplify the process in situations when it is not clear how the data arises.

4.2 Clustering with DP mixture model

Density estimation with the DPM relates closely to the clustering problem, as this is often carried out by fitting a mixture model to the data, but now the number of mixture components also becomes a parameter of interest. We will focus on fitting a Gaussian mixture model, when the data generating mechanism is based on a finite number of clusters K_0.

A parametric *ad hoc* approach would be to fit the mixture for several K values and use model selection to choose the "best" one. Within the Bayesian non-parametrics framework, we can leave the number of mixture components unspecified, fit a DPM and explore the posterior distribution for K.

This approach is often used in practice, however, the DPM is a misspecified model in this case; when using the Dirichet process the implicit modelling assumption is that as the number of observations $n \to \infty$, we will observe an infinite number of clusters. It turns out that the posterior distribution of K is not consistent in this case, i.e. it does not concentrate on the true K_0; see [8] for details. To empirically verify the behavior of this posterior, we generated data from a mixture of one-dimensional Gaussians, containing either two ($K_0 = 2$) or five ($K_0 = 5$) components, fitted a DPM and explored the posterior distribution over K (Figure 5).

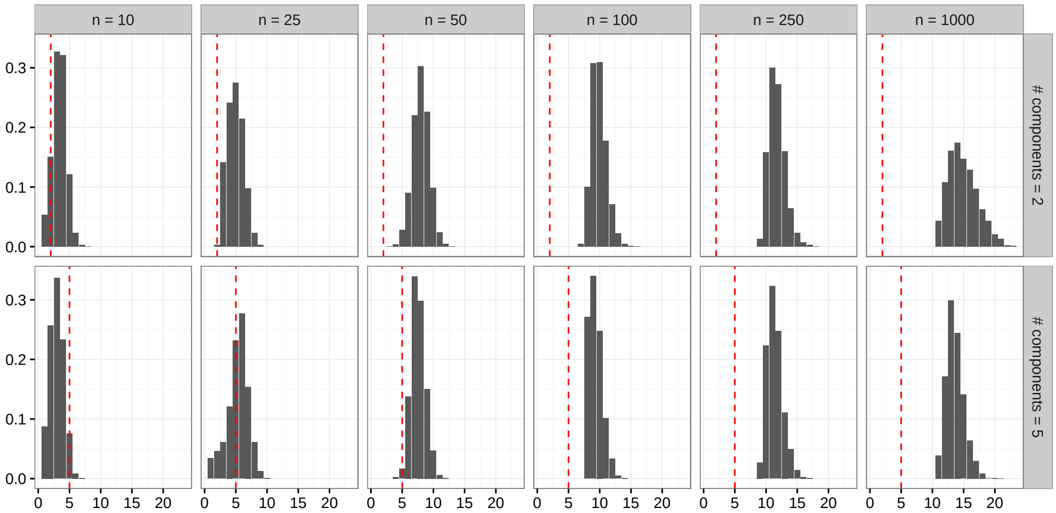

Figure 5: The posterior distribution on the number of clusters is inconsistent for a finite mixture. We generated data from a mixture $0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$ (upper panels) and $\sum_{k=-2}^{2} \frac{1}{5}\mathcal{N}(2k,1)$ (lower panels). The true K_0 is denoted by vertical dashed line. When the number of data points n increases, the posterior concentrates towards larger K values than the underlying truth.

Due to the misspecified model, it is not surprising that the DPM does not correctly identify the true number of components in a finite mixture. However, in general DPMs of normals are consistent for estimating the actual density, provided sufficient regularity of the true density [8].

5 Final Remarks

To conclude, we have motivated the use of the Dirichlet process and Pólya trees as non-parametric Bayesian priors, including some original discussion of the more technical details. Although these models are flexible, they are not perfect. For example the DP mixture model suffers from inconsistency when estimating the number of components, and the Lebesgue density of a measure generated from a Pólya tree is in general discontinuous. This discontinuity of the density can be tackled by random partitions [9], which also reduces the problem of partition dependence in Pólya trees. Finally, we note that the DPM model we have described can be extended significantly to a hierarchical model [2], to tackle the problem of grouped data, where dependencies between groups and within groups are captured by a Dirichlet process.

References

[1] Ferguson, T.S. (1974) *Prior Distributions on spaces of proability measures Ann. Statist. 2. 615-629*

[2] Teh, Y., Jordan, M., Beal, M. and Blei, D. (2006) Hierarchical Dirichlet Processes, *Journal of the American Statistical Association* Vol. 101, No. 476 (Dec., 2006), pp. 1566-1581.

[3] Orbanz, P. (2014) Lecture Notes on Bayesian Nonparametrics, `http://stat.columbia.edu/~porbanz/npb-tutorial.html`

[4] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2), 249-265.

[5] Lavine M. (1992). Some Aspects of Pólya Tree Distributions For Statistical Modelling *The annals of statistics Vol.20, No. 3, p. 1222-1235*

[6] Müller P. (2013). Non-Parametric Inference: Pólya Trees Chapter 4 `http://projecteuclid.org/download/pdfview_1/euclid.cbms/1362163749`

[7] Walker, S., Damien, P., Laud, P. and Smith, A. (1999) Bayesian Nonparametric Inference for Random Distributions and Related Functions *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3): 485-527.

[8] Miller, J. W., Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *In Advances in neural information processing systems* (pp. 199-206).

[9] Paddock, S. M., Ruggeri, F., Lavine, M. and West, M. (2003) Randomized Pólya tree models for nonparametric Bayesian inference. *Statistica Sinica*, 443-460.