Bayesian Nonparametrics

Tamás Papp, supervised by Dr Marco Battiston

1 Introduction

Bayesian nonparametrics is the study of Bayesian inference methods for nonparametric and semiparametric models — that is, models whose parameter space is infinite-dimensional or grows with the amount of data. The nonparametric Bayesian approach may be preferred due to the conceptual simplicity of the Bayesian paradigm (where all inference tools are produced by the posterior distribution) and the flexibility afforded by nonparametric models (which can avoid unverifiable assumptions imposed by parametric models).

One of the main theoretical results underpinning Bayesian statistics is de Finetti's theorem, which states that if a sequence of random variables is exchangeable — meaning the joint distribution of any finite subset of the random variables is invariant to permutation — then there exists a random parameter that renders the random variables conditionally independent. Furthermore, the joint distribution of any finite subset of the random variables can be expressed as a mixture over the random parameter. The key concept in Bayesian nonparametric statistics is to regard that parameter as infinite-dimensional, for instance being a probability distribution or a function, with the challenge being finding an appropriate prior for such objects.

Bayesian nonparametric modelling faced a surge in popularity in the late 1990s and early 2000s, as increases in computational capabilities made such methods feasible. In machine learning, this approach has enabled the study of more complex datasets in an unsupervised learning setting. Here, minimal assumptions are made about the data and the underlying latent structure (associations, categorisations or presence of certain features) is to be inferred. Unsupervised learning problems often arise in fields such as computer vision, natural language processing, and bioinformatics, in all of which the Bayesian nonparametric paradigm has been applied with particular success.

This report provides an overview of Bayesian nonparametric modelling approaches. Section 2 follows Ghosal and van der Vaart (2017) and introduces the Dirichlet process, colloquially known as the "normal distribution" of Bayesian nonparametrics, which is a measure on discrete probability measures and forms the building block for priors in the nonparametric paradigm.

The Chinese restaurant process, the predictive distribution of the Dirichlet process and the first step in constructing inference procedures, is also mentioned. Section 3 focuses on a simple mixture model which uses the Dirichlet process prior in order to cluster observations while avoiding assumptions on the number of clusters and technical complications with Bayesian inference. Following Neal (2000), ways of performing approximate inference by Markov chain Monte Carlo are mentioned in more detail. Section 4 follows Teh et al. (2006), extending the Dirichlet mixture model to a more complex, hierarchical setting where the data lies in pre-determined groups. Section 5 follows Griffiths and Ghahramani (2011) in their derivation of a nonparametric model for a generalised case of clustering where each observation can have multiple features. Finally, Section 6 offers some perspective on open problems in Bayesian nonparametrics and concludes.

2 Dirichlet process

The default prior on spaces of probability measures in Bayesian nonparametrics is the Dirichlet process. It arises as the natural generalisation of the finite-dimensional Dirichlet distribution.

Definition 1 (Dirichlet distribution)**.** We write Dirichlet$(k; \boldsymbol{\alpha})$ for the k-dimensional Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_k)$. It is supported on the k-simplex $\mathbb{S}_k = \{(x_1, \dots, x_k) : x_1, \dots, x_k \geq 0 \text{ and } \sum_{i=1}^k x_i = 1\}$ and has density

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

for $x_1, \dots, x_k \geq 0$ and $\sum_{i=1}^k x_i = 1$.

From its density, it is straightforward to deduce that the Dirichlet distribution is conjugate to the multinomial distribution. To be precise,

$$\left.\begin{aligned} \mathbf{p} &\sim \text{Dirichlet}(k; \boldsymbol{\alpha}) \\ \mathbf{N} \mid \mathbf{p} &\sim \text{Multinomial}(n, k; \mathbf{p}) \end{aligned}\right\} \implies \mathbf{p} \mid N \sim \text{Dirichlet}(k; \boldsymbol{\alpha} + \mathbf{N}).$$

The marginals of the Dirichlet are Beta-distributed, that is $p_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$, where $\alpha_0 := \sum_{i=1}^k \alpha_i$ is the total mass of the Dirichlet.

We could view a draw from a Dirichlet distribution as a discrete measure supported on finitely many points. This makes the Dirichlet a useful prior in mixture models with a known number of mixture components, but lackluster when the number of components is unknown and possibly growing with the size of the data. To adapt to such a case, we can extend the Dirichlet distribution to a Dirichlet process. The formal definition is as follows.

Definition 2 (Dirichlet process)**.** A random measure G on space $(\mathfrak{X}, \mathscr{X})$ has a Dirichlet process distribution $\mathrm{DP}(\alpha)$ with base measure α if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$(G(A_1), \ldots, G(A_k)) \;\sim\; \text{Dirichlet}\left(k; \alpha(A_1), \ldots, \alpha(A_k)\right).$$

The base measure α decomposes into total mass α_0 and mean measure G_0 (also called the base probability measure). We therefore equivalently write $DP(\alpha_0, G_0)$ when we wish to differentiate between the total mass and mean measure.

Sethuraman (1994) showed that the Dirichlet process has a "stick-breaking" representation. If $G \sim \mathrm{DP}(\alpha_0, G_0)$ then

$$G \overset{\mathrm{d}}{=} \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{1}$$

with $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$[1] and $\phi_k \overset{\mathrm{iid}}{\sim} G_0$ for all k. This GEM distribution gives rise to the intuitive name of this representation, with it dividing the "stick" of probability mass one into an infinite number of parts by breaking off lengths of size proportional to a draw from a Beta distribution. A draw from a Dirichlet process is therefore a random discrete measure with probability one.

Observations $\theta_1, \ldots \theta_n$ sampled independently from a draw of a Dirichlet process

$$\begin{aligned} \theta_1, \ldots, \theta_n \mid G \;&\sim\; G \\ G \;&\sim\; \mathrm{DP}(\alpha) \end{aligned}$$

are often called a "sample from the Dirichlet process". The Dirichlet process posterior is also a Dirichlet process,

$$G \mid \theta_1, \ldots, \theta_n \;\sim\; \mathrm{DP}\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right).$$

Viewed from a frequentist perspective, if data $\theta_1, \ldots, \theta_n$ arise from some true distribution P, then the above posterior converges in distribution to δ_P as $n \longrightarrow \infty$.

2.1 Chinese restaurant process

The joint distribution of $(\theta_1, \ldots, \theta_n)$ generated from a Dirichlet process can be described by the sequence of predictive distributions

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \;\sim\; \sum_{j=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\theta_j} + \frac{\alpha_0}{i-1+\alpha_0} G_0.$$

Though it might not be apparent at first, this implies that there are very few distinct values among any $(\theta_1, \ldots, \theta_n)$ — even if G_0 is continuous, so that if θ_i is conditionally drawn from G_0

[1] By $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$ we mean $\pi_k \overset{\mathrm{d}}{=} \beta_k \prod_{i=1}^{k-1}(1-\beta_i)$ with $\beta_i \overset{\mathrm{iid}}{\sim} \mathrm{Beta}(1, \alpha_0)$ for all i, k.

then it is almost surely distinct from previous samples. In particular, the number of distinct values is $O(\log n)$, by Ghosal and van der Vaart (2017).

A more intuitive description of this sequence of predictive distributions can be obtained by rewriting it in an altered form. If there are K distinct values among $\theta_1, \ldots, \theta_{i-1}$, we introduce auxiliary variables ϕ_k that label the distinct values and counts m_k for each such value, for $k \in \{1, \ldots, K\}$. The predictive distribution becomes

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0} \delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0. \tag{2}$$

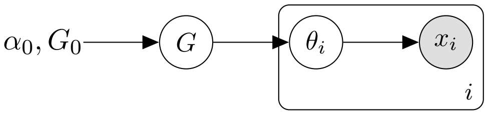
This generating process was named the “Chinese restaurant process” by Jim Pitman and Lester Dubins (Aldous, 1985). The metaphor is in reference to the seemingly infinite capacity of Chinese restaurants in San Francisco and its explanation is as follows. Customers indexed by i enter a restaurant sequentially and then choose to sit at some table, with tables being indexed by k. The first customer chooses a table at random ($\phi_1 \sim G_0$ associated to θ_1). All subsequent customers i choose to sit at previously occupied table k with probability proportional to the number of customers m_k already sitting there, or open a new table with probability proportional to α_0. If they open a new table, the number of occupied tables K is incremented by one, $\phi_K \sim G_0$ is drawn and associated to θ_i.

3 Dirichlet process mixtures

The Dirichlet process is a useful prior for mixture models where the number of mixture components is unknown or potentially growing with the amount of data. Such models are often considered when clustering data or estimating a density. By adopting a Dirichlet process prior for the mixture component parameters, unverifiable assumptions about the number of mixture components and technical complications with approximate posterior inference can be avoided. The Dirichlet process mixture model is

$$\begin{aligned} x_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim \mathrm{DP}\left(\alpha_0, G_0\right). \end{aligned} \tag{3}$$

Figure 1: DP mixture as a Bayesian network



To make the notation more precise, we use capital letters G_0, G and F for probability distributions. To focus on issues specific to the Dirichlet process, we assume F and G_0 are continuous and have associated densities $f(\cdot)$ and $g_0(\cdot)$ respectively. While priors on α_0, G_0 and on hyperparameters for F would often be introduced in practice, in similar spirit to the remark before we do not consider them.

An intuitive way of arriving at model (3) is by taking an infinite limit of a finite-dimensional mixture model. By viewing draws from a Dirichlet distribution as finite discrete measures, the following is a commonly used K-dimensional Bayesian mixture model:

$$\begin{aligned} x_i \mid z_i, \boldsymbol{\phi} &\sim F(\phi_{z_i}) \\ z_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(K; \alpha_0/K, \ldots, \alpha_0/K) \\ \phi_k &\sim G_0. \end{aligned} \tag{4}$$

In the limit of $K \longrightarrow \infty$, and associating $\theta_i = \phi_{z_i}$, this model converges to (3). The process of constructing a nonparametric model by taking the limit of a finite-dimensional one is more widely applicable in Bayesian modelling, enabling inference procedures to be deduced from the limit of appropriate conditionals.

We now focus on ways of performing inference with the Dirichlet process mixture model. Since exact computation of posterior expectations for this model are infeasible even for small sample sizes, we turn to Monte Carlo methods.

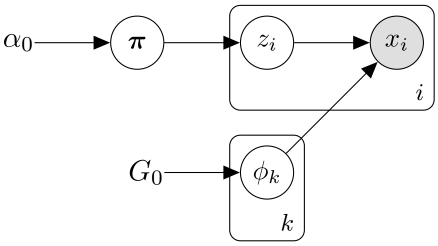
3.1 Inference

Inference in the Dirichlet process mixture model can be done by Markov chain Monte Carlo sampling from the posterior, with Neal (2000) reviewing a number of such samplers. He mentions that while it is possible to perform inference with formulation (3), there are potential issues in doing so. For instance, a Gibbs sampler derived from the Chinese restaurant process (2) (as Escobar and West (1995) do) would be rather inefficient, with the problem being that "there are often groups of observations that with high probability are associated with the same θ". Since such an algorithm only updates one θ_i at a time, it must pass through many low probability intermediate states with such groups split up, slowing down mixing severely.

Fortunately, this issue can be avoided by reformulating the model. Using the stick-breaking construction, an equivalent model to (3) is

$$
\begin{aligned}
x_i \mid z_i, \boldsymbol{\phi} &\sim F(\phi_{z_i}) \\
z_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\
\boldsymbol{\pi} &\sim \mathrm{GEM}(\alpha_0) \\
\phi_k &\sim G_0.
\end{aligned}
\tag{5}
$$

Figure 2: Reformulated DP mixture as a Bayesian network



In this reformulated model, the index variable z_i represents the cluster or latent class associated to observation x_i, so that $\theta_i = \phi_{z_i}$ as in (4). Note that the specific labels assigned to each z_i are of no importance, as they simply label whether observations are in the same group or not. We are then free to, for instance, force z_i to only take the smallest values in the natural numbers, as in the "no gaps" sampler of MacEachern and Müller (1998).

By restricting to samplers with the mixing proportions $\boldsymbol{\pi}$ integrated out, the efficiency issue with formulation (3) is avoided. Following Neal (2000), we therefore use formulation (5) throughout the remainder of this section and list several algorithms from the same paper. These are the two Gibbs samplers for the conjugate case and the author's three noteworthy contributions for the non-conjugate case: a Metropolis-Hastings sampler, a partial Gibbs and Metropolis-Hastings sampler and an augmented Gibbs sampler.

3.1.1 Samplers for the conjugate case

If F and G_0 are conjugate, it is possible to directly use Gibbs samplers. The relevant quantities are the marginals of $P(\mathbf{z}, \boldsymbol{\phi}|\mathbf{x})$, where the parameters in use are $\mathbf{z} = (z_1, \ldots, z_n)$, $\boldsymbol{\phi} = (\phi_k : k \in \mathbf{z})$ and the data is $\mathbf{x} = (x_1, \ldots, x_n)$. We write $\mathbf{z}^{-i}$ for the vector $\mathbf{z}$ without entry i and m_k^{-i} for the number of entries of $\mathbf{z}^{-i}$ that are equal to k. The marginals can be computed from the Chinese restaurant process and Bayes' rule, producing a "full" Gibbs sampler.

Algorithm 1 (Full Gibbs sampler)**.** The state of the Markov chain is $\mathbf{z} = (z_1, \ldots, z_n)$ and $\boldsymbol{\phi} = (\phi_k : k \in \mathbf{z})$. Repeat:

- For $i = 1, \dots, n$: Sample z_i from conditional

  $$\begin{aligned} \text{If } k \in \mathbf{z}^{-i}\text{: } P(z_i = k \mid \mathbf{z}^{-i}, \mathbf{x}, \boldsymbol{\phi}) &= b m_k^{-i} f(x_i, \phi_k) \\ P(z_i \notin \mathbf{z}^{-i} \mid \mathbf{z}^{-i}, \mathbf{x}, \boldsymbol{\phi}) &= b\alpha_0 \int f(x_i, \phi) g_0(\phi) \mathrm{d}\phi, \end{aligned}$$

  where b is the appropriate normalisation constant. If $z_i \notin \mathbf{z}^{-i}$, sample new ϕ_{z_i} from density $\propto f(x_i, \phi_{z_i}) g_0(\phi_{z_i})$ and add to $\boldsymbol{\phi}$. Discard any ϕ_k not associated with an observation.

- For all $k \in \mathbf{z}$: Update ϕ_k from $\phi_k \mid \{x_j : z_j = k\}$, that is with density $\propto \prod_{j:z_j=k} f(x_j|\phi_k) g_0(\phi_k)$.

If it is possible to completely integrate out the cluster parameters $\boldsymbol{\phi}$ analytically, and the exact values of the cluster parameters are of no interest, a simplified Gibbs sampler on the marginals of $P(\mathbf{z}|\mathbf{x})$ can be used.

Algorithm 2 (Simplified Gibbs sampler)**.** The state of the Markov chain is $\mathbf{z} = (z_1, \dots, z_n)$. Repeat: For $i = 1, \dots, n$, sample z_i from conditional

$$\begin{aligned} \text{If } k \in \mathbf{z}^{-i}\text{: } P(z_i = k \mid \mathbf{z}^{-i}, \mathbf{x}) &= b m_k^{-i} \int \prod_{j\,:\,z_j=k} f(x_j, \phi) g_0(\phi) \mathrm{d}\phi \\ P(z_i \notin \mathbf{z}^{-i} \mid \mathbf{z}^{-i}, \mathbf{x}) &= b\alpha_0 \int f(x_i, \phi) g_0(\phi) \mathrm{d}\phi, \end{aligned}$$

where b is the appropriate normalisation constant.

3.1.2 Samplers for the non-conjugate case

In the non-conjugate case, the Gibbs samplers can usually not be applied as the relevant integrals are analytically intractable. Neal (2000) therefore suggests alternate samplers, the first being a Metropolis-Hastings algorithm proposing from the conditional prior of z_i:

$$\begin{aligned} \text{If } k \in \mathbf{z}^{-i}\text{: } P(z_i = k \mid \mathbf{z}^{-i}) &= \frac{m_k^{-i}}{n - 1 + \alpha_0} \\ P(z_i \notin \mathbf{z}^{-i} \mid \mathbf{z}^{-i}) &= \frac{\alpha_0}{n - 1 + \alpha_0}. \end{aligned} \tag{6}$$

This is a symmetric proposal, as the distribution depends strictly on $\mathbf{z}^{-i}$, which remains unchanged. The Metropolis-Hastings acceptance ratio thus reduces to a ratio of likelihoods. To speed up mixing, an additional Gibbs update step of sampling ϕ_k within each cluster is added.

Algorithm 3 (Metropolis-Hastings sampler)**.** The state of the Markov chain is $\mathbf{z} = (z_1, \dots, z_n)$ and $\boldsymbol{\phi} = (\phi_k : k \in \mathbf{z})$. Repeat:

- For $i = 1, \dots, n$, repeat R times update of z_i: Propose z_i^* from conditional (6), if not among $\mathbf{z}^{-i}$ also sample $\phi_{z_i^*} \sim G_0$. Accept with probability $\min\left[1, f(x_i, \phi_{z_i^*})/f(x_i, \phi_{z_i})\right]$. Otherwise, keep z_i the same.

- For all $k \in \mathbf{z}$: Update ϕ_k from $\phi_k \mid \{x_i : z_i = k\}$ or perform any other update that leaves this distribution invariant.

Since in practice relatively low values of $\alpha_0 \simeq 1$ are used, representing equal weighting in the first Chinese restaurant process allocation, this sampler would create a new cluster relatively infrequently. Neal (2000) thus conjectures that a sampler that creates a new component more often in the step of updating z_i might be more efficient. This can be forced by using an altered proposal distribution, which proposes a strictly new component whenever z_i is not a singleton, that is when there is a $j \neq i$ such that $z_j = z_i$ (observation x_i is not in a class of its own). Direct usage of this chain, however, results in an inefficient sampler: when moving an observation from one existing group to another, the sampler must pass through an unlikely state where that observation is a singleton. This is the exact same issue as when Gibbs sampling in formulation (3). We can make such changes more likely, and thus improve the efficiency of the sampler, by performing partial Gibbs updates on the observations that are neither singletons nor allowed to become singletons. This creates a valid, ergodic Markov chain, as there is a non-zero probability of visiting every possible state.

Algorithm 4 (Partial Gibbs sampler)**.** The state of the Markov chain is $\mathbf{z} = (z_1, \ldots, z_n)$ and $\boldsymbol{\phi} = (\phi_k : k \in \mathbf{z})$. Repeat:

- For $i = 1, \ldots, n$, update z_i by Metropolis-Hastings:
  - If z_i is not a singleton, create new z_i^* and sample $\phi_{z_i^*} \sim G_0$. Accept z_i^* with probability $\min\left[1, \alpha_0 f(x_i, \phi_{z_i^*})/(n-1) f(x_i, \phi_{z_i})\right]$.
  - If z_i is a singleton, propose z_i^* from $\mathbf{z}^{-i}$ uniformly. Accept z_i^* with probability $\min\left[1, (n-1) f(x_i, \phi_{z_i^*})/\alpha_0 f(x_i, \phi_{z_i})\right]$.
  - Otherwise, keep z_i the same.
- For $i = 1, \ldots, n$, update z_i by Gibbs sampling: If z_i is not a singleton, replace with new value $k \in \mathbf{z}^{-i}$ from

$$P\left(z_i = k \mid \mathbf{z}^{-i}, \mathbf{x}, \boldsymbol{\phi}, z_i \in \mathbf{z}^{-i}\right) \ \propto\ m_k^{-i} f(x_i, \phi_k).$$

  Otherwise, keep z_i the same.
- For all $k \in \mathbf{z}$: Update ϕ_k from $\phi_k \mid \{x_i : z_i = k\}$ or perform any other update that leaves this distribution invariant.

An alternative approach to Metropolis-Hastings updates can also be taken to allow inference in non-conjugate models. By augmenting the sample space with auxiliary parameters, we can sample from the augmented distribution and then discard them to sample from the target distribution. When sampling z_i from the second expression in (6), we instead introduce M auxiliary variables sampled from G_0 to represent possible values z_i can take.

Algorithm 5 (Augmented Gibbs sampler)**.** The state of the Markov chain is $\mathbf{z} = (z_1, \ldots, z_n)$ and $\boldsymbol{\phi} = (\phi_k : k \in \mathbf{z})$. Repeat:

- For $i = 1, \ldots, n$, update z_i: Let K^{-i} count the distinct values in $\mathbf{z}^{-i}$. Relabel K^{-i} distinct values in $\mathbf{z}^{-i}$ by $\{1, \ldots, K^{-i}\}$. Sample $\phi_k \sim G_0$ for $K^{-i} + 1 < k \leq K^{-i} + M$. If z_i is not a singleton, resample $\phi_{K^{-i}+1} \sim G_0$. Draw new value k from

$$P\left(z_i = k \mid \mathbf{z}^{-i}, x_i, \phi_1, \ldots, \phi_h\right) \propto \begin{cases} m_k^{-i} f(x_i, \phi_k) & \text{for } 1 \leq k \leq K^{-i}, \\ \\ \dfrac{\alpha}{M} f(x_i, \phi_k) & \text{for } K^{-i} < k \leq K^{-i} + M. \end{cases}$$

  Discard the ϕ_k not associated with any observation.

- For all $k \in \mathbf{z}$: Update ϕ_k from $\phi_k \mid \{x_i : z_i = k\}$ or perform any other update that leaves this distribution invariant.

As for choosing between the three algorithms for the non-conjugate case, the author suggests they would dominate each other in different scenarios, based on the specific characteristics of the problem at hand. Numerical results show that the augmented Gibbs sampler has the potential to achieve significantly less correlated samples per iteration as the number of auxiliary variables M is increased, at an added cost per iteration.

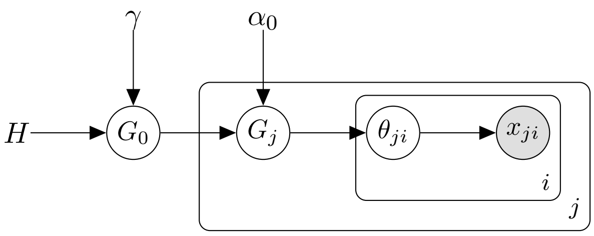
4 Hierarchical Dirichlet processes

If Dirichlet process are natural priors in nonparametric Bayesian clustering, hierarchical Dirichlet processes are the extension when clustering data with a hierarchical structure. Teh et al. (2006) consider such a setting. They assume observations lie in pre-determined groups, with observations in each group being drawn from a mixture model, allowing the groups to share mixture components. They offer two situations of interest as motivation. One is in the field of genetics, where identifying shared haplotypes among the genetic structure of groups of humans could shed light on the migration patterns of early humans. Another is in the field of information retrieval, where articles in journals could be analysed to see to what extent themes ("extreme value statistics", "survival analysis" etc.) are shared across the journals.

They propose modelling such data using a mixture model with a hierarchical Dirichlet process prior, guided by the natural clustering property of the Dirichlet process and the fact that it forces atoms to be shared among groups due to it being almost surely discrete. Labeling data x_{ji} to belong to observation i in group j, the model under their consideration is

$$\begin{aligned} x_{ji} \mid \theta_{ji} &\sim F(\theta_{ji}) \\ \theta_{ji} \mid G_j &\sim G_j \\ G_j &\sim \mathrm{DP}\left(\alpha_0, G_0\right) \\ G_0 &\sim \mathrm{DP}\left(\gamma, H\right). \end{aligned} \tag{7}$$

Figure 3: Hierarchical DP mixture as a Bayesian network



For simplicity, we assume that F, G_0 and H are continuous distributions, with density functions $f(\cdot), g_0(\cdot)$ and $h(\cdot)$.

Similarly to the Dirichlet process mixture model, the hierarchical model can be derived as the limit of a finite-dimensional model. MacKay and Peto (1995) propose the following hierarchical model for natural language processing:

$$\begin{aligned} x_{ji} \mid z_{ji}, \boldsymbol{\phi} &\sim F(\phi_{z_{ji}}) \\ z_{ji} \mid \boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \boldsymbol{\pi}_j \mid \boldsymbol{\beta} &\sim \text{Dirichlet}\left(|\boldsymbol{\beta}|; \alpha_0 \boldsymbol{\beta}\right) \\ \boldsymbol{\beta} &\sim \text{Dirichlet}(L; \gamma/L, \ldots, \gamma/L), \\ \phi_k &\sim H. \end{aligned}$$

As $L \longrightarrow \infty$, Teh et al. (2006) show this model converges to (7).

4.1 Chinese restaurant franchise process

Teh et al. (2006) derive the predictive distribution for the hierarchical Dirichlet process prior, terming it the "Chinese restaurant franchise" process, in complete analogy to and in extension of the Chinese restaurant process. The metaphor here is as follows. We have a chain of restaurants indexed by j, with customers indexed by i entering one restaurant in the chain and sitting down at a table, with tables being indexed by t. All restaurants serve a common menu of dishes, with each dish being indexed by k. Further notation is made more precise in the sequel.

Table 1: Notation used in CRF

Notation	*Description*
ϕ_k	Dishes from the menu
ψ_{jt}	Table-specific choice of dishes
m_{jk}	The number of tables in restaurant j serving dish k
n_{jtk}	The number of customers in restaurant j at table t having dish k
t_{ji}	Index of table associated to customer i in restaurant j, $\theta_{ji} = \psi_{jt_{ji}}$
k_{jt}	Index of dish associated to table t in restaurant j, $\psi_{jt} = \phi_{k_{jt}}$
z_{ji}	Index of dish associated to each customer, $z_{ji} = k_{jt_{ji}}$
Dot instead of index	Sum over that index, e.g. $m_{\cdot k}$ is the number of tables serving dish k

As in the Chinese restaurant process, customers enter a restaurant and either choose tables already occupied by customers at that restaurant with probability proportional to the number of customers sitting there or open a new table with probability proportional to α_0,

$$\theta_{ji} \mid \theta_{j1}, \dots, \theta_{j,i-1}, G_0 \ \sim \ \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0. \tag{8}$$

We call this the "table level" description of the Chinese restaurant franchise. Whenever a new table is opened, $m_{j\cdot}$ is incremented by one, $\psi_{jm_{j\cdot}} \sim G_0$ is drawn and we set $t_{ji} = m_{j\cdot}$ and $\theta_{ji} = \phi_{jm_{j\cdot}}$.

The innovation brought forward by the Chinese restaurant happens when a customer opens a new table: a dish is also ordered for the whole table, to be shared by all future customers at that table. This dish is now chosen from a global level Chinese restaurant process

$$\psi_{jt} \mid \psi_{11}, \dots, \psi_{j,t-1} \ \sim \ \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H, \tag{9}$$

where K is the total number of distinct dishes. We call this the "dish level" description of the Chinese restaurant franchise. Whenever a new dish is ordered, K is incremented by one and $\phi_k \sim H$ is drawn and associated to ψ_{jt}. Multiple tables in the same restaurant can serve the same dish.

We are ultimately interested in which mixture component is associated to each observation — that is which dish z_{ji} each customer eats — and we can derive that from the association between table and dish (9) and the association between customer and table (8). We therefore use both equations when performing inference.

4.2 Inference

The posterior distribution in the hierarchical Dirichlet process model is analytically intractable, so we sample it by Markov Chain Monte Carlo. By drawing insight from the Dirichlet mixture model inference paper of Neal (2000), samplers derived from the Chinese restaurant franchise as-is would be inefficient due to having to pass through many low-probability intermediate states. This is solved by sampling the index variables t_{ji} and k_{jt} instead, which is the approach taken by Teh et al. (2006).

We mention the first Gibbs sampler proposed by Teh et al. (2006) for the hierarchical Dirichlet process model, which uses the Chinese restaurant franchise reformulated in terms of the index variables. F and H are assumed to be conjugate, so this algorithm is in the spirit of the simplified Gibbs sampler (Algorithm 2) for the Dirichlet mixture model. We can therefore integrate the mixture component parameters $\boldsymbol{\phi}$ out and the only quantities of interest are the marginals of $P(\mathbf{t},\mathbf{k}|\mathbf{x})$, where $\mathbf{x} = (x_{ji} : \text{ all } j,i)$, $\mathbf{t} = (t_{ji} : \text{ all } j,i)$ and $\mathbf{k} = (k_{jt} : \text{ all } j,t)$, with index t running only over tables in use.

The complex interactions make the notation used for this algorithm somewhat involved. Define $\mathbf{x}_{jt} = (x_{ji} : \text{ all } j,i \text{ such that } t_{ji} = t)$. Placing a negative superscript on a count or set means the variable corresponding to the set or count is removed, for instance $n_{jt\cdot}^{-ji}$ is count $n_{jt\cdot}$ if customer i in restaurant j were removed. The conditional likelihood of x_{ji} under mixture component k given all data items except x_{ji} is labelled as

$$f_k^{-x_{ji}}(x_{ji}) \;:=\; f(x_{ji}|z_{ji} = k, \mathbf{x}^{-ji})$$

and can be evaluated by applying Bayes' rule and then integrating over parameters $\boldsymbol{\phi}$. The joint conditional likelihood of all data items in $\mathbf{x}_{jt}$ under mixture component k except $\mathbf{x}_{jt}$ is labelled $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$, and can be computed similarly.

To sample $\mathbf{t}$, we use posterior marginals derived from the table-level CRF (8)

$$\begin{aligned} \text{If } t \in \mathbf{t}^{-ji}\text{: } P(t_{ji} = t \mid \mathbf{t}^{-ji}, \mathbf{k}, \mathbf{x}) \;&=\; b n_{jt\cdot}^{-ji} f_{k_{jt}}^{-x_{ji}}(x_{ji}) \\ P(t_{ji} \notin \mathbf{t}^{-ji} \mid \mathbf{t}^{-ji}, \mathbf{k}, \mathbf{x}) \;&=\; b\alpha_0 f^{-x_{ji}}(x_{ji}|\mathbf{t}^{-ji}, t_{ji} \notin \mathbf{t}^{-ji}, \mathbf{k}), \end{aligned} \tag{10}$$

where b is the appropriate normalisation constant and, by the dish-level CRF (9) and summing over all mixture components, the conditional likelihood of x_{ji} for a new value of t_{ji} is

$$f^{-x_{ji}}(x_{ji}|\mathbf{t}^{-ji}, t_{ji} \notin \mathbf{t}^{-ji}, \mathbf{k}) = \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} f_k^{-x_{ji}}(x_{ji}) + \frac{\gamma}{m_{\cdot\cdot} + \gamma} f(x_{ji}).$$

If a new value $t_{ji} = t$ is sampled, then we also sample a new k_{jt} from

$$\begin{aligned} \text{If } k \in \mathbf{k}^{-jt}\text{: } P(k_{jt} = k \mid \mathbf{t}, \mathbf{k}^{-jt}, \mathbf{x}) \;&=\; b m_{\cdot k} f_k^{-x_{ji}}(x_{ji}) \\ P(k_{jt} \notin \mathbf{k}^{-jt} \mid \mathbf{t}, \mathbf{k}^{-jt}, \mathbf{x}) \;&=\; b\gamma f(x_{ji}), \end{aligned} \tag{11}$$

where b is the appropriate normalisation constant and we have used Bayes' rule and (9).

When sampling $\mathbf{k}$, note that changing k_{jt} changes the component membership of all customers at table t. The likelihood when setting $k_{jt} = k$ then becomes $f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt})$ and we sample the conditionals of $\mathbf{k}$ from

$$\begin{aligned} \text{If } k \in \mathbf{k}^{-jt}\text{: } P(k_{jt} = k \mid \mathbf{t}, \mathbf{k}^{-jt}, \mathbf{x}) &= b m_{\cdot k}^{-jt} f_k^{-\mathbf{x}_{jt}}(\mathbf{x}_{jt}) \\ P(k_{jt} \notin \mathbf{k}^{-jt} \mid \mathbf{t}, \mathbf{k}^{-jt}, \mathbf{x}) &= b\gamma f(\mathbf{x}_{jt}). \end{aligned} \tag{12}$$

Combining these sampling steps together, the following Gibbs sampler is obtained.

Algorithm 6 (Gibbs sampler)**.** The state of the Markov chain is $\mathbf{t} = (t_{j,i} : \text{all } j, i)$ and $\mathbf{k} = (k_{j,t} : \text{all } j, \text{ all } t \in \mathbf{t})$. Repeat:

- For all j, i: sample t_{ji} from (10). If sample $t_{ji} = t \notin \mathbf{t}^{-ji}$, also sample new k_{jt} from (11). If a table becomes unoccupied, then delete the corresponding k_{jt} and any unallocated mixture components k.
- For all j, all $t \in \mathbf{t}$: sample k_{jt} from (12).

Given that the clusters induced by the values z_{ji} are the main interest, one might wonder whether it is possible to obtain samplers for these indices directly, instead of having to reconstruct them indirectly from t_{ji} and k_{jt}. Teh et al. (2006) propose such a Gibbs sampler as well, which samples z_{ji} and m_{jk} using an augmented representation and has more straightforward bookkeeping. Note that Algorithm 6 updates multiple observations at a time whenever $\mathbf{k}$ is updated, whereas this direct assignment procedure updates only one membership at a time, potentially making it less efficient.

In the case of F and H being non-conjugate, variable $\boldsymbol{\phi}$ cannot be integrated out, so Algorithm 6 is not applicable. Metropolis-Hastings proposals or augmented representations as in Algorithms 3-5 of Section 3 could be applied to construct inference procedures in such a situation.

5 Indian buffet process

Both the Dirichlet process and hierarchical Dirichlet process mixture models assume there is a single latent feature associated to each observation. In contrast, many unsupervised learning problems represent each observation as having multiple features. Consider the case of an image recognition problem, where the image could contain several objects at several locations. While a single factor describing each image could reasonably be used, much more could be inferred by using a binary vector of features, each entry representing whether or not a specific object is present in the image. Using a Dirichlet process prior on parameters describing the latent structure would therefore be unsuitable in such a problem. At the same time, taking a nonparametric Bayesian approach is preferred, as it affords flexibility.

Griffiths and Ghahramani (2011) consider the setting of representing objects with an infinite number of features and extend the nonparametric Bayesian approach to this case. They assume objects are exchangeable and represent the sequence of features as an infinite binary vector associated to each object, which will have a finite number on non-zero entries. They derive a nonparametric prior for the matrix of features by taking the limit of a finite-dimensional beta-Bernoulli model

$$\begin{aligned} z_{ik} \mid \pi_k &\sim \text{Bernoulli}(\pi_k) \\ \pi_k \mid \alpha &\sim \text{Beta}\left(\frac{\alpha}{K}, 1\right), \end{aligned}$$

where the index $k = 1 \ldots, K$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ is the feature vector associated to observation i. This induces a probability distribution on $N \times K$ binary matrices $\mathbf{Z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_N^T)^T$. Taking the limit as $K \longrightarrow \infty$ of the above model, noting the expected number of non-zero entries in the matrix $\mathbf{Z}$ remains bounded above, the distribution converges to one over binary matrices of N rows of infinite feature vectors.

Similar to how partitions of objects form equivalence classes when clustering, it is useful to have equivalence classes of $N \times \infty$ binary matrices $\mathbf{Z}$. We define the left-ordered form of such a matrix by ordering its columns from left to right by the magnitude of the binary number expressed by that column, taking the first row as the most significant bit. Matrices $\mathbf{Z}$ with the same left-ordered form are put in the same equivalence class $[\mathbf{Z}]$. These equivalence classes are in bijection with the equivalence classes under exchangeability of features. Adjusting for the cardinality of each $[\mathbf{Z}]$, the distribution obtained by the infinite limit induces one over equivalence classes (equation 14 in Griffiths and Ghahramani (2011) gives an explicit formula).

This distribution is also recovered by a stochastic process Griffiths and Ghahramani (2011) call the "Indian buffet process", by (a geographically adjusted) analogy with the Chinese restaurant process. The metaphor refers to the many Indian restaurants in London that have a seemingly infinite number of dishes in their lunchtime buffets, which matches the infinite feature vector in the model.

The description of the Indian buffet process is as follows. N customers sequentially enter a restaurant with infinitely many dishes arranged in a line, which they sample starting from the same end. The first customer samples Poisson(α) consecutive dishes, then stops. The i^{th} customer moves along the buffet, sampling any previously chosen dish k with probability m_k/i, where m_k is the number of previous customers who sampled that dish. After running out of previously sampled dishes, the customer samples Poisson(α/i) new dishes consecutively, then stops. While customers are not exchangeable under this process, the distribution it induces for the left-ordered form equivalence classes has exactly the same form as the one for the infinite model defined above.

It is also possible to define an exchangeable Indian buffet process. We define the history of dish k at customer i by $(z_{1k}, \ldots, z_{i-1,k})$. In the exchangeable Indian buffet process, the only difference is that customer i makes a single decision for all K_h dishes with the same history h, choosing Binomial$(m_h/i, K_h)$ of them, starting from the left.

The exchangeable Indian buffet process allows straightforward computation of marginals of $\mathbf{Z}$. As customers are exchangeable, we can make the i^{th} object to correspond to the last customer in the buffet, so the marginals are

$$P(z_{ik} = 1 \mid \mathbf{z}_k^{-i}) \;=\; \frac{m_k^{-i}}{N}.$$

If we additionally had data $\mathbf{x}$ and specified a likelihood $f(\mathbf{x}|\mathbf{Z})$, one might assume this expression would allow us to perform inference via Gibbs sampling. Iteratively sampling from the marginals

$$P(z_{ik} = 1 \mid \mathbf{Z}^{-ik}, \mathbf{x}) \;\propto\; \frac{m_k^{-i}}{N} f(\mathbf{x} \mid \mathbf{Z}),$$

however, does not result in a valid Markov chain, as the order in which variables are sampled depends on the state of the chain. Regardless, Griffiths and Ghahramani (2011) successfully use this heuristic Gibbs-like strategy in their numerical demonstrations.

An insightful look into the fundamentals of the Indian buffet process is given by Thibaux and Jordan (2007). They establish that the Beta process is the underlying de Finetti mixing distribution that induces the Indian buffet process, similar to how the Dirichlet process is the underlying de Finetti mixing distribution for the Chinese restaurant process.

6 Conclusions and future work

This report aimed to provide insight on theoretical, modelling and computational aspects in Bayesian nonparametrics. We have focused on three of the most relevant papers in the area, that is the works of Neal (2000), Teh et al. (2006) and Griffiths and Ghahramani (2011). One common theme is that models which represent observations in terms of a finite set of latent features can be extended to more flexible, infinite-feature models, while still being able to perform approximate inference.

We now turn to listing potential directions for future work in Bayesian nonparametrics, the first of which is theoretical. Jordan (2011), in his poll of forty-eight eminent statisticians on open problems in Bayesian statistics, places Bayesian nonparametrics as the fifth most important topic for research in the field. Frequentist evaluation of Bayesian nonparametric models is mentioned as a common concern by many of the respondents, though the book of Ghosal and van der Vaart (2017) partially reconciles this. Prior specification and identifiability are also listed as concerns, in particular overcoming arbitrary hyperparameter specification.

Continuing on the theoretical side, there is a great interest in finding and cataloguing priors for various classes of problems where Bayesian nonparametrics could be applied. Similar to how Dirichlet processes are the go-to prior when performing clustering and Gaussian processes the default prior when performing regression, a new prior on infinite-dimensional objects could broaden the applicability of nonparametric Bayesian modelling.

From a modelling perspective, the strategy of taking the limit of a finite model could be applied to give rise to all three of the nonparametric models considered Sections 3-5. Griffiths and Ghahramani (2011) thus anticipate this could be successfully applied in other unsupervised learning problems, with the open problem being finding the next impactful model derived in such a way.

Finally, although increases in computational power have enabled inference when using Bayesian nonparametrics, there is still a significant computational overhead compared to the parametric setting. From a computational perspective, there is therefore interest in finding new, more efficient inference procedures. In cases where the number of observations is particularly large, adapting cutting edge Monte Carlo methods designed for big data to the nonparametric setting would be a promising direction for future work.

References

David J. Aldous. Exchangeability and Related Topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, 1985.

Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

Thomas L. Griffiths and Zoubin Ghahramani. The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Micheal I. Jordan. Open Problems in Bayesian Statistics. *ISBA Bulletin*, 18(1):1–4, March 2011.

Steven N. MacEachern and Peter Müller. Estimating Mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

David J. C. MacKay and Linda C. Bauman Peto. A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, 1(3):289–307, 1995.

Radford M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Jayaram Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Romain Thibaux and Michael I. Jordan. Hierarchical Beta Processes and the Indian Buffet Process. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2, pages 564–571, 2007.