Lectures on Bayesian nonparametrics: modeling, algorithms and some theory

VIASM summer school in Hanoi, August 2015

XuanLong Nguyen
University of Michigan

Abstract

This is an elementary introduction to fundamental concepts of Bayesian nonparametrics, with an emphasis on the modeling and inference using Dirichlet process mixtures and their extensions. I draw partially from the materials of a graduate course jointly taught by Professor Jayaram Sethuraman and myself in Fall 2014 at the University of Michigan.

Bayesian nonparametrics is a field which aims to create inference algorithms and statistical models, whose complexity may grow as data size increases. It is an area where probabilists, theoretical and applied statisticians, and machine learning researchers all make meaningful contacts and contributions. With the following choice of materials I hope to take the audience with modest statistical background to a vantage point from where one may begin to see a rich and sweeping panorama of a beautiful area of modern statistics.

1 Statistical inference

The main goal of statistics and machine learning is to make inference about some unknown quantity based on observed data. For this purpose, one obtains data X with distribution depending on the unknown quantity, to be denoted by parameter θ. In classical statistics, this distribution is completely specified except for θ. Based on X, one makes statements about θ in some optimal way. This is the basic idea of statistical inference. Machine learning originally approached the inference problem through the lense of the computer science field of artificial intelligence. It was eventually realized that machine learning was built on the same theoretical foundation of mathematical statistics, in addition to carrying a strong emphasis on the development of algorithms. Then as now, the focus on applications and methods among the two fields may be at times convergent and at times complementary.

A simple example A well-known example of statistical inference is to estimate the mean parameter θ from n-iid sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, assuming that the distribution of X given θ is the normal distribution $N(\theta, 1)$:

$$P(\boldsymbol{X}|\theta) = \frac{1}{(2\pi)^{n/2}} \exp -\left\{ \frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2 \right\}.$$

A popular estimate for θ (arising from either the maximum likelihood estimation (MLE) technique, or the method of moments) takes the form $\hat{\theta} = \overline{X}$. This method of inference is also called the frequentist method.

Bayesian inference The Bayesian paradigm says that we do not start from total ignorance. We always have some information about θ, which can be summarized by a probability distribution for θ. Technically, θ is treated as a random variable, and the a priori information about θ is given by a prior distribution, $\pi(\theta)$. The joint density of the data and the parameter is

$$\mathbb{P}(\boldsymbol{X}, \theta) = P(X_1, \ldots, X_n|\theta)\pi(\theta),$$

from which we deduce via Bayes' rule the conditional distribution of θ given data $\boldsymbol{X}$ — this distribution is also called the posterior distribution:

$$\mathbb{P}(\theta|\boldsymbol{X}) \propto P(X_1, \ldots, X_n|\theta)\pi(\theta).$$

In the equation above $\propto$ denotes "proportional", because we have omitted normalizing constant $\mathbb{P}(X)$ in the denominator, which does not change with θ appearing in the left hand side.

For instance, we may choose a normal prior: $\pi(\theta) = N(\mu, \tau^2)$. The posterior distribution, namely the conditional distribution of θ given $\boldsymbol{X}$, turns out to be (again) normal with mean μ_n and variance τ_n^2, where

$$\mu_n = \frac{n\overline{X} + \mu/\tau^2}{n + 1/\tau^2},\ \tau_n^2 = \frac{1}{n + 1/\tau^2}.$$

This calculation can be easily done by appealing to Bayes' rule.

The nice thing about the Bayesian paradigm is that we will be able to provide answers to other kinds of queries concerning θ once the posterior distribution has been obtained. For point estimation of θ, we can use the posterior expectation μ_n as the estimate, which minimizes the posterior squared error loss. It is clear that as the sample size n gets large, both Bayesian estimate μ_n and the frequentist answer $\hat{\theta}$ coincides. Moreover, the posterior variance τ_n vanishes, as the role of prior parameter τ is diminished. This is a desirable feature in Bayesian inference: the conclusion from more data should eventually overwhelm that of prior knowledge. The textbook by Robert [2007] is a good entry point to Bayesian statistics.

2 An example of discrete data

We introduce another simple example, this time for discrete data, that signals the appearance of the Dirichlet distribution, a key tool in Bayesian statistics in general, and Bayesian nonparametrics in particular.

Consider a finite sample space $\mathcal{X} = \{1, 2, \ldots, k\}$. The data will be elements of $\mathcal{X}$, and we are concerned with their underlying distribution. Thus, let Θ denote the set of probability measures (distributions) on $\mathcal{X}$, that is,

$$\Theta = \mathscr{P}(\mathcal{X}) := \Delta^{k-1} = \{(\theta_1, \ldots, \theta_k) : \theta_i \geq 0, i = 1, \ldots, k, \sum_{i=1}^{k} \theta_i = 1\}.$$

As before, assume that given θ, data $\boldsymbol{X} = (X_1, \ldots, X_n)$ represents a n-iid sample such that $P(X_i = j|\theta) = \theta_j$. Then the probability mass function is that of a multinomial distribution

$$P(X_1, \ldots, X_n|\theta) = \prod_{j=1}^{k} \theta_j^{N_j},$$

where N_j is a count statistic: $N_j = \sum_{i=1}^{n} \mathbb{I}(X_i = j)$. A frequentist estimate such as the MLE gives $\hat{\theta}_j = N_j/n$.

On the other hand, the Bayesian approach involves placing a prior distribution for θ. A common choice for the probability simplex Δ^{k-1} is the Dirichlet distribution.

Dirichlet distribution The Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha})$ with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ has the density

$$h(\theta_1, \ldots, \theta_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{\alpha_j - 1}, \ \alpha_j \geq 0, j = 1, \ldots, k, \sum \alpha_j > 0.$$

Γ in the above display denotes a Gamma function. Actually h is the pdf of $(\theta_1, \ldots, \theta_{k-1})$ due to the linear constraint with θ_k. It is also convenient to view the Dirichlet density as being proportional to the monomial form $h \propto \prod \theta_j^{\alpha_j}$, since this is the only part of the density formula that varies with $\theta_1, \ldots, \theta_k$.

We can also construct the Dirichlet distribution in the following way, via Gamma random variables. Let $Z_1, \ldots, Z_k$ be independent Gamma random variables with $Z_j \sim G(\alpha_j, 1), j = 1, \ldots, k$. Let $V = \sum_{j=1}^k Z_j$. Then the vector

$$U = (U_1, \ldots, U_k) := (Z_1/V, \ldots, Z_k/V)$$

is distributed by $\mathcal{D}(\boldsymbol{\alpha})$. Moreover, we can show that $U \perp V$. (These facts can be easily obtained by undergraduate level probability — directly calculating the joint density of (U, V) via a change of variable formula).

Return to our Bayesian inference treatment, where θ is endowed with the Dirichlet prior. Applying the Bayes' rule, the posterior distribution for θ takes the form:

$$\pi(\theta | X_1, \ldots, X_n) \propto P(X_1, \ldots, X_n | \theta) h(\theta_1, \ldots, \theta_k) \propto \prod_{j=1}^k \theta_j^{N_j + \alpha_j - 1}.$$

This is nothing but a Dirichlet distribution with parameters $(\alpha_j + N_j)$. The posterior mean takes the form $\hat{\theta}_j = \frac{N_j + \alpha_j}{n + \sum_{j=1}^k \alpha_j}$ which combines both a priori information via the $\boldsymbol{\alpha}$, and the information provided by the data $\boldsymbol{X}$ via count statistics N_j.

3 Clustering problem and mixture models

Clustering is a quintessential machine learning problem. It originated from classical statistics, but has become one of the main focuses of machine learning researchers, who are obsessed with clustering images, text documents, genes, cats, and people, etc.

The problem of clustering is often vaguely formulated as follows: given n data points $X_1, \ldots, X_n$ residing in some space, say $\mathbb{R}^d$, how do one subdivide these data into a number of clusters of points, in a way so that the data points belong to the same cluster are more similar than those from different clusters. A popular method is called the k-means algorithm, which is a simple and fast procedure for obtaining k clusters for a given number of k, but there is only limited theoretical basis for such an algorithm.

To provide a firm mathematical foundation for clustering, a powerful approach is to introduce additional probabilistic structures for the data. Such modeling is important to provide guarantee that we are doing the right thing under certain assumptions, but more importantly it opens up new venues for developing more sophisticated clustering algorithms as additional information about the data set or requirement about the inference become available. The most common statistical modeling tool is mixture models.

A mixture distribution admits the following density:

$$p(x | \phi) = \sum_{j=1}^k p_j f(x | \phi_j)$$

where f is a known density kernel, k is the number of mixing components. p_j and ϕ_j are the mixing probability and parameter associated with component j. Given n-iid sample $\overline{X} = (X_1, \ldots, X_n)$ from this mixture density, it is possible to obtain the parameters ϕ_j via MLE, which can be achieved by the Expectation-Maximization (EM) algorithm. In fact, the EM algorithm can be viewed to be a generalization of the popular k-means algorithm mentioned above.

The Bayesian approach involves endowing the parameters p_j and ϕ_j with a prior distribution. Concretely, consider a mixture of k Gaussians, then $f(x|\phi_j)$ can be taken as the normal density with mean and variance specified by ϕ_j. For instance, let $f(x|\phi_j) := N(\phi_j, 1)$. For prior specification for ϕ_j, as in Section 1, we take $\phi_j \sim N(\mu, \tau^2)$ independently for $j = 1, \ldots, k$, for some hyperparameter μ and τ. As for mixing probability vector $\boldsymbol{p} = (p_1, \ldots, p_k)$, as in Section 2, we take $\boldsymbol{p} \sim \mathcal{D}(\boldsymbol{\alpha})$ for some hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. The posterior distribution of p and ϕ can be obtained via Markov Chain Monte Carlo sampling algorithm, a topic that we will discuss in the sequel.

One important modeling issue is the choice of k. Various approaches exist ranging from parametric to nonparametric ideas. Parametric techniques typically evoke criteria such as AIC or BIC, which controls the complexity of the model in some way. Nonparametric techniques are more flexibile, as they allow for the model complexity (in this case driven by k) to grow as the data sample size n increases. Bayesian nonparametrics is an approach that achieves such inference in a Bayesian way. The question is how.

In our present example, k will be left unbounded. As a result, we need to give suitable prior distributions for the infinite number of parameters ϕ_j and p_j. Since the ϕ_js are unconstrained, it is simple to make a valid choice, by letting $\phi_j \sim N(\mu, \tau^2)$ independently for *all* j.

The nontrivial issue lies in specifying the prior for $\boldsymbol{p} = (p_1, p_2, \ldots)$, which is now an infinite sequence satisfying the constraint that $p_j \geq 0$ and $\sum_j p_j = 1$. With some moment of thought, it is possible to conceive the following specification based a random process of "stick-breaking": take a unit length stick, break it into two shorter pieces in a random fashion, one of which is assigned to be of length p_1, and the remaining part of length $1 - p_1$ is broken again randomly to obtain p_2, and so on. Formally, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ be iid $B(1, \alpha)$, where B denotes Beta distribution. Define

$$p_1 = \theta_1, p_n = \prod_{i=1}^{n-1} (1 - \theta_i)\theta_n, n = 2, 3, \ldots.$$

It is easy to check that the infinite sequence $\boldsymbol{p}$ constructed this way satisfies the constraint that $\sum_i p_i = 1$ almost surely. Moreover, the mixing distribution

$$P = \sum_{j=1}^{\infty} p_j \delta_{\phi_j}$$

is clearly a random probability measure on the space of $\mathbb{R}$.

It turns out (and a deep fact) that the random probability measure (PM) so defined has a particular distribution, the Dirichlet distribution on space of probability measures, and the random PM is accordingly called the Dirichlet process. Dirichlet processes are perhaps the most important building block of Bayesian nonparametrics — they provide a powerful tool for modeling random probability measures on common space such as $\mathbb{R}^d$, as well as more abstract spaces.

The stick-breaking construction given above is *not* how Dirichlet process was constructed originally by Thomas Ferguson in 1973 [Ferguson, 1973]. This explicit representation became available nearly a decade later, thanks to Jayaram Sethuraman (cf. Sethuraman [1994]). This deep connection underlies the great beauty and power of Dirichlet processes and Dirichlet distributions, which also exhibit many other attractive features. To see all this, we shall go back to take a second look at the Dirichlet on finite spaces.

4 Dirichlet measure on finite sample space

Dirichlet measure will be defined generally on $\mathbb{R}^d (d \geq 1)$, as well as more general spaces. But we will first define it for a finite sample space, as it is a simple re-expression from the finite-dimensional Dirichlet distribution on probability simplex Δ^{k-1} in Section 2.

Assume $\mathcal{X} = \{1, \ldots, k\}$ to be the sample space and let $\mathcal{B}$ be the set of all subsets of $\mathcal{X}$. Let $\mathcal{M}(\mathcal{X})$ be the set of all finite measures defined on $\mathcal{X}$, and $\mathscr{P}(\mathcal{X})$ the set of all probability measures (distributions) on $\mathcal{X}$. We write $\alpha(k)$ as an abbreviation for $\alpha(\{k\})$, and let $\boldsymbol{\alpha} = (\alpha(1), \ldots, \alpha(k)) \in \mathcal{M}(\mathcal{X})$. We shall make slight changes to the notation used in Section 2. Instead of using $\boldsymbol{\theta}$, we make use of P, which denotes the distribution for n-iid sample:

$$X_1, \ldots, X_n | P \overset{iid}{\sim} P.$$

Note that P is a distribution on $\mathcal{X}$. We say that P is random and distributed by the Dirichlet distribution, namely

$$P \sim \mathscr{D}_{\boldsymbol{\alpha}},$$

if the following holds

$$(P(1), \ldots, P(k)) \sim \mathcal{D}(\alpha(1), \ldots, \alpha(k)).$$

Note the slight but crucial change in the position of $\boldsymbol{\alpha}$ as a subscript in $\mathscr{D}_{\boldsymbol{\alpha}}$, and the use of letter $\mathscr{D}$. $\mathscr{D}_{\boldsymbol{\alpha}}$ represents a probability measure on the space of probability measures on the finite sample space $\mathcal{X}$.

Given $x \in \mathcal{X}$, $\delta_x \in \mathcal{M}(\mathcal{X})$ denotes the degenerate measure for which $\delta_x(B) = \mathbb{I}(x \in B)$ for any $B \in \mathcal{B}$. Then the quantity N_j from Section 2 can be written as $N_j = \sum_{i=1}^n \delta_{X_i}(j)$. With this modified notation, we have that the posterior distribution of P given n-data $\boldsymbol{X}$ is

$$\mathcal{D}(\alpha(1) + N_1, \ldots, \alpha(k) + N_k) \equiv \mathscr{D}_{\boldsymbol{\alpha} + \sum_{i=1}^n \delta_{X_i}}.$$

Summarizing, the Bayesian model of Section 2 has been re-written in a slightly more abstract, measure-theoretic form as follows:

$$P \sim \mathscr{D}_{\boldsymbol{\alpha}}, \tag{1}$$

$$X_1, \ldots, X_n | P \overset{iid}{\sim} P, \tag{2}$$

and found that the posterior distribution of (the random) P given the data is

$$P | X_1 = x_1, \ldots, X_n = x_n \sim \mathscr{D}_{\boldsymbol{\alpha} + \sum_{i=1}^n \delta_{x_i}}. \tag{3}$$

The equations in the above display express an initially standard parametric modeling and Bayesian inference in a way that anticipates the liberating spirit of nonparametric Bayesian analysis. Specifically, Eq. (2) is a standard i.i.d. statement of observed data, but the conditioning is on a probability measure P, rather than some finite dimensional parameter that defines P. Accordingly, Eq. (1) specifies the prior for P, instead of P's parameters. Finally, Eq. (3) expresses the conjugacy property of the Dirichlet prior on the space of probability measures on $\mathcal{X}$ with respect to the i.i.d. likelihood. It turns out that this conjugacy property continues to hold for much more general spaces $\mathcal{X}$. The conjugacy property is the main source behind the relevance and popularity of Dirichlet process prior in Bayesian nonparametrics.

Properties of $\mathscr{D}_{\boldsymbol{\alpha}}$ prior. These basic properties about finite dimensional Dirichlet distributions are simple to verify, by exploiting the representation based on Gamma variables.

Proposition 4.1. *Let $\boldsymbol{\alpha} \in \mathcal{M}(\mathcal{X})$ and let $(P(1), \ldots, P(k)) \sim \mathcal{D}(\alpha(1), \ldots, \alpha(k))$, then*

$$(P(1) + P(2), P(3), \ldots, P(k)) \sim \mathcal{D}(\alpha(1) + \alpha(2), \alpha(3), \ldots, \alpha(k)).$$

Proof. Write the Dirichlet variable $P(j) = Z_j / \sum_{j=1}^k Z_j$, where Z_j are independent gamma variables with shape parameter $\alpha(j)$, and scale 1. Then exploiting the fact that the sum of independent gamma variables with the same scale is again gamma, to arrive at the conclusion. □

The following is a slight generalization.

Corollary 4.1. *Let $A_1, \ldots, A_m$ be a partition of $\mathcal{X}$, that is, $\cup_i A_i = \mathcal{X}$ and $A_i \cap A_j = \emptyset$. If $P \sim \mathscr{D}_{\boldsymbol{\alpha}}$, then*

$$(P(A_1), \ldots, P(A_m)) \sim \mathcal{D}(\alpha(A_1), \ldots, \alpha(A_m)).$$

In particular,

$$(P(A_i), P(A_i^c)) \sim B(\alpha(A_i), \alpha(A_i^c))$$

where B is a Beta distribution.

Proposition 4.2. *Suppose that $P_1 \sim \mathscr{D}_{\boldsymbol{\alpha}_1}$, $P_2 \sim \mathscr{D}_{\boldsymbol{\alpha}_2}$, and $U \sim B(\boldsymbol{\alpha}_1(\mathcal{X}), \boldsymbol{\alpha}_2(\mathcal{X}))$ are mutually independent, where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{M}(\mathcal{X})$. Then*

$$UP_1 + (1 - U)P_2 \sim \mathscr{D}_{\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2}.$$

Note that if $P \sim \mathscr{D}_{\delta_j}$ then P is the degenerate probability measure δ_j (with probability one). The same conclusion holds if we say $P \sim \mathscr{D}_{\gamma\delta_j}$ for any $\gamma > 0$.

Corollary 4.2. *Let $P \sim \mathscr{D}_{\boldsymbol{\alpha}}$ and $U \sim B(1, \boldsymbol{\alpha}(\mathcal{X}))$ be independent. Then*

$$U\delta_j + (1 - U)P \sim \mathscr{D}_{\boldsymbol{\alpha} + \delta_j}.$$

Proposition 4.3. *For a measurable set $\mathcal{C} \subset \mathscr{P}(\mathcal{X}, \mathcal{B})$, the probability that P belongs to $\mathcal{C}$ under $\mathscr{D}_{\boldsymbol{\alpha}}$ is*

$$\mathscr{D}_{\boldsymbol{\alpha}}(\mathcal{C}) = \sum_{j=1}^k \frac{\alpha(j)}{\alpha(\mathcal{X})} \mathscr{D}_{\boldsymbol{\alpha} + \delta_j}(\mathcal{C}).$$

Proof. If $X|P \sim P$, and $P \sim \mathscr{D}_{\boldsymbol{\alpha}}$, we have that $\mathbb{P}(X = j) = \mathbb{E}(\mathbb{P}(X = j|P)) = \mathbb{E}P(j) = \alpha(j)/\alpha(\mathcal{X})$. Now,

$$\mathscr{D}_{\boldsymbol{\alpha}}(\mathcal{C}) = \sum_{j=1}^k \mathbb{P}(P \in \mathcal{C}|X = j)\mathbb{P}(X = j).$$

Appealing to the previous proposition yields the desired result. □

We can rewrite the above proposition as

$$\mathbb{E}\mathscr{D}_{\boldsymbol{\alpha} + \delta_X}(\mathcal{C}) = \mathscr{D}_{\boldsymbol{\alpha}}(\mathcal{C}),$$

where X is a random variable satisfying $\mathbb{P}(X = j) = \alpha(j)/\alpha(\mathcal{X})$. Combining the previous proposition and corollary, we obtain immediately:

Corollary 4.3. *Let $P \sim \mathscr{D}_{\boldsymbol{\alpha}}$. $U \sim B(1, \alpha(\mathcal{X}))$. X is a discrete random variable with $\mathbb{P}(X = j) = \alpha(j)/\alpha(\mathcal{X})$. If U, P and X are mutually independent, then*

$$U\delta_X + (1-U)P \sim \mathscr{D}_{\boldsymbol{\alpha}}.$$

Now, if $U_1, U_2, \ldots$ are iid beta variables $B(1, \alpha(\mathcal{X}))$, $X_1, X_2, \ldots$ are iid discrete variables distributed according to $\alpha(\cdot)/\alpha(\mathcal{X})$, and $P_1, P_2, \ldots$, are iid samples of $\mathscr{D}_{\boldsymbol{\alpha}}$, and that the P_j, X_j and U_j are all mutually independent. By iterating the conclusion of the previous corollary, we obtain the following identities where the equality is in distribution:

$$\begin{aligned} P &\stackrel{d}{=} U_1\delta_{X_1} + (1-U_1)P_1 \\ &\stackrel{d}{=} U_1\delta_{X_1} + (1-U_1)U_2\delta_{X_2} + (1-U_1)(1-U_2)P_2 \\ &\stackrel{d}{=} U_1\delta_{X_1} + (1-U_1)U_2\delta_{X_2} + (1-U_1)(1-U_2)U_3P_3 \\ &\stackrel{d}{=} U_1\delta_{X_1} + (1-U_1)U_2\delta_{X_2} + (1-U_1)(1-U_2)U_3\delta_{X_3} + \ldots, \end{aligned}$$

where the last inquality follows from the fact that the product $(1-U_1)(1-U_2)\ldots$ tends to 0 with probability 1. One can see that this is nothing but the stick-breaking representation of the Dirichlet process P that we introduced in Section 3. Of course, we have only proved the connection of Dirichlet measure to the stick-breaking representation for finite sample space $\mathcal{X}$. Extending this result to general spaces requires the more powerful machinery of measure-theoretic probability theory.

5 Probability space and measures

When speaking of probability distributions, we usually have pictures of probability mass functions or density functions. In fact, these are exceptions rather than the norm. To define general distributions on various spaces, and to eventually speak of "random distributions", we need more advanced probability concepts involving sigma algebras and measures.

A probability measure on a space such as Ω is a function applied to the collection of subsets of Ω. This collection of subsets has to be sufficiently rich, such that they are closed under standard set operations such as (countable) intersection, union and complementation, so we can talk about assigning probability mass to these subsets in a reasonable way. In the formal theory of probability, this collection of subsets is called a sigma algebra. Specifically, a collection of subsets $\mathcal{B}$ of Ω is said to be a sigma algebra if it satisfies the following conditions:

(i) $\Omega \in \mathcal{B}$.

(ii) If $A \in \mathcal{B}$ then $A^c \in \mathcal{B}$.

(iii) For any $A_1, A_2, \ldots, A_n, \ldots \in \mathcal{B}$, we have $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$.

The smallest sigma algebra that includes as its members all open sets of a sample space Ω (e.g., $\Omega = \mathbb{R}$) is called the Borel sigma algebra of subsets of Ω. The Borel sigma algebra is sufficient for most practical and theoretical purposes.

Definition 5.1. *P is a probability measure on the measurable space denoted by $(\Omega, \mathcal{B})$ if it is a function defined on $\mathcal{B}$ and*

(i) $0 \leq P(A) \leq 1$.

(ii) $P(\Omega) = 1$.

(iii) $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ *for disjoint sets* $A_1, \ldots \in \mathcal{B}$.

Removing the first two conditions while retaining the third one gives a definition of measures. Thus probability measures are specially case of positive measures that assign 1 to the entire set Ω. The third condition, also known as *countable additivity* property, is equivalent to the following two conditions:

- $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ for disjoint sets $A_1, A_2 \in \mathcal{B}$.
- $P(A_n) \downarrow 0$ for sequence of sets $A_n \in \mathcal{B}$ such that $A_n \downarrow \emptyset$.

Specifying a valid probability distribution (measure) on a given space is nontrivial matter, but there are standard tools to handle this with ease. For instance, take $\Omega = \mathbb{R}$. it suffices to specify probability values (subject to the above conditions) to all half-open interval for the form $(a, b]$, where $a < b$ are pairs of rational numbers. Then it is possible to extend this specification to all subsets in the Borel sigma algebra of $\mathbb{R}$ by appealing to countable additivity property. The fact that the extension can be achieved is due to an application of the monotone class theorem, which is a standard tool in measure theory.

Given a probability measure P on $\mathbb{R}$, the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ is defined as $F(x) = P(X \leq x)$. Conversely, any function $F : \mathbb{R} \to [0, 1]$ that satisfies the following properties:

(i) F is a non-decreasing function.

(ii) F is right continuous.

(iii) $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to +\infty} F(x) = 1$,

there exists a unique probability measure on P that corresponds to the cdf F. P is specified by first assigning probability values to half-open intervals of rational pairs $(a, b]$ such that $P((a, b]) := F(b) - F(a)$. Then for any finite union of disjoint pairs $(a_i, b_i]$ for $i = 1, \ldots, n$, $n \in \mathbb{N}$, define $P(\cup_{i=1}^{n}(a_i, b_i])) = \sum_{i=1}^{n} F(b_i) - F(a_i)$. Due to the right continuity of F, we can specify $P((a, b])$ for any pair of real numbers a, b, and consequently assign probability value for any finite union of disjoint half-open interval of real numbers. Finally, using monotone class theorem one can show that there exists a unique probability assignment to all element of the Borel sigma algebra of $\mathbb{R}$ that is consistent with all probability assignments we have just described. (The content of this statement is known as Caratheodory's extension theorem).

When the probability space $\Omega = \mathbb{R}^d$ or other abstract spaces, we cannot rely upon the specification of distribution function F, which does not exist, but the basic idea remains the same: one may first specify probability values to a countable number of elements of $\mathcal{B}$ based on an approximating and countable set of Ω, while making sure that the countable additivity property holds for the specification. This is often easy to achieve since there are only a countable number of identities to verify. Then one may extend the probability assignment to all subsets of $\mathcal{B}$, by appealing to the monotone class theorem. We do not have to worry about these issues for the rest of our lectures.

6 Random measures and Dirichlet processes

Having defined a (probability) measure as function defined on a (Borel) sigma algebra of Ω for which countable additivity holds, how do we make sense of random measures?

A random measure may be regarded as a family of random variables indexed by the Borel sets of Ω. In particular, if P is a random measure, this entails the collection of random variables $\{P(A)|A \in \mathcal{B}\}$. But there is much more. We expect that $(P(A_1), P(A_2), \ldots, P(A_n))$ is a n-dimensional random variable for any given subsets $A_1, \ldots, A_n \subset \mathcal{B}$, $n \in \mathbb{N}$. The collection of such random variables has to be consistent in distribution with one another. It is necessary (but in general not sufficient) that

(i) $P(A \cup B) = P(A) + P(B)$ almost surely for all A, B disjoint Borel subsets of Ω

(ii) $P(A_n) \downarrow 0$ almost surely for any sequence of vanishing Borel subsets A_n of Ω.

A precise definition is a bit more abstract: Consider measure space $(\Omega, \mathcal{B})$ (e.g., $\Omega = \mathbb{R}$ and $\mathcal{B}$ the Borel sigma algebra of $\mathbb{R}$). Let $\mathscr{P} := \mathscr{P}(\Omega)$ denote the space of probability measures on Ω. To speak of measures of measures in $\mathscr{P}$, we need a suitable notion of sigma algebra of subsets of $\mathscr{P}$. This is the smallest sigma algebra that contains sets of the form $\{P \in \mathscr{P} : P(B) < r | B \in \mathcal{B}, r \in \mathbb{R}\}$, and denoted by $\sigma(\mathscr{P})$. Finally, to speak of randomness (for measure), we need an underlying measure space $(\Theta, \mathcal{E})$.

Definition 6.1. *A random probability measure P on $(\Omega, \mathcal{B})$ is an $\mathscr{P}$-valued random variable with respect some underlying probability space $(\Theta, \mathcal{E})$, that is, P is a function of two arguments in $\Theta \times \mathcal{B}$ such that*

$$\begin{cases} P(\omega, \cdot) \in \mathscr{P} \text{ for each } \omega \in \Theta \\ P(\cdot, A) \text{ is a random variable on } (\Theta, \mathcal{E}) \text{ for each } A \in \mathcal{B}. \end{cases}$$

Two random measures P_1 and P_2 are equivalent almost surely, denoted by $P_1 = P_2$ a.s., if for any $A \in \mathcal{B}$, there holds

$$P_1(\omega, A) = P_2(\omega, A) \text{ almost surely}.$$

Note that the equivalence condition in this definition is weaker than requiring that $P_1(\omega, A) = P_2(\omega, A)$ for all $A \in \mathcal{B}$ almost surely. As is customary, instead of using $P(\omega, A)$ we often suppress argument ω when talking about random variables $P(A)$ where $A \in \mathcal{B}$.

Verifying that a function $P : \Theta \times \mathcal{B} \to \mathbb{R}$ is a valid random measure is nontrivial, because it involves checking that $P(\cdot, A)$ be a valid random variable for uncountably many sets $A \in \mathcal{B}$. It is a deep result which establishes that, when Ω is a complete separable metric space, the collection of non-negative valued random variables $\{P(A)|A \in \mathcal{B}\}$ that satisfies the conditions (i) and (ii) presented at the beginning of this section can be extended to a corresponding random measure, again denoted by P. Moreover, such random measure is unique.

We are in a position to give Ferguson's original definition of Dirichlet processes, one of the most famous instances of random (probability) measures.

Definition 6.2. *Let $\boldsymbol{\alpha}$ be a non-negative measure on Ω. P is a random measure on Ω such that for any partition $(A_1, \ldots, A_n), n \in \mathbb{N}$ of Ω, $(P(A_1), \ldots, P(A_n))$ is a random vector distributed according to the Dirichlet distribution $\mathcal{D}(\alpha(A_1), \ldots, \alpha(A_n))$.*

It is easy to check that the collection of variables $\{P(A)|A \in \mathcal{B}\}$ defined in the definition can be constructed to satisfy the consistency conditions (i) and (ii) (cf. Proposition 4.1). Then, by appealing to the existence and uniqueness theorem mentioned in the above paragraph, one may be able to conclude that such random probability P exists and is unique. We shall call the random measure P a Dirichlet process. The distribution for the random variable P is referred to as Dirichlet measure or Dirichlet distribution. Notationally, we write

$$P \sim \mathscr{D}_{\boldsymbol{\alpha}}.$$

If we write $\boldsymbol{\alpha} = \alpha\boldsymbol{\beta}$, where $\alpha = \boldsymbol{\alpha}(\Omega) > 0$ and $\boldsymbol{\beta}$ a probability measure on Ω, then $\boldsymbol{\beta}$ is also called the base probability measure, which represents the mean of the Dirichlet process, while α is called concentration parameter, which describes the concentration of Dirichlet processes around their mean.

For the rest of this section we prove several important consequences of Ferguson's definition. The first is on the conjugacy of the Dirichlet prior.

Theorem 6.1. *If $P \sim \mathscr{D}_{\alpha\boldsymbol{\beta}}$, and $X|P \sim P$, then*

$$P|X = x \sim \mathscr{D}_{\alpha\boldsymbol{\beta}+\delta_x}.$$

Proof. Let Q be the distribution of (P, X), and Q_P, Q_X the marginal distribution of P and X, respectively. Q_X^P and Q_P^X are conditional distribution of X given P, and P given X, respectively. By our assumption, $Q_P = \mathscr{D}_{\alpha\boldsymbol{\beta}}$, $Q_X^P = P$. For any measurable $A \subset \Omega$, $Q_X(A) = Q(X \in A) = \mathbb{E}[Q(X \in A)|P] = \mathbb{E}P(A) = \boldsymbol{\beta}(A)$.

To study the posterior distribution of P given X, let $(A_1, \ldots, A_n)$ be a partition of Ω, and take any $C \in \mathcal{B}^n$, the conditional probability that $(P(A_1), \ldots, P(A_k)) \in C$ given X is any function $f(X)$ such that it is $\mathcal{B}$ measurable, satisfying

$$Q((P(A_1), \ldots, P(A_k)) \in C, X \in B) = \int \mathbb{I}(x \in B) f(x) Q_X(dx).$$

Write $BA_i := B \cap A_i$, $BA_i^c = A_i \setminus BA_i$, $P_i := P(A_i)$. Note that

$$\begin{aligned} Q((P_1, \ldots, P_n) \in C, X \in B) &= \sum_{i=1}^n Q((P_1, \ldots, P_n) \in C, X \in BA_i) \\ &= \sum_{i=1}^n \mathbb{E}\Big\{ Q[(P_1, \ldots, P_n) \in C, X \in BA_i|P] \Big\} \\ &= \sum_{i=1}^n \mathbb{E}\Big\{ \mathbb{I}[(P_1, \ldots, P_n) \in C] P(X \in BA_i) \Big\} \end{aligned}$$

Fix i. Note that $(A_1, \ldots, A_{i-1}, BA_i, BA_i^c, \ldots, A_n)$ forms a partition of Ω, and by definition of Dirichlet processes, $(P_1, \ldots, P_{i-1}, P(BA_i), P(BA_i^c), \ldots, P_n) \sim \mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(BA_i), \alpha\beta(BA_i^c), \ldots, \alpha\beta(A_n))$.

Therefore,

$$\begin{aligned} & \mathbb{E}\Big\{ \mathbb{I}[(P_1, \ldots, P_n) \in C] P(X \in BA_i) \Big\} \\ = & \int_{(y_1, \ldots, y_{i1} + y_{i2}, \ldots, y_n) \in C} y_{i1} \mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_{i-1}), \alpha\beta(AB_i), \alpha\beta(AB_i^c), \ldots, \alpha\beta(A_n))\, dy_1 \ldots dy_n \\ = & \int_{(y_1, \ldots, y_{i1} + y_{i2}, \ldots, y_n) \in C} y_1^{\alpha\beta(A_1)-1} \ldots y_{i1}^{\alpha\beta(BA_i)} y_{i2}^{\alpha\beta(BA_i^c)-1} \ldots y_n^{\alpha\beta(A_n)-1}\, dy_1 \ldots dy_n \\ = & \frac{\beta(BA_i)\Gamma(\alpha+1)}{\Gamma(\alpha\beta(A_1) \ldots \Gamma(\alpha\beta(BA_i)+1)\Gamma(\alpha\beta(BA_i^c) \ldots \Gamma(\alpha\beta(A_n))} \times \\ & \int_{(y_1, \ldots, y_{i1} + y_{i2}, \ldots, y_n) \in C} y_1^{\alpha\beta(A_1)-1} \ldots y_{i1}^{\alpha\beta(BA_i)} y_{i2}^{\alpha\beta(BA_i^c)-1} \ldots y_n^{\alpha\beta(A_n)-1}\, dy_1 \ldots dy_n, \end{aligned}$$

where the last equation is due to the identities that $\int_{\Delta^{n-1}} y_1^{\beta_1-1} \ldots y_n^{\beta_n-1}\, dy_1 \ldots dy_n = \Gamma(\beta_1) \ldots \Gamma(\beta_n)/\Gamma(\beta_1 + \ldots + \beta_n)$, and that $\Gamma(u+1) = u\Gamma(u)$ for any $u > 0$.

So, if we allow

$$(Y_1,\ldots,Y_{i-1},Y_{i1},Y_{i2},\ldots,Y_n)\sim\mathcal{D}(\alpha\beta(A_1),\ldots\alpha\beta(A_{i-1}),\alpha\beta(BA_i)+1,\alpha\beta(BA_i^c),\alpha\beta(A_{i+1}),\ldots,\alpha\beta(A_n)),$$

then by Prop. 4.1, $(Y_1,\ldots,Y_{i1}+Y_{i2},\ldots,Y_n)\sim\mathcal{D}(\alpha\beta(A_1),\ldots\alpha\beta(A_{i-1}),\alpha\beta(A_i)+1,\alpha\beta(A_{i+1}),\ldots,\alpha\beta(A_n))$. Thus,

$$\begin{aligned}
&\mathbb{E}\Big\{\mathbb{I}[(P_1,\ldots,P_n)\in C]P(X\in BA_i)\Big\}\\
=\;& Q(X\in BA_i)\mathcal{D}(\alpha\beta(A_1),\ldots,\alpha\beta(A_{i-1}),\alpha\beta(A_i)+1,\alpha\beta(A_{i+1}),\ldots,\alpha\beta(A_n))(C)\\
=\;& \mathbb{E}\Big\{\mathbb{I}(X\in BA_i)\mathcal{D}(\alpha\beta(A_1)+\delta_X(A_1),\ldots,\alpha\beta(A_n)+\delta_X(A_n))(C)\Big\},
\end{aligned}$$

where the last equality in the above display is due to $BA_i\subset A_i$. Summing over $i=1,\ldots,n$ we get

$$Q((P(A_1),\ldots,P(A_n))\in C,X\in B)=\mathbb{E}\Big\{\mathbb{I}(X\in B)\mathcal{D}(\alpha\beta(A_1)+\delta_X(A_1),\ldots,\alpha\beta(A_n)+\delta_X(A_n))(C)\Big\}$$

for all $B\in\mathcal{B}$. Hence, $\mathcal{D}(\alpha\boldsymbol{\beta}+\delta_X(A_1),\ldots,\alpha\boldsymbol{\beta}+\delta_X(A_n))$ is a version of the conditional distribution of $(P(A_1),\ldots,P(A_n))$ given X. By Ferguson's definition, this conclusion is immediate. □

The following result, due to David Blackwell, may be a surprise for many.

Proposition 6.1. *If P is a Dirichlet process on $\mathbb{R}$, then P is discrete almost surely.*

Proof. Consider the set $D=\{(P,x):P(\{x\})>0\}$. This set is measurable, moreover by Fubini's theorem

$$Q(D)=\int Q_P^x(D_x)Q_X(dx)=\int\mathscr{D}_{\alpha\boldsymbol{\beta}+\delta_x}(D_x)Q_X(dx),$$

where $D_x:=\{P:P(\{x\})>0\}$. Under $\mathscr{D}_{\alpha\boldsymbol{\beta}+\delta_x}$, we have $(P(\{x\}),P(\mathbb{R}\backslash\{x\}))\sim\mathcal{D}(1+\alpha\boldsymbol{\beta}(\{x\}),\alpha\boldsymbol{\beta}(\mathbb{R}\backslash\{x\}))$, which yields that $P(\{x\})>0$ almost surely. Thus, $\mathscr{D}_{\alpha\boldsymbol{\beta}+\delta_x}(D_x)=1$, and so $Q(D)=1$. Again, by Fubini's theorem, we also have

$$Q(D)=\int Q_X^p(D_p)Q_P(dp)=\int P(D_p)\mathscr{D}_{\alpha\boldsymbol{\beta}}(dp)$$

where $D_p=\{x:P(\{x\})>0\}$. Since $Q(D)=1$, we have $P(D_p)=1$ almost surely under $\mathscr{D}_{\alpha\boldsymbol{\beta}}$. That is, P is discrete almost surely under $\mathscr{D}_{\alpha\boldsymbol{\beta}}$. □

7 Characterizations via Pólya sequence

Exchangeablity and de Finetti's theorem A sequence of random variables $X_1,X_2,\ldots$ in $(\Omega,\mathcal{B})$ is exchangeable if for all $n\geq 2$, and all permutations i of n elements, $(X_1,\ldots,X_n)$ and $(X_{i(1)},\ldots,X_{i(n)})$ have the same joint distribution. One of the deepest results in probability is de Finetti's theorem, which states that "an infinite exchangeble sequence is a mixture of i.i.d. sequences". That is, there exists an unique random measure P (with probability distribution $\boldsymbol{\alpha}$) such that conditional on P, $X_1,\ldots,X_n,\ldots$ are iid sample according to P. Precisely, for any measurable sets $A_1,\ldots,A_n$,

$$Q(X_1\in A_1,\ldots,X_n\in A_n)=\int P(A_1)\ldots P(A_n)\alpha(dP).$$

P is referred to as the directing random measure of the exchangeable sequence $X_1, \ldots$. There are a number of proofs for this result — a common proof makes use of the reverse martingale convergence theorem. See the textbook by Kallenberg [2005].

The celebrated theorem of de Finetti also plays a foundational role in Bayesian statistics, because it suggests the need for a prior distribution in the statistical modeling of exchangeable data. Note that exchangeability is a much weaker assumption than the i.i.d. assumption (often evoked in frequentist statistics), because according de Finetti's theorem, the former is equivalent to the weaker notion of conditional i.i.d.

Consider the following infinite sequence obtained by a two-stage procedure:

$$\begin{aligned} P &\sim \mathscr{D}_{\boldsymbol{\alpha}} \\ X_1, \ldots, X_n, \ldots | P &\overset{iid}{\sim} P, \end{aligned}$$

what can we say about the distribution of the sequence $X_1, \ldots, X_n$, which is obtained by marginalizing out the random measure P? Clearly, $X_1, \ldots$ generated this way is an infinite exchangeable sequence, for which Dirichlet process is the (unique) directing random measure. It turns out that the joint distribution of $X_1, \ldots$ can be specified explicitly by a sampling scheme known as as Pólya sequence. This allows us to have an alternative and very simple characterization of Dirichlet processes.

Pólya sequence Let $\boldsymbol{\alpha}$ be a measure on a complete separable metric space Ω. We shall say that a sequence $\{X_n, n \geq 1\}$ of random variables with values in Ω is a Pólya sequence with parameter $\boldsymbol{\alpha}$ if for every $A \subset \Omega$,

$$\begin{aligned} Q(X_1 \in A) &= \boldsymbol{\alpha}(A)/\boldsymbol{\alpha}(\Omega) \\ Q(X_{n+1} \in A | X_1, \ldots, X_n) &= \boldsymbol{\alpha}_n(A)/\boldsymbol{\alpha}_n(\Omega), \end{aligned}$$

where $\boldsymbol{\alpha}_n = \boldsymbol{\alpha} + \sum_{i=1}^n \delta_{X_i}$. Note that, for finite Ω, say $\Omega = \{1, \ldots, k\}$, the sequence $\{X_n\}$ represents the results of successive draws from an urn where initially the urn has $\boldsymbol{\alpha}(x)$ balls of color x and, after each draw, the ball drawn is replaced and another ball of its same color is added to the urn.

The following theorem is due to Blackwell and James MacQueen [Blackwell and MacQueen, 1973].

Theorem 7.1. *Let $\{X_n\}$ be a Pólya sequence with parameter $\boldsymbol{\alpha}$. Then*

(a) $m_n := \boldsymbol{\alpha}_n/\boldsymbol{\alpha}(\Omega)$ converges with probability one as $n \to \infty$ to a limiting discrete measure P.

(b) P is a Dirichlet process with parameter $\boldsymbol{\alpha}$, and

(c) Given P, the variables $X_1, X_2, \ldots$ are independently and identically distributed with distribution P.

Proof. Here is a proof for the case Ω is finite, say $\Omega = \{1, 2, \ldots, k\}$. If $\{X_n\}$ is a Pólya sequence, it is an easy calculation (by induction) to obtain that

$$\begin{aligned} Q(X_1 = x_1, \ldots, X_n = x_n) &= Q(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) Q(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \\ &= \prod_{x=1}^{k} \boldsymbol{\alpha}(x)^{[n(x)]} / \boldsymbol{\alpha}(\Omega)^{[n]}, \end{aligned}$$

where $n(x)$ denotes the number of i with $x_i = x$, and $a^{[k]} = a(a+1)\ldots(a+k-1)$. This formula also shows that the Pólya sequence is exchangeable.

On the other hand, if $X_1, \ldots, X_n | P \sim P$ where P is a Dirichlet process with parameter $\boldsymbol{\alpha}$, then

$$\begin{aligned} Q(X_1 = x_1, \ldots, X_n = x_n) &= \mathbb{E}Q(X_1 = x_1, \ldots, X_n = x_n | P) \\ &= \mathbb{E}\prod_{x=1}^{k} P(k)^{n(x)} \\ &= \prod_{x=1}^{k} \boldsymbol{\alpha}(x)^{[n(x)]} / \boldsymbol{\alpha}(\Omega)^{[n]}, \end{aligned}$$

where the last identity is a classical formulae for the moments of Dirichlet distribution $\mathcal{D}(\boldsymbol{\alpha}(1), \ldots, \boldsymbol{\alpha}(k))$. Therefore, the directing random measure for the exchangeable Pólya sequence is a Dirichlet process.

It remains to prove (a). We have shown that the Pólya sequence $X_1, \ldots, X_n, \ldots$ is an iid sequence with the law P, where P is a random measure (Dirichlet process). If π_n is the empirical distribution of $X_1, \ldots, X_n$, that is: $\pi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, then it follows from the strong law of large numbers that π_n converges in distribution to P with probability one as $n \to \infty$. Since $m_n = (\boldsymbol{\alpha} + \sum_{i=1}^{n} \delta_{X_i})/(\boldsymbol{\alpha}(\Omega) + n) = (\mu + n\pi_n)/(\boldsymbol{\alpha}(\Omega) + n)$, (a) follows. □

Chinese restaurant process Let $\boldsymbol{\alpha} = \alpha\boldsymbol{\beta}$ be a measure on Ω (that is, $\alpha = \boldsymbol{\alpha}(\Omega) > 0$, and $\boldsymbol{\beta}$ is a probability measure on Ω). The Pólya sequence $\{X_n\}$ is equivalently rewritten as follows: First, draw $X_1 \sim \boldsymbol{\beta}$. For $n \geq 1$, draw $X_{n+1} = X_i$ with probability $1/(\alpha + n)$ for each $i = 1, \ldots, n$, and draw $X_{n+1} \sim \boldsymbol{\beta}$ with the remaining probability $\alpha/(\alpha + n)$.

This construction makes explicit the clustering effects of the Pólya sequence. As long as $\boldsymbol{\beta}$ is an non-atomic probability measure, at step $n + 1$, the probability that X_{n+1} taking the same value as one of the previous elements in the sequence is $n/(\alpha+n)$. X_{n+1} is assigned a new value with the remaining probability. If we are interested only in how the sequence $\{X_n\}$ are subdivided into subsets of equal values without considering the actual values, what we have obtained is a random partition of the natural numbers $1, 2, \ldots$. The random partition constructed this way was given by a colorful name, *Chinese restaurant process*.

The Chinese restaurant process (CRP) makes random table seating assignments for an infinite sequence of customers (numbered by the naturals 1,2,...) arriving at a Chinese restaurant which has infinite number of tables: the first customer arrives and seats at an arbitrary table there. The following customers arrive and choose their table by the following rule: either one of the non-empty tables is chosen with probability proportional to the current number of customers seating at that table; otherwise, that customer chooses a new table with probability proportional to α.

Note that the CRP makes use of only scalar parameter $\alpha = \boldsymbol{\alpha}(\Omega)$. It is simple to observe how one can construct the Pólya sequence from the CRP: View Ω as the menu of dishes, and $\boldsymbol{\beta}$ a probability distribution on dishes. Given a sequence of customers indexed by the naturals, whose seating assignments are given by the Chinese restaurant process with parameter $\alpha > 0$. For each table, a dish is ordered independently from distribution $\boldsymbol{\beta}$, and shared by all customers seating at the table. Let X_i be the dish that i is having. Then, $X_1, \ldots, X_n, \ldots$ is a Pólya sequence with parameter $\alpha\boldsymbol{\beta}$.

8 Other properties of Dirichlet processes

Sethuraman's explicit representation We have seen the original definition of Dirichlet processes by Thomas Ferguson, who showed that such a random measure exists and is unique. We also know that Dirichlet process is a discrete probability measure almost surely. Another way of showing the existence of Dirichlet

processes is via the Pólya sequence for which a Dirichlet processes act as the directing measure. Both the definition and the Pólya urn's characterization remain quite abstract in that they do not allow us to get a handle on the explicit representation of the Dirichlet processes. The explicit representation of Sethuraman resolved this difficulty — it also contributed to the booming popularity of Dirichlet processes in particular and Bayesian nonparametrics in general.

Theorem 8.1. *Let Ω be a metric space, $\boldsymbol{\alpha} = \alpha\boldsymbol{\beta}$ a measure on Ω, where $\alpha = \boldsymbol{\alpha}(\Omega)$. Let $\Theta := \{\theta_1, \theta_2, \ldots\}$ be an iid sequence with common distribution Beta$(1, \alpha)$, and $\boldsymbol{Y} = (Y_1, Y_2, \ldots)$ be iid with common distribution $\boldsymbol{\beta}$. Let Θ be independent of $\boldsymbol{Y}$. Define $\boldsymbol{p} = (p_1, p_2, \ldots)$ by $p_1 = \theta_1$, $p_n = \theta_n \prod_{j=1}^{n-1}(1 - \theta_j)$, for $n = 2, 3, \ldots$. Define a random probability measure $P(\cdot)$ by*

$$P = \sum_{i=1}^{\infty} p_i \delta_{Y_i}.$$

Then the distribution of P is the Dirichlet measure $\mathscr{D}_{\alpha\boldsymbol{\beta}}$.

This theorem can be proved by directly verifying the original definition of Ferguson. It is interesting to note that Ferguson requires Ω to be complete separable metric space, so that a general argument of existence of random measures can be invoked. However, no such assumption is required in Sethuraman's theorem. To prove Sethuraman's construction is indeed that of a Dirichlet process, one can verify that the random measure P defined in the above theorem satisfies the following properties: the vector of probability mass P on any finite partition of Ω follows a finite dimensional Dirichlet distribution. Moreover, if $X|P \sim P$, then the conditional distribution of such a vector induced by P is again another Dirichlet distribution with suitably updated parameters. The proof is reminiscent of the calculations we have done in Section 4 for finite sample spaces. See Sethuraman [1994] for the details.

A Markov chain characterization Sethuraman's construction allows us to write the following equality in distribution:

$$P \stackrel{d}{=} \theta_1 \delta_{Y_1} + (1 - \theta_1) P$$

where $\theta_1 \sim B(1, \alpha)$, $Y_1 \sim \boldsymbol{\beta}$, and P is a Dirichlet process distributed by $\mathscr{D}_{\alpha\boldsymbol{\beta}}$, such that θ_1, Y_1 and P are independent.

This equation can be recognized as the equation for a stationary measure for the Markov chain defined via the recursion:

$$P_n = \theta_n \delta_{Y_n} + (1 - \theta_n) P_{n-1}; n \geq 1$$

where $P_0 \in \mathscr{P}(\Omega)$ is arbitrary and $\{(\theta_n, Y_n)\}$ is an iid sequence with the same distribution as (θ_1, Y_1) above. It was shown by Feigin and Tweedie [1989] that this Markov chain has an unique invariant measure which we may identify as the Dirichlet process distributed by $\mathscr{D}_{\alpha\boldsymbol{\beta}}$.

Support of Dirichlet measure Dirichlet measures place mass on the space of measures $\mathscr{P}(\Omega)$. We are interested in where a Dirichlet measure concentrates its mass in $\mathscr{P}(\Omega)$.

Definition 8.1. *The support of a probability measure μ is the smallest closed set M with $\mu(M) = 1$. Equivalently, M is the support of μ if and only if each of the following conditions is met:*

(i) M is a closed set.

(ii) $\mu(M) = 1$.

(iii) For any $x \in M$, there exists a neighborhood, $S(x, \epsilon)$ of x for some $\epsilon > 0$ with $\mu(S(x, \epsilon)) > 0$.

We denote the support of μ by $\operatorname{spt} \mu$.

The following result can be found in Ghosh and Ramamoorthi [2002]:

Proposition 8.1. *Let M be the support of $\boldsymbol{\alpha}$, a measure on Ω. Then, the support of $\mathscr{D}_{\boldsymbol{\alpha}}$ is*

$$\mathscr{P} = \{P | \operatorname{spt} P \subset M\}.$$

As a consequence, if $\boldsymbol{\alpha}$ is any probability measure with support of the entire real line, then the support of Dirichlet measure is the entire space of probability measures on $\mathbb{R}$. This holds, despite the fact that Dirichlet processes are discrete almost surely. This partly explains that Dirichlet is a sufficiently rich prior for the space of probability measures, making it a suitable choice for Bayesian nonparametric modeling.

One can say more about the concentration of mass on the space of probability measures under the Dirichlet measure. In particular, if Ω is a bounded set, and $\boldsymbol{\alpha}$ behaves like a uniform distribution on Ω, then $\mathscr{D}_{\boldsymbol{\alpha}}$ also behaves like a uniform distribution on $\mathscr{P}(\Omega)$. See Lemma 5 of Nguyen [2013] for a precise statement, which gives a lower bound on the probability mass on a neighborhood centered on any element of $\mathscr{P}(\Omega)$, where the neighborhood is specified by a Wasserstein metric. This result is also crucial for establishing an asymptotic theory for posterior inference with models based on Dirichlet process prior — a topic we shall touch on in the last section.

9 Markov Chain Monte Carlo algorithms

We are ready to discuss statistical modeling using the tools introduced in the previous sections. We recall the statistical inference problems described in Section 3. Recall the clustering problem which is approached using mixture models. We do not want to specify the number of mixing components. We will take a Bayesian nonparametric approach, by endowing a prior distribution on the space of mixing measures. That prior is the Dirichlet process. The resulting mixture model is known as Dirichlet process mixtures, first studied by Antoniak [1974] and Lo [1984].

Let $f(x|\phi)$ be a known density kernel. For simplicity, we assume that both x and ϕ are variables taking values in $\Omega = \mathbb{R}^d$, for some $d \geq 1$. A mixture model specifies the density as follows:

$$p(x|p_1, \ldots; \phi_j \ldots) = \sum_{j=1}^{\infty} p_j f(x|\phi_j),$$

where the collection of parameters (p_j, ϕ_j) are represented by the mixing measure

$$G = \sum_{i=1}^{\infty} p_j \delta_{\phi_j}.$$

Accordingly, the mixture density given in the previous equation is to be denoted by $p_G = \sum p_j f(x|\phi_j)$.

We shall endow G with the Dirichlet process prior. As a result, the full Bayesian model is given as follows

$$G|\alpha, G_0 \sim \mathscr{D}_{\alpha G_0} \tag{4}$$

$$X_1, \ldots, X_n|G \overset{iid}{\sim} p_G. \tag{5}$$

Here G_0 is called a hyperparameter, and it can be taken to be a parametric distribution on Ω.

Example f is a normal kernel with varying mean and variance: that is, $\phi_j = (\mu_j, \nu_j)$, and $f(\cdot|\phi_j) := N(\cdot|\mu_j, \nu_j)$. G_0 is a parametric distribution on the mean and variance parameters:

$$G_0(d\mu, d\nu) = G_0(d\nu) \times G_0(d\mu|\nu) := \text{InvGamma}(s/2, S/2) \times N(m, \tau\nu),$$

where (s, S) are scale-shape parameters for the inverse Gamma distribution for ν, while m and $\tau\nu$ specify the conditional distribution of μ given ν. This choice of hyperparameters for G_0 is standard in parametric Bayesian statistics, because G_0 is the parametric conjugate prior for the location-scale normal likelihood f. In fact, it can be verified easily that by combining prior G_0 with the kernel density f, one obtain

$$\int f(y|\mu, \nu)G_0(d\mu, d\nu) = T_s(y; m, M),$$

which is the density of Student's t distribution with mode m, scale $M^{1/2}$ and s degrees of freedom, where $M = (1 + \tau)S/s$.

Once the model is set up, it remains to compute the posterior distribution of quantities of interest. In this case, we are interested in the mixing measure G, as well as the induced data density p_G. While it is generally not possible to compute posterior distributions exactly, one may produce samples from the posterior by Markov Chain Monte Carlo methods. We proceed to discuss two prominent MCMC approaches to sampling from posterior distributions induced by DP mixtures.

9.1 Marginal method

Calculating the posterior of mixing measure G given the data is a nontrivial task, because the posterior distribution in question is on the space of measures, whose dimensionality is potentially unbounded. There is a simple way to get around — one can obtain the posterior of finitely many quantities of interest not by accessing to G directly. In fact, it is possible to marginalize/integrate out the latent variable G. This approach is therefore called the "marginal" approach.

The tool that we use is the Pólya sequence characterization of Dirichlet processes — we have seen from Section 7 how one can easily draw sample of the Pólya sequence without having to sample from the directing random measure G. To this end, we will not work with the model representation (4) directly, but the following equivalent model:

$$G|\alpha, G_0 \quad \sim \quad \mathscr{D}_{\alpha G_0} \tag{6}$$

$$\theta_1, \ldots, \theta_n|G \quad \overset{iid}{\sim} \quad G \tag{7}$$

$$X_i|\theta_i \quad \overset{indep}{\sim} \quad f(\cdot|\theta_i) \text{ for } i = 1, \ldots, n. \tag{8}$$

Latent variables $\theta_1, \ldots, \theta_n$ represent the parameter with each $X_1, \ldots, X_n$ are respectively associated. Our computational goal is to contruct a Markov chain for $\{\theta_1, \ldots, \theta_n\}$ that converges to the posterior distribution $\mathbb{P}(\theta_1, \ldots, \theta_n|D_n)$, where $D_n := (X_1, \ldots, X_n)$. In particular, we may take a Gibbs sampling approach, according to which each step of the Markov chain, a sample for $\{(\theta_1, \ldots, \theta_n\}$ is obtained by drawing θ_i from the conditional distribution $\mathbb{P}(\theta_i|\theta_{-i}, D_n)$ for each $i = 1, \ldots, n$. Here θ_{-i} denotes all elements of the (θ_j) sequence except element θ_i.

For each $i = 1, \ldots, n$, distribution $\mathbb{P}(\theta_i|\theta_{-i}, D_n)$ is referred to as a "conditional posterior" (the use of word "posterior" is due to the conditioning of data set D_n, while "conditional" is due to the conditioning of parameters θ_{-i}). The conditional posterior can be obtained by invoking Bayes' rule, which combines

the "conditional prior" with "likelihood": the former is nothing but the conditional probability that we have learned from Pólya sequence:

$$\theta_i|\theta_{-i} \sim \alpha G_0 + \sum_{j\neq i} \delta_{\theta_j},$$

and the latter is given by $Y_i|\theta_i \sim f(\cdot|\theta_i)$. By Bayes' rule, and conditional independence, we have

$$\begin{aligned} \mathbb{P}(d\theta_i|\theta_{-i}, D_n) &\propto \mathbb{P}(d\theta_i|\theta_{-i})\mathbb{P}(X_i|\theta_i) \\ &\propto \alpha f(X_i|\theta)G_0(d\theta) + \sum_{j\neq i} f(X_i|\theta_j)\delta_{\theta_j}. \end{aligned}$$

Note the proportional notation $\propto$. To actually draw a sample from the distribution in the above equation, we need to be able to compute the normalizing constant, which is:

$$C = \alpha \int f(X_i|\theta)G_0(d\theta) + \sum_{j\neq i} f(X_i|\theta_j).$$

When G_0 is chosen to be a conjugate prior to kernel density f, the integral in the above display can be computed exactly. For instance, we have seen from the previous section that combining normal likelihood with normal/inverse gamma prior leads to Student's t density for the first integral that defines C. Otherwise, further tricks are required, and there are various solutions available in such situations. Early contributors of such methods include Escobar and West [1995] and MacEachern and Mueller [1998].

Summarizing, the sampling algorithm consists of the following single line of codes: For each MCMC step, do as follows:

(1) for $i = 1, \ldots, n$, draw θ_i given existing θ_{-i} and D_n by the conditional distribution defined in the equation above.

The simplicity and elegance of this MCMC sampling algorithm belies its inefficiency. In fact, the Markov chain may mix very slowly, which means it takes a long time of Markov chain sampling to obtain a valid sample from a posterior distribution of interest. More sophisticated techniques exists to alleviate such inefficiency —- see the paper by Neal [2000].

9.2 Conditional method

The marginal method does not directly give us a sample of the posterior distribution for mixing measure G. To achieve this, we rely on the explicit representation of Dirichlet process, discussed in Section 8. The resultant MCMC sampling technique is called *conditional method*, which involves constructing a Markov chain by conditioning on the explicit representation of G, as opposed to marginalizing it out.

Using Sethuraman's representation, we can generate P from random vector $\boldsymbol{V} = (V_1, \ldots)$ and $\boldsymbol{Z} = (Z_1, \ldots)$ such that

$$P = \sum_{j=1}^{\infty} p_j \delta_{Z_j},$$

where $V_1, V_2, \ldots \overset{iid}{\sim} B(1, \alpha)$, and $Z_1, Z_2, \ldots \overset{iid}{\sim} G_0$, and $p_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$, for $j = 1, 2, \ldots$.

The latent variables $\theta_1, \ldots, \theta_n$ can be generated as $Z_{r_1}, \ldots, Z_{r_n}$, where $r_1, \ldots, r_n$ given $\boldsymbol{V}$ and $\boldsymbol{Z}$ are independent with $\mathbb{P}(r_i = h|\boldsymbol{V}, \boldsymbol{Z}) = p_h$ for $h = 1, 2, \ldots$; $i = 1, \ldots, n$. Accordingly, the joint distribution of all variables in the model, including data vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$(\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{r}, \boldsymbol{X}) \sim B^{\infty}(1, \alpha) \times G_0^{\infty} \times \prod_{i=1}^{n} p_{r_i} \times \prod_{i=1}^{n} f(X_i|Z_{r_i}). \tag{9}$$

We seek to devise a Markov chain that converges to the posterior distribution of $\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{r}$ given data $\boldsymbol{X}$. There is a difficulty: there are an infinite number of variables to handle, which cannot possibly be sampled simultaneously. It turns out that this is not necessary, thanks to a clever trick due to Walker [2007]. The idea is to introduce an auxiliary sequence of random variables, $\boldsymbol{u} = (u_1, u_2, \ldots)$ with $0 \leq u_i \leq 1, i = 1, \ldots, n$ such that the distribution of $(\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{r}, \boldsymbol{X})$ is given through

$$(\boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{r}, \boldsymbol{X}) \sim B^{\infty}(1, \alpha) \times G_0^{\infty} \times \prod_{i=1}^{n} \mathbb{I}(u_i < p_{r_i}) \times \prod_{i=1}^{n} p_{r_i} \times \prod_{i=1}^{n} f(X_i|Z_{r_i}). \tag{10}$$

It is clear that integrating out all u_i in the joint distribution given by (10) leads to the joint distribution (9).

What one gains the introduction of auxiliary variables $\boldsymbol{u}$ is that, when $\boldsymbol{u}$ is conditioned on, we only need to choose labels r_i from the finite set:

$$H(u_i) := \{h : p_h > u_i\}, i = 1, \ldots, n.$$

If one thinks of a bar graph in which the height of each bar represents the magnitude of $p_j, j = 1, \ldots, \infty$, then restricting label r_i to only to $H(u_i)$ corresponds visually to "slicing" out the portion below the height u_i, and making only the bars taller than u_i to remain. For this reason, this sampling technique is called "slice sampling".

The Gibbs sampling algorithm consists of making the following steps:

(1) Sampling $\boldsymbol{u}$ conditional on $\boldsymbol{V}, \boldsymbol{r}, \boldsymbol{X}$: for each $i = 1, \ldots, n$, draw $u_i \overset{indep}{\sim} \text{Uniform}[0, p_{r_i}]$.

(2) Sampling $\boldsymbol{V}$ conditional on $\boldsymbol{r}, \boldsymbol{Z}, \boldsymbol{X}$: we only need to sample v_j for $j = 1, \ldots,$ up to the index h for which $p_h > u_i, i = 1, \ldots, n$. Let $m_h := \#\{i : r_i = h\}$ be the number of indices i such that $r_i = h$, then draw

$$V_h \sim \text{B}(1 + m_h, \alpha + \sum_{k>h} m_k), \ h = 1, \ldots, K.$$

For this, K needs to be at least as large as $\max(r_i : i = 1, \ldots, n)$. Later, if required, $V_{K+1}, V_{K+2}, \ldots$ can be generated independently from $\text{B}(1, \alpha)$.

(3) Sampling $\boldsymbol{Z}$ conditional on $\boldsymbol{V}, \boldsymbol{u}, \boldsymbol{r}, \boldsymbol{X}$: For $h = 1, \ldots, K$, draw Z_h independently according to

$$Z_h \propto \prod_{i:r_i=h} f(X_i|Z_h) G_0(dZ_h).$$

(4) Lastly, sampling $\boldsymbol{r}$ conditional on $\boldsymbol{V}, \boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{X}$: for $i = 1, \ldots, n$, draw a sample of r_i independently according to $r_i = h|\boldsymbol{V}, \boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{X} \propto f(X_i|Z_h)\mathbb{I}(u_i < p_h)$. This means that

$$r_i = h|\boldsymbol{V}, \boldsymbol{u}, \boldsymbol{Z}, \boldsymbol{X} \propto f(X_i|Z_h)$$

on support $H(u_i) = \{h|p_h > u_i\}$.

[[A technical note: If K is not large enough, it might happen that for some h in $H(u_i)$, p_h and Z_h have not been sampled. If this is the case, we need to increase K such that all $p_{K+1}, p_{K+2}, \ldots$ are less than $u_i, i = 1, 2, \ldots, n$, so that all eligible p_h, Z_h are sampled. Thus, if $1 - \sum_{n=1}^{K} p_n = \sum_{n=K+1}^{\infty} p_n \geq \min(u_i; i = 1, \ldots, n)$, we generate $V_{K+1} \sim \mathrm{B}(1, \alpha)$, $Z_{K+1} \sim G_0$, replace $K := K + 1$ and repeat the above steps until $1 - \sum_{n=1}^{K} p_n < \min(u_i; i = 1, \ldots, n)$.]]

Before ending this section, we note that the mixing properties of Markov chains constructed via either marginal or conditional method for DP mixtures remain unknown from a theoretical standpoint.

10 Hierarchical and nested Dirichlet processes

So far, the underlying mathematical foundation and algorithms of Dirichlet process mixtures have been presented. It is comforting to know that a DP mixture of normals can approximate any probability distribution that has a sufficiently smooth density. Thus, DP mixtures represent a rich class of statistical models. Coupling with the ease in posterior computation of model parameters, as we have seen in the previous section, DP mixtures have become one of the most powerful tools in Bayesian nonparametric modeling. But there are much more: they can be easily extended to address additional structures we may know about the data and their inferential problems. And therein lie the true extent of their full potential. We shall give a few canonical examples in this section — more examples can be found in the excellent collection of articles edited by Hjort et al. [2010].

10.1 Hierarchical Dirichlet processes

Suppose that we are given not one data set, but a collection of data sets $D_1, \ldots, D_m$. Each of these data sets may be modeled by a mixture model with a common kernel density f, and varying mixing measures $G_1, \ldots, G_m$. However, we might not do well on each of the data sets, as far as inference is concerned, especially for those with small sample size. Using hierarchical models, we may establish a probabilistic linkage among the m mixture models on the related data sets and are likely to gain improvement in our inference. Assume, in particular, that the m data sets are exchangeable so that $G_1, \ldots, G_m$ are conditionally i.i.d., as a consequence of de Finetti's theorem. Thus, we may view $G_1, \ldots, G_m$ as random measures that are conditionally i.i.d. according to a distribution. Assume further that $G_1, \ldots, G_m$ are Dirichlet processes independently sampled from a common Dirichlet measure with base measure $\boldsymbol{\alpha} = \alpha G_0$ on some suitable space Ω:

$$G_1, \ldots, G_m | \boldsymbol{\alpha} \overset{iid}{\sim} \mathscr{D}_{\alpha_0 G_0},$$

α_0 is a fixed hyperparameter (which may be endowed with a prior), and G_0 is a (random) probability measure on Ω that requires further distributional specifications.

We may make a parametric assumption on G_0 (such as letting $G_0 := N(\mu_0, \nu_0)$, and endow prior distributions on G_0's parameters). We may be more daring, and decide to go fully nonparametric: G_0 is endowed with another Dirichlet process prior with some base parameter γH:

$$G_0 | \gamma, H \sim \mathscr{D}_{\gamma H},$$

where $\gamma > 0$ and H is a parametric distribution on Ω. The specification the collection of mixing measures G_i via the two equations given above define what is widely known as the Hierarchical Dirichlet processes

of Teh et al. [2006]. This model and with it the liberating spirit of Bayesian nonparametrics have been adopted, successfully applied and extended to a vast number of application domains, such as image analysis, population genetics, natural language processing, robotics and environmental sciences.

Stick-breaking representation The tools developed in the previous sections on Dirichlet processes can be applied to characterize the hierarchical Dirichlet processes in much similar ways.

Since $G_0 \sim \mathscr{D}_{\gamma H}$, we can write G_0 as

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ are defined by the usual stick-breaking construction via beta variables $B(1, \gamma)$, and $\phi_1, \phi_2, \ldots$ are and i.i.d. sample of H.

Moreover, for each $j = 1, 2, \ldots$, since $G_j | G_0 \sim \mathscr{D}_{\alpha G_0}$, G_j has the same support as that of G_0, and therefore, it may be written as

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}.$$

It is not difficult to show that the collection of (ϕ_{jk}) satisfies the following law: if we view both $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ and $\boldsymbol{\pi}_j := (\pi_{j1}, \pi_{j2}, \ldots)$ as distributions on the integers, then $\boldsymbol{\pi} | \boldsymbol{\beta} \sim \mathscr{D}_{\alpha_0 \boldsymbol{\beta}}$.

An obvious but quite important feature of the HDP is how the mixing measures G_j share with one another the same set of supporting atoms (ϕ_k), while the probability mass are varying as random samples of another Dirichlet measure. This enables the sharing of "statistical strength" among different groups of data, and allows the information from one group to improve the inference of another group.

The stick-breaking representation described above can also be utliized to devise a conditional method of MCMC sampling for posterior inference with HDP model.

Chinese restaurant franchise Recall that Dirichlet processes arise as the directing random measure of Pólya sequences (of exchangeable observations), whose specification can be separated into the Chinese restaurant process of table assignment for an infinite sequence of customers arriving at a restaurant, and the i.i.d. selection of dishes for each table in the restaurant. As can be expected, the hierarchical Dirichlet processes also have a similar interpretation, aptly named by Teh et al. [2006] as the Chinese restaurant franchise.

A Chinese restaurant franchise is a random process involving an infinite collection of restaurants, each of which receives an infinite sequence of customers, who are going to be assigned to an infinite collection of tables. Moreover a dish is ordered in a random fashion for each assigned table. The details are as follows: The customers arrive at each restaurant j and assigned to a table according to the Chinese restaurant process governed by parameter α. The main novelty is in how a dish is selected for each table — whenever a customer is assigned to a new (empty) table in a restaurant, a dish is also selected in the following manner: with probability proportional to γ one selects the dish according to distribution H, while with probability proportional to the number of tables serving each distinct dish across the restaurant franchise, one selects among the existing dishes (with replacement). It can be seen that the random dishes selected by customers at restaurant j corresponds to an i.i.d. sample selected according to the measure G_j.

The culinary interpretation described above provides a basis for devising a simple marginal method of MCMC sampling for posterior inference with the HDP model.

10.2 Nested Dirichlet process

The nested Dirichlet process proposed by Rodriguez et al. [2008] is an even more audacious proposal for modeling a collection of exchangeable random probability measures. The first step is to assume that the collection of random measures $G_1, G_2, \ldots$ on Ω are an i.i.d. sample from a measure P of probability measures. To be clear, (G_j) are random elements in $\mathscr{P}(\Omega)$, while P is a (random) element taking value in $\mathscr{P}(\mathscr{P}(\Omega))$:

$$G_1, G_2, \ldots | P \overset{iid}{\sim} P.$$

To endow a prior for P using the Dirichlet process: let G_0 be a probability measure on Ω, we assume that

$$P|\alpha_0, G_0 \sim \mathscr{D}_{\gamma \mathscr{D}_{\alpha_0 G_0}}.$$

$\alpha_0 > 0$ and γ are two hyperparameters for the model. We may specify G_0 parametrically. It is also possible to endow G_0 with another Dirichlet process prior as in the HDP.

The nested Dirichlet process presents an extremely flexible prior for an i.i.d. collection of random measures, but this flexibility also implies a weakness. Because there are very little structure imposed other than the i.i.d. specification, the posterior inference may be very inefficient from a statistical standpoint. By speaking of statistical efficiency, we are entering a traditional domain of concern by theoretical statisticians.

11 Optimal transport based asymptotic theory

In addition to model specification and derivation of model-based inference algorithms, statisticians are also concerned with a perennial question: are we doing the right thing? Does the proposed procedure guarantee improved estimates as more data become available? How fast does the estimates converge to the true parameters of the model that is assumed to generate the observed data?

Suppose that data set $D_n = (X_1, \ldots, X_n)$ is an n-iid sample from the mixture density $p_G = \sum_{j=1}^{\infty} p_j f(x|\phi_j)$, where the mixing measure G takes some true value $G = G_0$. G_0 represents the truth that we want to estimate. Using Bayesian method, we may endow G with a Dirichlet process prior, and using MCMC algorithm we can obtain samples from posterior distribution of G given the data D_n. The posterior distribution of G is said to be consistent if it concentrates most its mass on the true G_0, as sample size n tends to infinity. To study the statistical efficiency of the posterior inference, we want to know how fast is the concentration of the posterior mass around the truth G_0.

General method for deriving posterior consistency and convergence of data densities have been fairly well-established (cf. [Ghosal et al., 2000]). For mixture models and general hierarchical models, one may be more interested in the behavior of the posterior distributions of latent variables as the sample size gets large. This has become an active research area lately. There are some progress that we now describe briefly.

We shall state a result of [Nguyen, 2013] for Dirichlet process mixtures. To study the convergence rate of the posterior of mixing measure G, a suitable choice of metric is called the Wasserstein metric, which arises in the theory of optimal transportation [Villani, 2008]. Let G and G' be two probability measures on a metric space Ω. A coupling between G and G' is a joint distribution which projects to marginal distribution G and G'. Then, the r-order Wasserstein distance between G and G' is viewed as the optimal cost of moving all the mass of G to that of G':

$$W_r(G, G') := \left(\inf_{\tau \in \kappa(G, G')} \|\theta - \theta'\|^r \, \tau(d\theta, d\theta') \right)^{1/r},$$

where $\kappa(G, G')$ denotes the set of all couplings between G and G'.

It turns out that the posterior concentration behavior the Dirichlet process mixtures depends strongly on the smoothness of the kernel density f. Let Ω a bounded subset of $\mathbb{R}^d$, the base probability measure G_0 of the Dirichlet prior behaves like an uniform distribution on Ω. Consider only kernel functions of the form $f(x|\phi) := f(x - \phi)$, where f can be either a ordinary smooth or a supersmooth function with parameter $\beta > 0$. For ordinary smooth kernel densities (e.g., Laplace kernel), let $\epsilon_n = (\log n/n)^{2/(d+2)(4+(2\beta+1)d')}$ for any $d' > d$. For supersmooth kernel densities (e.g., Gaussian kernel), let $\epsilon = (\log n)^{-1/\beta}$.

Theorem 11.1. *Let $\epsilon_n \downarrow 0$ be defined as above. Under some additional mild technical assumptions, the posterior distribution of G, which is induced from the Dirichlet process mixture model given the n-iid data, concentrates around the true G_0 at the rate of at least ϵ_n. That is,*

$$\mathbb{P}(W_2(G_0, G) \geq \epsilon_n | X_1, \ldots, X_n) \longrightarrow 0$$

in p_{G_0}-probability as $n \to \infty$.

A useful implication of this result is that when the kernel density function f is very smooth, the inference of latent mixing measure G and its associated parameter is highly inefficient. In other words, if one is interested only in clustering structures and estimation of mixing parameters, one should consider working with rough kernel densities rather than smooth ones. Very smooth kernel densities may be useful in approximating data densities very well via mixture modeling, but they become a detrimental huddle when it comes to the estimation of mixing measures' parameters.

To establish an asymptotic theory for hierarchical and nonparametric Bayesian model such as the hierarchical Dirichlet processes, one need to develop a suitable notion of metric on the space of probability measures of probability of measures. It turns out that optimal transport based distances continue to be a natural choice. In particular, we can also define a suitable notion of W_2 distance for two Dirichlet measures $\mathscr{D}_{\alpha G}$ and $\mathscr{D}_{\alpha' G'}$ (see Section 3 of Nguyen [to appear]). We can show the following inequality

$$W_2(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha' G'}) \geq W_2(G, G')$$

for any pairs of probability measures G, G', and $\alpha, \alpha' > 0$. Remarkably, equality is achieved if $\alpha = \alpha'$. That is, for any pair G, G' and $\alpha > 0$,

$$W_2(\mathscr{D}_{\alpha G}, \mathscr{D}_{\alpha G'}) = W_2(G, G').$$

This identity represents an invariant property for Dirichlet measures. It is an open question to ask whether the inequality suggests another unique characterization for Dirichlet measures. In any cases, these properties are crucial in the asymptotic theory for the HDP, as shown by Nguyen [to appear].

References

C. Antoniak. Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1: 353–355, 1973.

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

P. Feigin and R. Tweedie. Linear functionals and markov chains associated with dirichlet processes. *Math. Proc. Camb. Phil. Soc.*, 105:579–585, 1989.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28 (2):500–531, 2000.

J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics*. Springer, 2002.

N. Hjort, C. Holmes, P. Mueller, and S. Walker (Eds.). *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer-Verlag, New York, 2005.

A.Y. Lo. On a class of bayesian nonparametric estimates i: Density estimates. *Annals of Statistics*, 12(1): 351–357, 1984.

S. MacEachern and P. Mueller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

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

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400, 2013.

X. Nguyen. Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure. *Bernoulli*, arxiv.org/abs/1301.0802, to appear.

C. P. Robert. *The Bayesian Choice: From decision-theoretic foundations to computational implementations*. Springer, 2nd edition, 2007.

A. Rodriguez, D. Dunson, and A.E. Gelfand. The nested Dirichlet process. *J. Amer. Statist. Assoc.*, 103 (483):1131–1154, 2008.

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

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics–Simulation and Computation*, 36(1):45–54, 2007.