Notes on Bayesian Nonparametrics

In these notes[1] we present some of the most commonly used nonparametric Bayesian methods. These methods place priors on infinite dimensional spaces. The priors are based on certain stochastic processes called Dirichlet processes and Gaussian processes. In many cases, we cannot write down explicit formulas for the priors. Instead, we give explicit algorithms for drawing from the prior and the posterior.

 These notes are written at a more advanced and technical level than what we present in S&DS 365. They are offered to give further detail and as a complement to what is presented in class; students are not responsible for the parts that are not discussed in lecture. Please also see the readings in "Probabilistic Machine Learning."

1. What is Nonparametric Bayes?

In parametric Bayesian inference we have a model $\mathcal{M} = \{f(y \mid \theta) : \theta \in \Theta\}$ and data $Y_1, \ldots, Y_n \sim f(y \mid \theta)$. We put a prior distribution $\pi(\theta)$ on the parameter θ and compute the posterior distribution using Bayes' rule:

$$\pi(\theta \mid Y) = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{m(Y)} \tag{1}$$

where $Y = (Y_1, \ldots, Y_n)$, $\mathcal{L}_n(\theta) = \prod_i f(Y_i \mid \theta)$ is the likelihood function and

$$m(y) = m(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n \mid \theta)\pi(\theta)d\theta = \int \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)d\theta$$

is the marginal distribution for the data induced by the prior and the model. We call m the *induced marginal*. The model may be summarized as:

$$\begin{aligned} \theta &\sim \pi \\ Y_1, \ldots, Y_n \mid \theta &\sim f(y \mid \theta). \end{aligned}$$

We use the posterior to compute a point estimator such as the posterior mean of θ. We can also summarize the posterior by drawing a large sample $\theta_1, \ldots, \theta_N$ from the posterior $\pi(\theta \mid Y)$ and the plotting the samples.

In nonparametric Bayesian inference, we replace the finite dimensional model $\{f(y \mid \theta) : \theta \in \Theta\}$ with an infinite dimensional model such as

$$\mathcal{F} = \left\{ f : \int (f''(y))^2 dy < \infty \right\} \tag{2}$$

Typically, neither the prior nor the posterior have a density function with respect to a dominating measure. But the posterior is still well defined.

[1] These notes were written by Larry Wasserman and John Lafferty.

On the other hand, if there is a dominating measure for a set of densities $\mathcal{F}$ then the posterior can be found by Bayes theorem:

$$\pi_n(A) \equiv \mathbb{P}(f \in A \,|\, Y) = \frac{\int_A \mathcal{L}_n(f) d\pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f) d\pi(f)} \tag{3}$$

where $A \subset \mathcal{F}$, $\mathcal{L}_n(f) = \prod_i f(Y_i)$ is the likelihood function and π is a prior on $\mathcal{F}$. An estimate of f is the posterior mean

$$\widehat{f}(y) = \int f(y) d\pi_n(f). \tag{4}$$

A posterior $1-\alpha$ region is any set A such that $\pi_n(A) = 1-\alpha$.

If there is no dominating measure for $\mathcal{F}$ then the posterior still exists but cannot be obtained by simply applying Bayes' theorem. Several questions arise:

1. How do we construct a prior π on an infinite dimensional set $\mathcal{F}$?
2. How do we compute the posterior? How do we draw random samples from the posterior?
3. What are the properties of the posterior?

The answers to the third question are subtle. In finite dimensional models, the inferences provided by Bayesian methods usually are similar to the inferences provided by frequentist methods. Hence, Bayesian methods inherit many properties of frequentist methods: consistency, optimal rates of convergence, frequency coverage of interval estimates etc. In infinite dimensional models, this is no longer true, and the inferences provided by Bayesian methods do not necessarily coincide with frequentist methods.

2. Distributions on Infinite Dimensional Spaces

To use nonparametric Bayesian inference, we will need to put a prior π on an infinite dimensional space. For example, suppose we observe $X_1, \ldots, X_n \sim F$ where F is an unknown distribution. We will put a prior π on the set of all distributions $\mathcal{F}$. In many cases, we cannot explicitly write down a formula for π as we can in a parametric model. This leads to the following problem: how can we describe a distribution π on an infinite dimensional space? One way to describe such a distribution is to give an explicit algorithm for drawing from the distribution π. In a certain sense, "knowing how to draw from π" takes the place of "having a formual for π."

The Bayesian model can be written as

$$\begin{aligned} F &\sim \pi \\ X_1, \ldots, X_n \,|\, F &\sim F. \end{aligned}$$

The model and the prior induce a marginal distribution m for $(X_1, \ldots, X_n)$,

$$m(A) = \int \mathbb{P}_F(A) d\pi(F)$$

where

$$\mathbb{P}_F(A) = \int I_A(x_1, \ldots, x_n) dF(x_1) \cdots dF(x_n).$$

We call m the *induced marginal*. Another aspect of describing our Bayesian model will be to give an algorithm for drawing $X = (X_1, \ldots, X_n)$ from m.

After we observe the data $X = (X_1, \ldots, X_n)$, we are interested in the posterior distribution

$$\pi_n(A) \equiv \pi(F \in A \,|\, X_1, \ldots, X_n). \tag{5}$$

Once again, we will describe the posterior by giving an algorithm for drawing randonly from it.

To summarize: in some nonparametric Bayesian models, we describe the prior distribution by giving an algorithm for sampling from the prior π, the marginal m and the posterior π_n.

3. Four Nonparametric Problems

We will focus on four specific problems. The four problems and their most common frequentist and Bayesian solutions are:

Statistical Problem	**Frequentist Approach**	**Bayesian Approach**
Estimating a cdf	empirical cdf	Dirichlet process
Estimating a density	kernel density estimator	Dirichlet process mixture
Estimating a regression function	kernel smoother	Gaussian process
Estimating many multinomials	EM or empirical Bayes	hierarchical DPM

4. Estimating a cdf

Let $X_1, \ldots, X_n$ be a sample from an unknown cdf (cumulative distribution function) F where $X_i \in \mathbb{R}$. The usual frequentist estimate of F is the *empirical distribution function*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x). \tag{6}$$

The DKW inequality says that for every $\epsilon > 0$ and every F,

$$\mathbb{P}_F\Big(\sup_x |F_n(x) - F(x)| > \epsilon\Big) \leq 2e^{-2n\epsilon^2}. \tag{7}$$

Setting $\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$ we have

$$\inf_F \mathbb{P}_F\left(F_n(x) - \epsilon_n \leq F(x) \leq F_n(x) + \epsilon_n \ \text{ for all } x\right) \geq 1 - \alpha \tag{8}$$

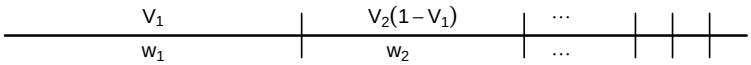

FIG 1. *The stick breaking process shows how the weights $w_1, w_2, \ldots$ from the Dirichlet process are constructed. First we draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$. Then we set $w_1 = V_1,\ w_2 = V_2(1 - V_1),\ w_3 = V_3(1 - V_1)(1 - V_2), \ldots$.*

where the infimum is over all cdf's F. Thus, $\big(F_n(x) - \epsilon_n, F_n(x) + \epsilon_n\big)$ is a $1 - \alpha$ confidence band for F.

To estimate F from a Bayesian perspective we put a prior π on the set of all cdf's $\mathcal{F}$ and then we compute the posterior distribution on $\mathcal{F}$ given $X = (X_1, \ldots, X_n)$. The most commonly used prior is the *Dirichlet process prior* which was invented by the statistician Thomas Ferguson in 1973.

The distribution π has two parameters, F_0 and α and is denoted by $\text{DP}(\alpha, F_0)$. The parameter F_0 is a distribution function and should be thought of as a prior guess at F. The number α controls how tightly concentrated the prior is around F_0. The model may be summarized as:

$$\begin{aligned} F &\sim \pi \\ X_1, \ldots, X_n \,|\, F &\sim F \end{aligned}$$

where $\pi = \text{DP}(\alpha, F_0)$.

How to Draw From the Prior

To draw a single random distribution F from $\text{Dir}(\alpha, F_0)$ we do the following steps:

1. Draw $s_1, s_2, \ldots$ independently from F_0.
2. Draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$.
3. Let $w_1 = V_1$ and $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ for $j = 2, 3, \ldots$.
4. Let F be the discrete distribution that puts mass w_j at s_j, that is, $F = \sum_{j=1}^{\infty} w_j \delta_{s_j}$ where δ_{s_j} is a point mass at s_j.

It is clear from this description that F is discrete with probability one. The construction of the weights $w_1, w_2, \ldots$ is often called the *stick breaking process*. Imagine we have a stick of unit length. Then w_1 is is obtained by breaking the stick a the random point V_1. The stick now has length $1 - V_1$. The second weight w_2 is obtained by breaking a proportion V_2 from the remaining stick. The process continues and generates the whole sequence of weights $w_1, w_2, \ldots$. See Figure 1. It can be shown that if $F \sim \text{Dir}(\alpha, F_0)$ then the mean is $\mathbb{E}(F) = F_0$.

You might wonder why this distribution is called a Dirichlet process. Recall that a random vector $P = (P_1, \ldots, P_k)$ has a Dirichlet distribution with parameters $(\alpha, g_1, \ldots, g_k)$ (with $\sum_j g_j = 1$) if the distribution of P has density

$$f(p_1, \ldots, p_k) = \frac{\Gamma(\alpha)}{\prod_{j=1}^{k} \Gamma(\alpha g_j)} \prod_{j=1}^{k} p_j^{\alpha g_j - 1}$$

over the simplex $\{p = (p_1, \ldots, p_k) :\ p_j \geq 0, \sum_j p_j = 1\}$. Let $(A_1, \ldots, A_k)$ be any partition of $\mathbb{R}$

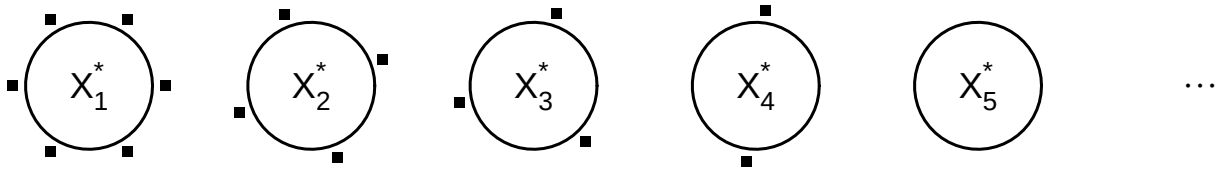

FIG 2. *The Chinese restaurant process. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.*

and let $F \sim \mathrm{DP}(\alpha, F_0)$ be a random draw from the Dirichlet process. Let $F(A_j)$ be the amount of mass that F puts on the set A_j. Then $(F(A_1), \ldots, F(A_k))$ has a Dirichlet distribution with parameters $(\alpha F_0(A_1), \ldots, \alpha F_0(A_k))$. In fact, this property characterizes the Dirichlet process.

4.1. How to Sample From the Marginal

One way is to draw from the induced marginal m is to sample $F \sim \pi$ (as described above) and then draw $X_1, \ldots, X_n$ from F. But there is an alternative method, called the *Chinese Restaurant Process* or *infinite Pólya urn* (Blackwell and MacQueen, 1973). The algorithm is as follows.

1. Draw $X_1 \sim F_0$.
2. For $i = 2, \ldots, n$: draw

$$X_i \,|\, X_1, \ldots X_{i-1} = \begin{cases} X \sim F_{i-1} & \text{with probability } \frac{i-1}{i+\alpha-1} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{i+\alpha-1} \end{cases}$$

   where F_{i-1} is the empirical distribution of $X_1, \ldots X_{i-1}$.

The sample $X_1, \ldots, X_n$ is likely to have ties since F is discrete. Let $X_1^*, X_2^*, \ldots$ denote the unique values of $X_1, \ldots, X_n$. Define cluster assignment variables $c_1, \ldots, c_n$ where $c_i = j$ means that X_i takes the value X_j^*. Let $n_j = |\{i : \ c_j = j\}|$. Then we can write

$$X_n = \begin{cases} X_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1}. \end{cases}$$

In the metaphor of the Chinese restaurant process, when the nth customer walks into the restaurant, he sits at table j with probability $n_j/(n+\alpha-1)$, and occupies a new table with probability $\alpha/(n+\alpha-1)$. The jth table is associated with a "dish" $X_j^* \sim F_0$. Since the process is exchangeable, it induces (by ignoring X_j^*) a partition over the integers $\{1, \ldots, n\}$, which corresponds to a clustering of the indices. See Figure 2.

4.2. How to Sample From the Posterior

Now suppose that $X_1, \ldots, X_n \sim F$ and that we place a $\mathrm{Dir}(\alpha, F_0)$ prior on F.

Theorem 4.1. *Let $X_1, \ldots, X_n \sim F$ and let F have prior $\pi = \mathrm{Dir}(\alpha, F_0)$. Then the posterior π for F given $X_1, \ldots, X_n$ is $\mathrm{Dir}\left(\alpha + n, \overline{F}_n\right)$ where*

$$\overline{F}_n = \frac{n}{n+\alpha} F_n + \frac{\alpha}{n+\alpha} F_0. \tag{9}$$

Since the posterior is again a Dirichlet process, we can sample from it as we did the prior but we replace α with $\alpha + n$ and we replace F_0 with $\overline{F}_n$. Thus the posterior mean is $\overline{F}_n$ is a convex combination of the empirical distribution and the prior guess F_0. Also, the predictive distribution for a new observation X_{n+1} is given by $\overline{F}_n$.

To explore the posterior distribution, we could draw many random distribution functions from the posterior. We could then numerically construct two functions L_n and U_n such that

$$\pi\big(L_n(x) \leq F(x) \leq U_n(x) \text{ for all x} \,|\, X_1, \ldots, X_n\big) = 1 - \alpha.$$

This is a $1 - \alpha$ Bayesian confidence band for F. Keep in mind that this is not a frequentist confidence band. It does *not* guarantee that

$$\inf_F \mathbb{P}_F(L_n(x) \leq F(x) \leq U_n(x) \text{ for all x}) = 1 - \alpha.$$

When n is large, $\overline{F}_n \approx F_n$ in which case there is little difference between the Bayesian and frequentist approach. The advantage of the frequentist approach is that it does not require specifiying α or F_0.

Example 4.2. Figure 3 shows a simple example. The prior is $\mathrm{DP}(\alpha, F_0)$ with $\alpha = 10$ and $F_0 = N(0, 1)$. The top left plot shows the discrete probabilty function resulting from a single draw from the prior. The top right plot shows the resulting cdf along with F_0. The bottom left plot shows a few draws from the posterior based on $n = 25$ observations from a N(5,1) distribution. The blue line is the posterior mean and the red line is the true F. The posterior is biased because of the prior. The bottom right plot shows the empirical distribution function (solid black) the true F (red) the Bayesian postrior mean (blue) and a 95 percnt frequentist confidence band.

5. Density Estimation

Let $X_1, \ldots, X_n \sim F$ where F has density f and $X_i \in \mathbb{R}$. Our goal is to estimate f. The Dirichlet process is not a useful prior for this problem since it produces discrete distributions which do not even have densities. Instead, we use a modification of the Dirichlet process. But first, let us review the frequentist approach.

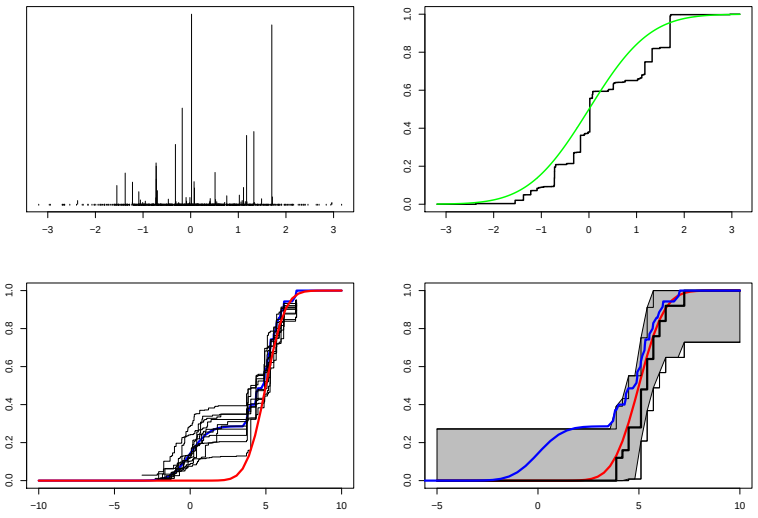

FIG 3. *The top left plot shows the discrete probabilty function resulting from a single draw from the prior which is a* $\mathrm{DP}(\alpha, F_0)$ *with* $\alpha = 10$ *and* $F_0 = N(0,1)$. *The top right plot shows the resulting cdf along with* F_0. *The bottom left plot shows a few draws from the posterior based on* $n = 25$ *observations from a N(5,1) distribution. The blue line is the posterior mean and the red line is the true* F. *The posterior is biased because of the prior. The bottom right plot shows the empirical distribution function (solid black) the true* F *(red) the Bayesian postrior mean (blue) and a 95 percent frequentist confidence band.*

The most common frequentist estimator is the kernel estimator

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where K is a kernel and h is the bandwidth. A related method for estimating a density is to use a mixture model

$$f(x) = \sum_{j=1}^{k} w_j f(x; \theta_j).$$

For example, of $f(x;\theta)$ is Normal then $\theta = (\mu, \sigma)$. The kernel estimator can be thought of as a mixture with n components. In the Bayesian approach we would put a prior on $\theta_1, \ldots, \theta_k$, on $w_1, \ldots, w_k$ and a prior on k. We could be more ambitious and use an infinite mixture

$$f(x) = \sum_{j=1}^{\infty} w_j f(x; \theta_j).$$

As a prior for the parameters we could take $\theta_1, \theta_2, \ldots$ to be drawn from some F_0 and we could take $w_1, w_2, \ldots,$ to be drawn from the stick breaking prior. (F_0 typically has parameters that require further priors.) This infinite mixture model is known as the Dirichlet process mixture model (Escobar and West, 1995). This infinite mixture is the same as the random distribution $F \sim \mathrm{DP}(\alpha, F_0)$ which had the form $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$ except that the point mass distributions δ_{θ_j} are replaced by smooth densities $f(x \,|\, \theta_j)$.

The model may be re-expressed as:

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, F_0) && (10)\\ \theta_1, \ldots, \theta_n \,|\, F &\sim F && (11)\\ X_i \,|\, \theta_i &\sim f(x \,|\, \theta_i), \quad i = 1, \ldots, n. && (12) \end{aligned}$$

(In practice, F_0 itself has free parameters which also require priors.) Note that in the DPM, *the parameters θ_i of the mixture are sampled from a Dirichlet process. The data X_i are not sampled from a Dirichlet process.* Because F is sampled from from a Dirichlet process, it will be discrete. Hence there will be ties among the θ_i's. (Recall our earlier discussion of the Chinese Restaurant Process.) The $k < n$ distinct values of θ_i can be thought of as defining clusters. The beauty of this model is that the discreteness of F automatically creates a clustering of the θ_j's. In other words, we have implicitly created a prior on k, the number of distinct θ_j's.

5.1. How to Sample From the Prior

Draw $\theta_1, \theta_2, \ldots, F_0$ and draw $w_1, w_2, \ldots,$ from the stsick breaking process. Set $f(x) = \sum_{j=1}^{\infty} w_j f(x; \theta_j)$. The density f is a random draw from the prior. We could repeat this process many times resulting in many randomly drawn densities from the prior. Plotting these densities could give some intuition about the structure of the prior.

5.2. How to Sample From the Prior Marginal

The prior marginal m is

$$\begin{aligned} m(x_1, x_2, \ldots, x_n) &= \int \prod_{i=1}^{n} f(x_i \,|\, F)\, d\pi(F) && (13)\\ &= \int \prod_{i=1}^{n} \left(\int f(x_i \,|\, \theta)\, p(\theta \,|\, F)\, dF(\theta) \right) dP(G) && (14) \end{aligned}$$

If we wnat to draw a sample from m, we first draw F from a Dirichlet process with parameters α and F_0, and then generate θ_i independently from this realization. Then we sample $X_i \sim f(x \,|\, \theta_i)$.

As before, we can also use the Chinese restaurant representation to draw the θ_j's sequentially. Given $\theta_1, \ldots, \theta_{i-1}$ we draw θ_j from

$$\alpha F_0(\cdot) + \sum_{i=1}^{n-1} \delta_{\theta_i}(\cdot).$$

Let θ_j^* denote the unique values among the θ_i, with n_j denoting the number of elements in the cluster for parameter θ_i^*; that is, if $c_1, c_2, \ldots, c_{n-1}$ denote the cluster assignments $\theta_i = \theta_{c_i}^*$ then $n_j = |\{i : c_i = j\}|$. Then we can write

$$\theta_n = \begin{cases} \theta_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1}. \end{cases}$$

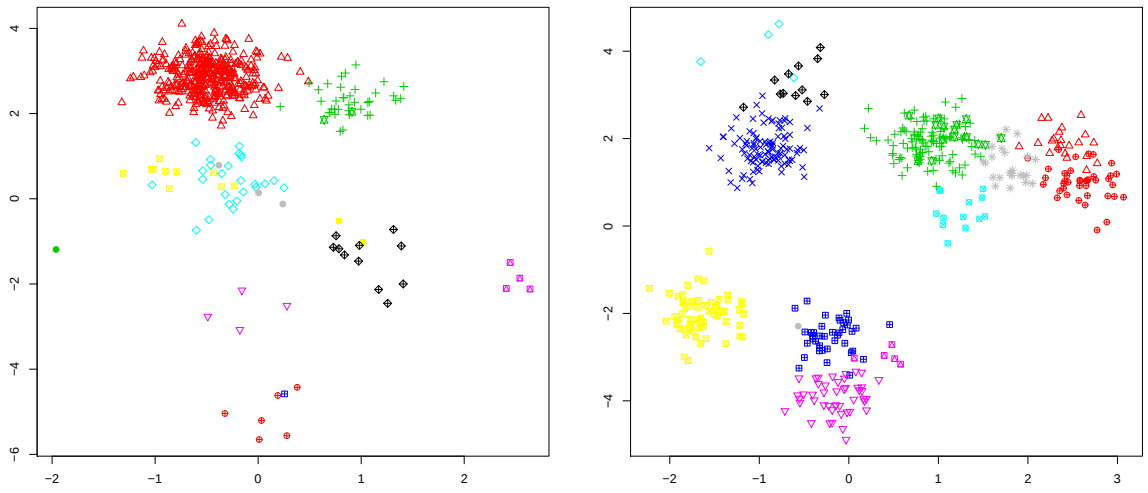

FIG 4. *Samples from a Dirichlet process mixture model with Gaussian generator,* $n = 500$.

5.3. How to Sample From the Posterior

We sample from the posterior by Gibbs sampling. Our ultimate goal is to approximate the predictive distribuiton of a new observation x_{n+1}:

$$\widehat{f}(x_{n+1}) \equiv f(x_{n+1} \,|\, x_1, \ldots, x_n).$$

This density is our Bayesian density estimator.

The Gibbs sampler for the DP mixture is straightforward in the case where the base distribution F_0 is conjugate to the data model $f(x \,|\, \theta)$. Recall that if $f(x \,|\, \theta)$ is in the exponential family it can be written in the (canonical) natural parameterization as

$$f(x \,|\, \theta) = h(x) \exp\left(\theta^T x - a(\theta)\right)$$

The *conjugate prior* for this model takes the form

$$p(\theta \,|\, \lambda = \{\lambda_1, \lambda_2\}) = g(\theta) \exp\left(\lambda_1^T \theta - \lambda_2 a(\theta) - b(\lambda_1, \lambda_2)\right)$$

Here $a(\theta)$ is the moment generating function (log normalizing constant) for the original model, and $b(\lambda)$ is the moment generating function for the prior. The parameter of the prior has two parts, corresponding to the two components of the vector of sufficient statistics $(\theta, -a(\theta))$. The parameter λ_1 has the same dimension as the parameter θ of the model, and λ_2 is a scalar. To verify conjugacy, note that

$$\begin{aligned} p(\theta \,|\, x, \lambda) &\propto p(x \,|\, \theta)\, p(\theta \,|\, \lambda) && (15) \\ &\propto h(x) \exp(\theta^T x - a(\theta)) g(\theta) \exp(\lambda_1^T \theta - \lambda_2 a(\theta) - b(\lambda_1, \lambda_2)) && (16) \\ &\propto g(\theta) \exp((x + \lambda_1)^T \theta - (\lambda_2 + 1) a(\theta)) && (17) \end{aligned}$$

The factor $h(x)$ drops out in the normalization. Thus, the parameters of the posterior are $\lambda = (\lambda_1 + x, \lambda_2 + 1)$.

Example 5.3. Take $p(\cdot \mid \mu)$ be normal with known variance. Thus,

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{18}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)}_{h(x)} \exp\left(\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \tag{19}$$

Let $\nu = \frac{\mu}{\sigma^2}$ be the natural parameter. Then

$$p(x \mid \nu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)}_{h(x)} \exp\left(x\nu - \frac{\nu^2\sigma^2}{2}\right) \tag{20}$$

Thus, $a(\nu) = \nu^2\sigma^2/2$. The conjugate prior then takes the form

$$p(\mu \mid \lambda_1, \lambda_2) = g(\mu) \exp\left(\lambda_1\mu - \lambda_2\frac{\mu^2\sigma^2}{2} - b(\lambda_1, \lambda_2)\right) \tag{21}$$

where $b(\lambda_1, \lambda_2)$ is chosen so that the prior integrates to one.

Under conjugacy, the parameters $\theta_1, \ldots, \theta_n$ can be integrated out, and the Gibbs sampling is carried out with respect to the cluster assignments $c_1, \ldots, c_n$. Let c_{-i} denote the vector of the $n-1$ cluster assignments for all data points other than i. The Gibbs sampler cycles through indices i according to some schedule, and sets $c_i = k$ according to the conditional probability

$$p(c_i = k \mid x_{1:n}, c_{-i}, \lambda) \tag{22}$$

This either assigns c_i to one of the existing clusters, or starts a new cluster. By the chain rule, we can factor this conditional probability as

$$p(c_i = k \mid x_{1:n}, c_{-i}, \lambda) = p(c_i = k \mid c_{-i})\, p(x_i \mid x_{-i}, c_{-i}, c_i = k, \lambda) \tag{23}$$

The class assignment probability $p(c_i = k \mid c_{-i})$ is governed by the Pólya urn scheme:

$$p(c_i = k \mid c_{-i}) \propto \begin{cases} \#\{j : c_j = k, j \neq i\} & \text{if } k \text{ is an existing cluster} \\ \alpha & \text{if } k \text{ is a new cluster} \end{cases} \tag{24}$$

The conditional probability of x_i is, by conjugacy, given by

$$p(x_i \mid x_{-i}, c_{-i}, c_i = k, \lambda) = \tag{25}$$

$$= p(x_i \mid \text{other } x_j \text{ in cluster } k, \lambda) \tag{26}$$

$$= \int p(x_i \mid \theta) p(\theta \mid \text{other } x_j \text{ in cluster } k, \lambda) \tag{27}$$

$$= \frac{\exp\left(b\left(\lambda_1 + \sum_{j\neq i} 1[c_j = k]x_j + x_i,\ \lambda_2 + \sum_{j\neq i} 1[c_j = k] + 1\right)\right)}{\exp\left(b\left(\lambda_1 + \sum_{j\neq i} 1[c_j = k]x_j,\ \lambda_2 + \sum_{j\neq i} 1[c_j = k]\right)\right)} \tag{28}$$

The probability of x_i conditioned on the event that it starts a new cluster is

$$p(x_i \mid F_0) = \int p(x_i \mid \theta)\, dF_0(\theta) \tag{29}$$

$$= \exp\left(b(x_i + \lambda_1, \lambda_1 + 1)\right) \tag{30}$$

The algorithm iteratively updates the cluster assignments in this manner, until convergence.

After appropriate convergence has been determined, the approximation procedure is to collect a set of partitions $c^{(b)}$, for $b = 1, \ldots, B$. The predictive distribution is then approximated as

$$p(x_{n+1} \mid x_{1:n}, \lambda, \alpha, F_0) \approx \frac{1}{B} \sum_{b=1}^{B} p(x_{n+1} \mid c_{1:n}^{(b)}, x_{1:n}, \lambda, \alpha, F_0)$$

where the probabilities are computed just as in the Gibbs sampling procedure, as described above.

If the base measure F_0 is not conjugate, MCMC is significantly more complicated and problematic in high dimensions. See Neal (2000) for a discussion of MCMC algorithms for this case.

The Mean Field Approximation. An alternative to sampling is to use an approximation. Recall that in the mean field variational approximation, we treat all of the variables as independent, and assume a fully factorized variational approximation q. The strategy is then to maximize the lower bound on the data likelihood, or equivalently to minimize the KL divergence $D(q \,\|\, p)$ with respect to the variational parameters that determine q.

In this setting, the variables we are integrating over are θ_j^* and V_j, for the infinite sequence $j = 1, 2, \ldots,$ together with the mixture component indicator variables Z_i, for $i = 1, 2, \ldots, n$. Since it is of course not possible to implement an infinite model explicitly, we take a finite variational approximation that corresponds to breaking the stick into T pieces. Thus, we take

$$q(V_{1:T}, \theta_{1:T}^*, Z_{1:n}) = \prod_{t=1}^{T-1} q_{\gamma_t}(V_t) \prod_{t=1}^{T} q_{\tau_t}(\theta_t^*) \prod_{i=1}^{n} q_{\phi_i}(Z_i) \tag{31}$$

where each factor has its own variational parameter. Each q_{γ_t} is a beta distribution, each q_{τ_t} is a conjugate distribution over θ_t^*, and each q_{ϕ_i} is a $(T-1)$-dimensional multinomial distribution. Note that while there are T mixture components in the variational approximation, the model itself is not truncated.

Let λ denote the parameters of the conjugate distribution F_0, as we did above for the Gibbs sampler. According to the standard variational procedure, we then bound the log marginal probability of the data from below as

$$\log p(x_{1:n} \mid \alpha, \lambda) \geq \tag{32}$$

$$\mathbb{E}_q[\log p(V \mid \alpha)] + \mathbb{E}_q[\log p(\theta^* \mid \lambda)] + \sum_{i=1}^{n} \left(\mathbb{E}_q[\log \pi_{Z_i}] + \mathbb{E}_q[\log p(x_i \mid \theta_{Z_i}^*)]\right)$$

$$+ \sum_{t=1}^{T-1} H(q_{\gamma_t}) + \sum_{t=1}^{T} H(q_{\tau_t}) + \sum_{i=1}^{n} H(q_{\phi_i}) \tag{33}$$

where H denotes entropy. For details on a coordinate ascent algorithm to optimize this lower bound as a function of the variational parameters, see (Blei and Jordan, 2005).

To estimate the predictive distribution, note first that the true predictive distribution under the stick breaking representation is given by

$$p(x_{n+1} \mid x_{1:n}, \alpha, \lambda) = \int \sum_{t=1}^{\infty} \pi_t(v)\, p(x_{n+1} \mid \theta_t^*)\, dP(v, \theta^* \mid x, \lambda, \alpha) \tag{34}$$

We approximate this by replacing the true stick breaking distribution with the variational distribution. Since, under the variational approximation, the mixture is truncated and the V and θ^* variables are conditionally independent, the approximated predictive distribution is thus

$$p(x_{n+1} \mid x_{1:n}, \alpha, \lambda) \approx \sum_{t=1}^{T} \mathbb{E}_q[\pi_t(V)]\, \mathbb{E}_q(p(x_{n+1} \mid \theta_t^*)). \tag{35}$$

6. Gaussian process regression

Consider the nonparametric regression model

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where $\mathbb{E}(\epsilon_i) = 0$. The frequentist kernel estimator for m is

$$\widehat{m}(x) = \frac{\sum_{i=1}^n Y_i\, K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}$$

where K is a kernel and h is a bandwidth. The Bayesian version requires a prior π on the set of regression functions $\mathcal{M}$. A common choice is the *Gaussian process prior*.

A stochastic process $m(x)$ indexed by $x \in \mathcal{X} \subset \mathbb{R}^d$ is a *Gaussian process* if for each $x_1, \ldots, x_n \in \mathcal{X}$ the vector $(m(x_1), m(x_2), \ldots, m(x_n))$ is normally distributed:

$$(m(x_1), m(x_2), \ldots, m(x_n)) \sim N(\mu(x), K(x))$$

where $K_{ij}(x) = K(x_i, x_j)$ is a Mercer kernel. When $x_1, \ldots, x_n$ are fixed we will denote the $n \times n$ matrix with entries $K(x_i, x_j)$ by $\mathbb{K}$.

What functions have high probability according to the Gaussian process prior? The prior favors $m^T \mathbb{K}^{-1} m$ being small. Suppose we consider an eigenvector v of $\mathbb{K}$, with eigenvalue λ, so that $\mathbb{K}v = \lambda v$. Then we have that

$$\frac{1}{\lambda} = v^T \mathbb{K}^{-1} v \tag{36}$$

Thus, eigenfunctions of the Mercer kernel K with *large* eigenvalues are favored by the prior. These correspond to smooth functions; the eigenfunctions that are very wiggly correspond to small eigenvalues.

Let's assume that $\mu = 0$, so the prior mean function is zero. Then for given $x_1, x_2, \ldots, x_n$ the density of the Gaussian process prior of $m = (m(x_1), \ldots, m(x_n))$ is

$$\pi(m) = (2\pi)^{-n/2}|\mathbb{K}|^{-1/2} \exp\left(-\frac{1}{2} m^T \mathbb{K}^{-1} m\right).$$

Under the change of variables $m = \mathbb{K}\alpha$, we have that $\alpha \sim N(0, \mathbb{K}^{-1})$ and thus

$$\pi(\alpha) = (2\pi)^{-n/2}|\mathbb{K}|^{-1/2} \exp\left(-\frac{1}{2} \alpha^T \mathbb{K} \alpha\right).$$

Under the standard Gaussian noise model, we observe $Y_i = m(x_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Thus, the log-likelihood is

$$\log p(Y \mid m) = -\frac{1}{2\sigma^2} \sum_i (Y_i - m(x_i))^2 + C$$

where $C = -\log(\sqrt{2\pi\sigma^2})$ is an additive constant that does not enter into our calculations. The log-posterior is

$$\begin{aligned} \log p(Y \mid m) + \log \pi(m) &= -\frac{1}{2\sigma^2} \|Y - \mathbb{K}\alpha\|_2^2 - \frac{1}{2} \alpha^T \mathbb{K} \alpha + C & (37) \\ &= -\frac{1}{2\sigma^2} \|Y - \mathbb{K}\alpha\|_2^2 - \frac{1}{2} \|\alpha\|_K^2 + C & (38) \end{aligned}$$

(note that the irrelevant constant C can change from line to line).

In Bayesian *maximum a posteriori* (MAP) inference, one estimates the mode of the posterior. Here, MAP estimation corresponds to Mercer kernel regression, which regularizes the squared error by the RKHS norm $\|\alpha\|_K^2$. The posterior mean is

$$\mathbb{E}(\alpha \mid Y) = \left(\mathbb{K} + \sigma^2 I\right)^{-1} Y$$

and thus

$$\widehat{m} = \mathbb{E}(m \mid Y) = \mathbb{K}\left(\mathbb{K} + \sigma^2 I\right)^{-1} Y.$$

We see that $\widehat{m}$ is nothing but a linear smoother and is, in fact, very similar to the frequentist kernel smoother. Unlike kernel regression, where we just need to choose a bandwidth h, here we need to choose the entire covariance function $K(x, y)$. This is what controls the level of Bayesian uncertainty in the model.

To calculate the posterior mean and variance at new points, we need a basic fact about the multidimensional Gaussian distribution.

Gaussian Conditionals

If (X_1, X_2) are jointly Gaussian with distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\right)$$

then the conditional distributions are also Gaussian and given by

$$\begin{aligned} X_1 \,|\, X_2 &\sim N\left(\mu + CB^{-1}(X_2 - \nu), A - CB^{-1}C^T\right) \\ X_2 \,|\, X_1 &\sim N\left(\nu + C^T A^{-1}(X_1 - \mu), B - C^T A^{-1} C\right) \end{aligned}$$

The matrix $A - CB^{-1}C^T$ is called the *Schur complement* of B.

We will use this to compute the predictive distribution for a new point $Y_{n+1} = m(x_{n+1}) + \epsilon_{n+1}$. Let k be the vector

$$k = (K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1})).$$

Then using the above fact we see that $(Y_1, \ldots, Y_{n+1})$ is jointly Gaussian with covariance

$$\begin{pmatrix} \mathbb{K} + \sigma^2 I & k \\ k^T & k(x_{n+1}, x_{n+1}) + \sigma^2 \end{pmatrix}.$$

Therefore, the conditional distribution of Y_{n+1} is given by the following expression.

Predictive distribution for a Gaussian process

The predictive distribution for a Gaussian process conditioned on data $x_1, \ldots, x_n$ and responses $Y_1, \ldots, Y_n$ is

$$Y_{n+1} \,|\, Y_{1:n}, x_{1:n} \sim N\left(k^T(\mathbb{K} + \sigma^2 I)^{-1} Y,\ k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(\mathbb{K} + \sigma^2 I)^{-1} k\right).$$

Thus, the posterior mean and variance are

$$\begin{aligned} \mathbb{E}(Y_{n+1} \,|\, x_{1:n}, Y_{1:n}) &= k^T(\mathbb{K} + \sigma^2 I)^{-1} Y \\ \text{Var}(Y_{n+1} \,|\, x_{1:n}, Y_{1:n}) &= k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(\mathbb{K} + \sigma^2 I)^{-1} k. \end{aligned}$$

Note that the above variance differs from the variance estimated using the frequentist method. However, Bayesian Gaussian process regression and kernel regression often lead to similar results. The advantages of the kernel regression is that it requires a single parameter h that can be chosen by cross-validation and its theoretical properties are simple and well-understood.

From a computational perspective, Gaussian process regression (just like Mercer kernel regression) can be computationally demanding, because the matrix inverse $(\mathbb{K} + \sigma^2 I)^{-1}$ in general requires $O(n^3)$ operations.

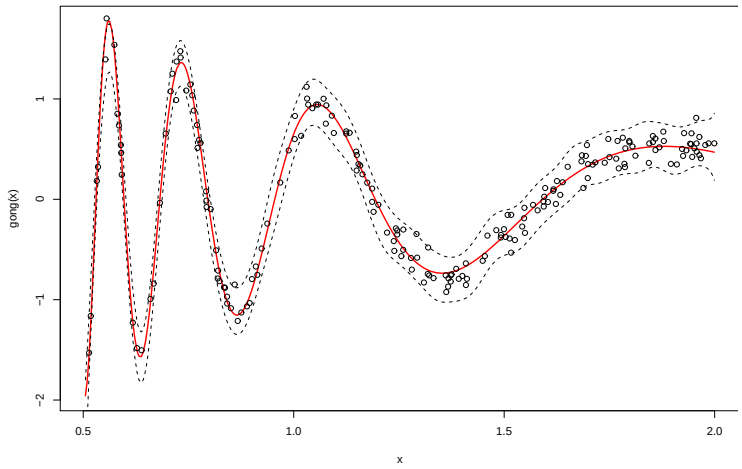

FIG 5. *Mean of a Gaussian process*

7. Estimating Many Multinomials

In many domains, the data naturally fall into groups. In such cases, we may want to model each group using a mixture model, while sharing the mixing components from group to group. For instance, text documents are naturally viewed as groups of words. Each document might be modeled as being generated from a mixture of *topics*, where a topic assigns high probability to words from a particular semantic theme, with the topics shared across documents—a finance topic might assign high probability to words such as "earnings," "dividend," and "report." Other examples arise in genetics, where each individual has a genetic profile exhibiting a pattern of haplotypes, which can be modeled as arising from a mixture of several ancestral populations.

As we've seen, Dirichlet process mixtures enable the use of mixture models with a potentially infinite number of mixture components, allowing the number of components to be selected adaptively from the data. A *hierarchical Dirichlet process mixture* is an extension of the Dirichlet process mixture to grouped data, where the mixing components are shared between groups.

Hierarchical modeling is an important method for "borrowing strength" across different populations. In a simple hierarchical model that allows for different disease rates in m different cities, where n_i people are selected from the ith city, and we observe how many people X_i have the disease being studied. We can think of the probability θ_i of the disease as a random draw from some distribution G_0, so that the hierarchical model can be written as

$$\begin{aligned} \text{For each } j = 1, \ldots, m \quad &: \\ \theta_j \quad &\sim \quad G_0 \\ X_j \,|\, n_j, \theta_j \quad &\sim \quad \text{Binomial}(n_i, \theta_j). \end{aligned}$$

It is then of interest to estimate the parameters θ_i for each city, or the overall disease rate $\int \theta \, d\pi(\theta)$, tasks that can be carried out using Gibbs sampling. This hierarchical model is shown in Figure 6.

To apply such a hierarchical model to grouped data, suppose that $F(\theta)$ is a family of distributions

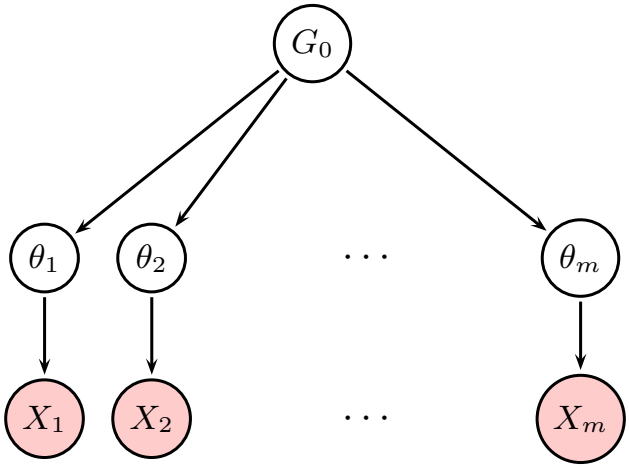

FIG 6. *A hierarchical model. The parameters θ_i are sampled conditionally independently from G_0, and the observations X_i are made within the ith group. The hierarchical structure statistically couples together the groups.*

for $\theta \in \Theta$, and $G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ is a (potentially) infinite mixture of the distributions $\{F(\theta_i)\}$, where $\sum_i \pi_i = 1$ and $\theta_i \in \Theta$. We denote sampling from this mixture as $X \sim \text{Mix}(G, F)$, meaning the two-step process

$$\begin{aligned} Z \,|\, \pi &\sim \text{Mult}(\pi) \\ X \,|\, Z &\sim F(\theta_Z). \end{aligned}$$

Here's a first effort at forming a nonparametric Bayesian model for grouped data. For each group, draw G_j from a Dirichlet process $\text{DP}(\gamma, H)$. Then, sample the data within group j from the mixture model specified by G_j. Thus:

For each $j = 1, \ldots, m$:

(a) Sample $G_j \,|\, \gamma, H \sim \text{DP}(\gamma, H)$

(b) For each $i = 1, \ldots, n_j$:

Sample $X_{ji} \,|\, G_j \sim \text{Mix}(G_j, F), \quad i = 1 \ldots, n_j.$

This process, however, does not satisfy the goal of statistically tying together the groups: each G_j is discrete, and for $j \neq k$, the mixtures G_j and G_k will not share any atoms, with probability one.

A simple and elegant solution, proposed by Teh et al. (2006), is to add a layer to the hierarchy, by first sampling G_0 from a Dirichlet process $\text{DP}(\gamma, H)$, and then sampling each G_j from the Dirichlet process $\text{DP}(\alpha_0, G_0)$. Drawing G_0 from a Dirichlet process ensures (with probability one) that it is a discrete measure, and therefore that the discrete measures $G_j \sim \text{DP}(\alpha_0, G_0)$ have the opportunity to share atoms. This leads to the following procedure. The model is shown graphically in Figure 7.

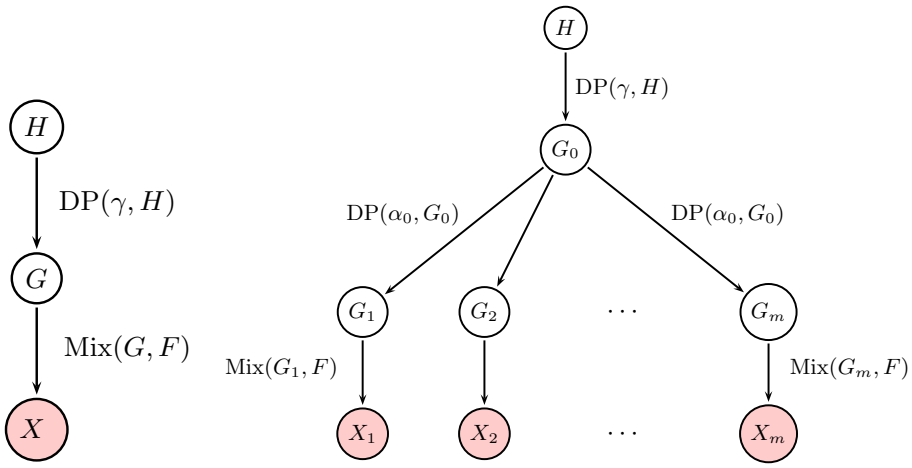

FIG 7. *Left: A Dirichlet process mixture. Right: A hierarchical Dirichlet process mixture. An extra layer is added to the hierarchy to ensure that the mixtures G_j share atoms, by forcing the measure G_0 to be discrete.*

Generative process for a hierarchical Dirichlet process mixture

1. Sample $G_0 \,|\, \gamma, H \sim \text{DP}(\gamma, H)$
2. For each $j = 1, \ldots, m$:
   (a) Sample $G_j \,|\, \alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0)$
   (b) For each $i = 1, \ldots, n_j$:
   $$\text{Sample } X_{ij} \,|\, G_j \sim \text{Mix}(G_j, F), \quad i = 1 \ldots, n_j.$$

The hierarchical Dirichlet process can be viewed as a hierarchical K-component mixture model, in the limit as $K \to \infty$, in the following way. Let

$$\begin{aligned}
\theta_k \,|\, H &\sim H, \quad k = 1, \ldots, K \\
\beta \,|\, \gamma &\sim \text{Dir}(\gamma/K, \ldots, \gamma/K) \\
\pi_j \,|\, \alpha_0, \beta &\sim \text{Dir}(\alpha_0\beta_1, \ldots, \alpha_0\beta_K), \quad j = 1, \ldots, m \\
X_{ji} \,|\, \pi_j, \theta &\sim \text{Mix}\left(\sum_{k=1}^{K} \pi_{jk}\delta_{\theta_k}, F\right), \quad i = 1, \ldots, n_j.
\end{aligned}$$

The marginal distribution of X converges to the hierarchical Dirichlet process as $K \to \infty$. This finite version was used for statistical language modeling by MacKay and Peto (1994).

7.1. Gibbs Sampling for the HDP

Suppose that the distribution H, which generates models θ, is conjugate to the data distribution F; this allows θ to be integrated out, circumventing the need to directly sample it. In this case, it is possible to derive several efficient Gibbs sampling algorithms for the hierarchical Dirichlet process

mixture. We summarize one such algorithm here; see Teh et al. (2006) for alternatives and further details.

The Gibbs sampler iteratively samples the variables $\beta = (\beta_1, \beta_2, \ldots)$, which are the weights in the stick breaking representation of G_0, and the variables $z_j = (z_{j1}, z_{j2}, \ldots, z_{jn_j})$ indicating which mixture component generates the jth data group $x_j = (x_{j1}, x_{j2}, \ldots, x_{jn_j})$. In addition, note that in the mixture $G_j = \sum_{k=1}^{\infty} \pi_{ji} \delta_{\theta_{ji}}$ for the jth group, an atom θ_k of G_0 can appear multiple times. The variable m_{jk} indicates the number of times component k appears in G_j; this is also stochastically sampled in the Gibbs sampler. A dot is used to denote a marginal count; thus, $m_{\cdot k} = \sum_{j=1}^{m} m_{jk}$ is the number of times component k appears in the mixtures $G_1, \ldots, G_m$. We denote

$$n_{j\cdot k} = \sum_{i=1}^{n_j} 1[z_{ji} = k]$$

which is the number of times component k is used in generating group j. These variables can be given mnemonic interpretations in terms of the "Chinese restaurant franchise" (Teh et al., 2006), an extension of the Chinese restaurant process metaphor to grouped data. Finally, the superscript $\backslash ji$ denotes that the ith element of the jth group is held out of a calculation. In particular,

$$n_{j\cdot k}^{\backslash ji} = \sum_{i' \neq i} 1[z_{ji'} = k]$$

Finally, we use the notation

$$f_k^{\backslash ji}(x_{ji}) = \frac{\int f(x_{ji} \,|\, \theta_k) \prod_{la \neq ji,\, z_{la}=k} f(x_{la} \,|\, \theta_k)\, h(\theta_k)\, d\theta_k}{\int \prod_{la \neq ji,\, z_{la}=k} f(x_{la} \,|\, \theta_k)\, h(\theta_k)\, d\theta_k}$$

to denote the conditional density of x_{ji} under component k, given all of the other data generated from this component. Here $f(\cdot \,|\, \theta)$ is the density of $F(\theta)$ and $h(\theta)$ is the density of $H(\theta)$. Under the conjugacy assumption, the integrals have closed form expressions.

Using this notation, the Gibbs sampler can be expressed as follows. At each point in the algorithm, a (random) number K of mixture components are active, with weights $\beta_1, \ldots, \beta_K$ satisfying $\sum_{j=1}^{K} \beta_j \leq 1$. A weight $\beta_u \geq 0$ is left for an as yet "unassigned" component k_{new}. These weights are updated according to

$$(\beta_1, \ldots, \beta_K, \beta_u) \quad \sim \quad \text{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma)$$

With β fixed, the latent variable z_{ji} for the ith data point in group j is sampled according to

$$p(z_{ji} = k \,|\, z^{\backslash ji}, \beta) = \begin{cases} \left(n_{j\cdot k}^{\backslash ji} + \alpha_0 \beta_k\right) f_k^{\backslash ji}(x_{ji}) & \text{if component } k \text{ previously used} \\ \alpha_0 \beta_u f_{k_{\text{new}}}^{\backslash ji}(x_{ji}) & \text{if } k = k_{\text{new}}. \end{cases}$$

Finally, the variable m_{jk} is updated according to the conditional distribution

$$p(m_{jk} = m \,|\, z, \beta) = \frac{\Gamma(\alpha_0 \beta_k)}{\Gamma(\alpha_0 \beta_k + n_{j\cdot k})} s(n_{j\cdot k}, m) (\alpha_0 \beta_k)^m$$

where $s(n, m)$ are unsigned Stirling numbers of the first kind, which count the number permutations of n elements having m disjoint cycles.

8. Exercises

1. Let $w_1, w_2, \ldots$ be the weights generated from the stick-breaking process. Show that $\sum_{j=1}^{\infty} w_j = 1$ with probability 1.
2. Let $F \sim \mathrm{DP}(\alpha, F_0)$. Show that $\mathbb{E}(F) = F_0$. Show that the prior gets more concentrated around F_0 as $\alpha \to \infty$.
3. Find a bound on
$$\mathbb{P}(\sup_x |\overline{F}_n(x) - F(x)| > \epsilon)$$
where $\overline{F}_n$ is defined by (9).
4. Consider the Dirichlet process $\mathrm{DP}(\alpha, F_0)$.
(a) Set $F_0 = N(0,1)$. Draw 100 random distributions from the prior and plot them. Ty several diferent values of α.
(b) Draw $X_1, \ldots, X_n \sim F$ where $F = N(5,3)$. Compute and plot the empirical distribution function and plot a 95 percent confidence band. Now compute the Bayesian posterior using a $\mathrm{DP}(\alpha, F_0)$ prior with $F_0 = N(0,1)$. Note that, to make this realistic, we are assuming that the prior guess F_0 is not equal to the true (but unknown) F. Plot the Bayes estimator $\overline{F}_n$. (Try a few different values of α.) Compute a 95 percent Bayesian confidence band. Repeat the entire process many times and see how often the Bayesian confidence bands actually contains F.
5. In the hierarchical Dirichlet process, we first draw $G_0 \sim \mathrm{DP}(\gamma, H)$. The stick breaking representation allows us to write this as
$$G_0 = \sum_{i=1}^{\infty} \beta_i \delta_{\theta_i}.$$
The next level in the hierarchy samples $G_j \sim \mathrm{DP}(\alpha, G_0)$.
   (a) Show that, under the stick breaking representation for G_j,
$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}$$
   where $\pi_j = (\pi_{j1}, \pi_{j2}, \ldots) \sim \mathrm{DP}(\alpha_0, \beta)$.
   (b) Show that π_j can equivalently be constructed as
$$\begin{aligned} V_{jk} &\sim \text{Beta}\left(\alpha_0 \beta_k, \alpha_0\left(1 - \sum_{k=1}^{k} \beta_i\right)\right) \\ \pi_{jk} &= V_{jk} \prod_{i=1}^{k-1} (1 - V_{ji}) \end{aligned}$$

References

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Blei, D. M. and Jordan, M. I. (2005). Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144.

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

MacKay, D. J. C. and Peto, L. C. (1994). A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Society*, 101(476):1566–1581.