Nonparametric Bayesian Methods
10716, Spring 2020

Pradeep Ravikumar (amending notes by Larry Wasserman)

1 What is Nonparametric Bayes?

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

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

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

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

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:

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

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 $\pi(\theta|Y)$ and the plotting the samples.

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

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

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 $\mathcal{F}$. If there is no dominating measure for $\mathcal{F}$ then the posterior still exists but cannot be obtained by simply applying Bayes' theorem. 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 $\pi_n(A) = 1-\alpha$.

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. The inferences provided by Bayesian methods do not necessarily coincide with frequentist methods and they do not necessarily have properties like consistency, optimal rates of convergence, or coverage guarantees.

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 in some space of distributions $\mathcal{F}$. We will put a prior π on the set of all distributions in $\mathcal{F}$. 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 we 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 $(X_1, \ldots, X_n)$,

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

where

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

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 Three Nonparametric Problems

We will focus on three 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 smoother	Dirichlet process mixture
Estimating a regression function	kernel smoother	Gaussian process

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

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

Recall 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

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

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 $\mathrm{DP}(\alpha, F_0)$. The parameter F_0 is a distribution function, and the number α controls how tightly concentrated the prior is around F_0. You can then think of F as some "noisy" draw around F_0, similar to how $Z \sim N(\theta, I)$ is a noisy sample around some fixed θ, except that here we have the infinite-dimensional analogue where we draw distributions. Before defining the Dirichlet Process, let us first build some further intuition, and which would also shed light on why it 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 $\mathbb{R}$. Suppose that for a random distribution F, we let $F(A_j)$ be the amount of mass that F puts on the set A_j. Consider the requirement that for any such partition, the random distribution F should satisfy the condition that $(F(A_1), \ldots, F(A_k))$ have a Dirichlet distribution with parameters $(\alpha, F_0(A_1), \ldots, F_0(A_k))$. This property precisely characterizes a random draw from the Dirichlet process $\mathrm{DP}(\alpha, F_0)$.

Formal Specification of the Prior. To draw a single random distribution F from $\mathrm{Dir}(\alpha, F_0)$ we do the following steps:

1. Draw $s_1, s_2, \ldots$ independently from F_0.
2. Draw $V_1, V_2, \ldots \sim \mathrm{Beta}(1, \alpha)$. (Note that $V_i \in [0, 1]$.)
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

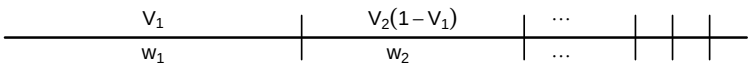

Figure 1: The stick breaking process shows how the weights $w_1, w_2, \ldots$ from the Dirichlet process are constructed. First we draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$. Then we set $w_1 = V_1,\ w_2 = V_2(1 - V_1),\ w_3 = V_3(1 - V_1)(1 - V_2), \ldots$.

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 $\mathbb{E}(F) = F_0$.

Remark. In practice, we can consider a finite truncation of the above, and draw a random draw from the Dirichlet process as:

1. Draw $s_1, \ldots, s_N$ independently from F_0.
2. Draw $V_1, \ldots, v_{N-1} \sim \text{Beta}(1, \alpha)$. (Note that $V_i \in [0, 1]$.)
3. Let $w_1 = V_1$ and $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$ for $j = 2, \ldots, N - 1$, and $w_N = 1 - \sum_{j=1}^{N-1} w_j$.
4. Let F be the discrete distribution that puts mass w_j at s_j, that is, $F = \sum_{j=1}^{N} w_j \delta_{s_j}$ where δ_{s_j} is a point mass at s_j.

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:

$$\begin{aligned} F &\sim \text{DP}(\alpha, F_0) \\ X_1, \ldots, X_n | F &\sim F \end{aligned}$$

where $\pi =$.

But there is an alternative method, called the Chinese Restaurant Process or infinite Pólya urn (Blackwell 1973). The algorithm is as follows.

1. Draw $X_1 \sim F_0$.

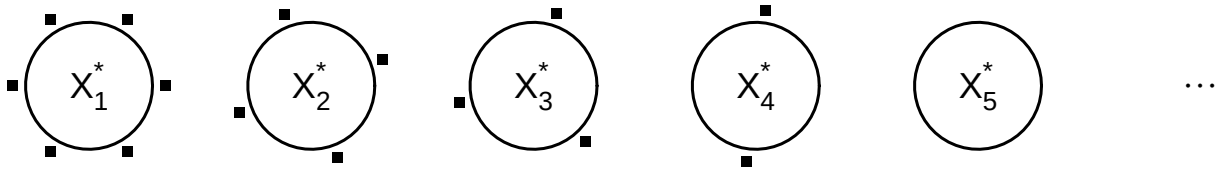

Figure 2: The Chinese restaurant process. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.

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.

How to Sample From the Posterior. Now suppose that $X_1, \ldots, X_n \sim F$ and that we place a $\text{Dir}(\alpha, F_0)$ prior on F.

Theorem 1 *Let $X_1, \ldots, X_n \sim F$ and let F have prior $\pi = \text{Dir}(\alpha, F_0)$. Then the posterior π for F given $X_1, \ldots, X_n$ is* $\text{Dir}\left(\alpha + n, \overline{F}_n\right)$ *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 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.

Specifically, recall that $F \sim \mathrm{DP}(\alpha, F_0)$ has the form $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$, which could be thought of an infinite mixture model with point mass distributions δ_{θ_j}, for which as we noted above a density does not exist. A natural extension is to instead consider an infinite mixture model with smoother components than point mass distributions:

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

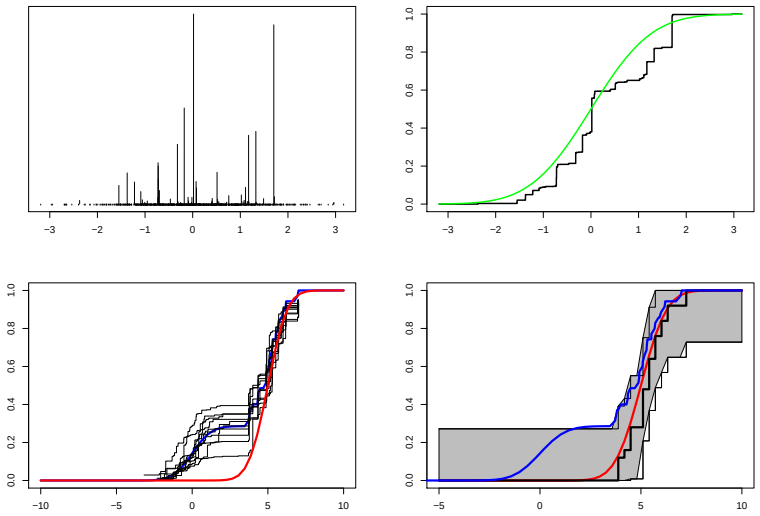

Figure 3: The top left plot shows the discrete probabilty function resulting from a single draw from the prior which is a DP(α, F_0) with $\alpha = 10$ and $F_0 = N(0,1)$. The top right plot shows the resulting cdf along with F_0. The bottom left plot shows a few draws from the posterior based on $n = 25$ 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 percnt frequentist confidence band.

where $g(x;\theta)$ is some smooth density such as that of the Normal distribution, for which $\theta_j = (\mu_j, \sigma_j)$. Here we wish to draw a random density f, which we could do so by largely following the same strategy as that of the Dirichlet process.

Formal Specification of the Dirichlet Process Mixture Prior. Draw $\theta_1, \theta_2, \ldots, F_0$, and draw $w_1, w_2, \ldots,$ from the stsick breaking process. Set $f(x) = \sum_{j=1}^{\infty} w_j g(x;\theta_j)$. 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.

This infinite mixture model is known as the Dirichlet process mixture model. As discussed above, this infinite "Dirichlet Process mixture" is similar to the random distribution $F \sim \mathrm{DP}(\alpha, F_0)$ drawn from a Dirichlet Process 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 $f(x|\theta_j)$.

Let us now briefly review the frequentist approach to density estimation.

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.

The kernel estimator can be thought of as a mixture with n components. A more general finite mixture model would take the form

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

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. The Dirichlet process mixture model could thus be viewed as infinite analogue of a Bayesian mixture model with a specific prior on the infinite set of mixture weights $\{w_j\}$, and the mixture component weights $\{\theta_j\}$.

How to Sample From the Prior Marginal. Given the construction of the Dirichlet Process Mixture, we can draw samples via:

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

(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*

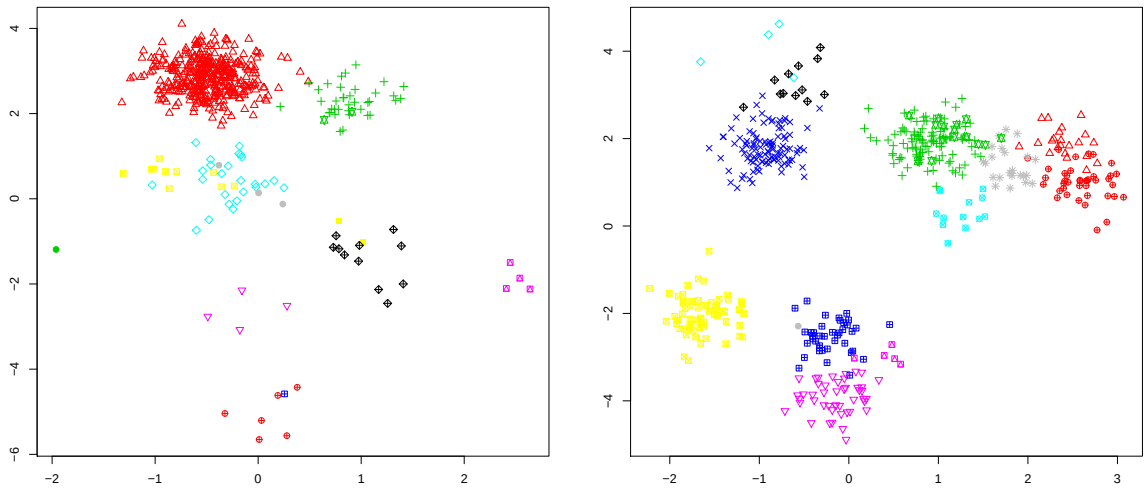

Figure 4: Samples from a Dirichlet process mixture model with Gaussian generator, $n = 500$.

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.

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

$$\frac{\alpha}{\alpha + n - 1} F_0(\cdot) + \frac{1}{\alpha + n - 1} \sum_{i=1}^{n-1} \delta_{\theta_i}(\cdot). \tag{13}$$

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

$$\theta_n = \begin{cases} \theta_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1}. \end{cases} \tag{14}$$

How to Sample From the Posterior. Unlike the Dirichlet Process case, here the posterior does not have a simple form, in part due to the presence of more general densities $g(x;\theta)$ rather than simple point mass distributions δ_θ. It is thus common to sample from the posterior by Gibbs sampling.

We briefly discuss an example using a mixture of Normals, via the approach of Ishwaran et al. (2002). The first step (in this particular approach) is to replace the infinite mixture with a large but finite mixture as discussed earlier. Thus we replace the stick-breaking process with $V_1, \ldots, V_{N-1} \sim \text{Beta}(1, \alpha)$ and $w_1 = V_1, w_2 = V_2(1-V_1), \ldots$. This generates $w_1, \ldots, w_N$ which sum to 1. Replacing the infinite mixture with the finite mixture is a numerical trick not an inferential step and has little numerical effect as long as N is large. For example, they show tha when $n = 1,000$ it suffices to use $N = 50$. A full specification of the resulting model, including priors on the hyperparameters is:

$$\begin{aligned}
\theta &\sim N(0, A) \\
\alpha &\sim \text{Gamma}(\eta_1, \eta_2) \\
\mu_1, \ldots, \mu_N &\sim N(\theta, B^2) \\
\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_N^2} &\sim \text{Gamma}(\nu_1, \nu_2) \\
K_1, \ldots, K_n &\sim \sum_{j=1}^{N} w_j \delta_j \\
X_i &\sim N(\mu_{K_i}, \sigma_{K_i}^2) \quad i = 1, \ldots, n
\end{aligned}$$

The hyperparemeters $A, B, \gamma_1, \gamma_2, \nu_1, \nu_2$ still need to be set. Compare this to kernel density estimation whihc requires the single bandwidth h. Ishwaran et al use $A = 1000$, $\nu_1 = \nu_2 = \eta_1 = \eta_2 = 2$ and they take B to be 4 ties the standard deviation of the data. It is now possible to wite down a Gibbs sampling algorithm for sampling from the posterior.

6 Nonparametric Regression

Consider the nonparametric regression model

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{15}$$

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

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} Y_i \; K\left(\frac{||x - X_i||}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{||x - X_i||}{h}\right)} \tag{16}$$

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:

$$(m(x_1), m(x_2), \ldots, m(x_n)) \sim N(\mu(x), K(x)) \tag{17}$$

where $K_{ij}(x) = K(x_i, x_j)$ is a Mercer kernel.

Let's assume that $\mu = 0$. 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

$$\pi(m) = (2\pi)^{-n/2}|K|^{-1/2} \exp\left(-\frac{1}{2} m^T K^{-1} m\right) \tag{18}$$

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

$$\frac{1}{\lambda} = v^T K^{-1} v \tag{19}$$

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

In this Bayesian setup, MAP estimation corresponds to Mercer kernel regression, which regularizes the squared error by the RKHS norm $\|\alpha\|_K^2$. The posterior mean, which due to Gaussianity, is also the posterior mode, and hence maximizes the log-posterior above, can be easily seen to be

$$\widehat{m} = \mathbb{E}(m|Y) = K\left(K + \sigma^2 I\right)^{-1} Y. \tag{20}$$

We see that $\widehat{m}$ is nothing but a linear smoother.

Comparison to Kernel Regression. We can see that this is in fact, very similar to the frequentist kernel smoother, which can be written as $\widehat{m} = KD^{-1}Y$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^n K_{ij}$.

Unlike kernel regression, where we just need to choose a bandwidth h, here we need to choose the function $K(x, y)$. This is a delicate matter.

Comparison to RKHS regression. It is also instructive to compare Gaussian processes and RKHS regression. To do so, we first consider the spectral representation of the kernel function:

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y),$$

so that $\{\lambda_i\}$ are its eigenvalues, and $\{\psi_i\}$ are the corresponding eigenfunctions. The Gaussian process can then be written as an infinite Bayesian linear regression model:

$$m(x) = \sum_{i=1}^{\infty} \theta_i \psi_i(x),$$

where $\theta_i \sim N(0, \lambda_i)$. This can be seen by noting that for this Bayesian linear regression model: $m(x)$ is Gaussian, with mean 0, and variance

$$\sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(x) = k(x,x).$$

Also $(m(x), m(y))$ is jointly Gaussian with covariance

$$\sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y) = k(x,y),$$

so that $m(x)$ is precisely the Gaussian process with kernel k.

On the other hand, as we have seen in earlier lecture, RKHS regression with the kernel k again estimates regression functions of the form $m(x) = \sum_{i=1}^{\infty} \theta_i \psi_i(x)$, with bounded RKHS norm

$$\|m\|_k = \sum_{i=1}^{\infty} \theta_i^2/\lambda_i.$$

However for a Gaussian process, its expected RKHS norm will be infinite since

$$\mathbb{E}\|m\|_k = \sum_{i=1}^{\infty} \mathbb{E}(\theta_i^2)/\lambda_i = \infty,$$

since $\mathbb{E}(\theta_i^2) = \lambda_i$ for a Gaussian process. Thus regression functions drawn from a Gaussian process are likely to be much less smooth compared to RKHS regression functions.

Predictive Distribution. Now, to compute the predictive distribution, given a Gaussian process prior, for a new point $Y_{n+1} = m(x_{n+1}) + \epsilon_{n+1}$, we note that $(Y_1, \ldots, Y_n) \sim N(0, K + \sigma^2 I)$. Let k be the vector

$$k = (K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1})) \tag{21}$$

Then $(Y_1, \ldots, Y_{n+1})$ is jointly Gaussian with covariance

$$\begin{pmatrix} K + \sigma^2 I & k \\ k^T & k(x_{n+1}, x_{n+1}) + \sigma^2 \end{pmatrix} \tag{22}$$

Therefore, conditional distribution of Y_{n+1} is

$$Y_{n+1}|Y_{1:n}, x_{1:n} \sim N\left(k^T(K + \sigma^2 I)^{-1}Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(K + \sigma^2 I)^{-1}k\right) \tag{23}$$

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-valdiation and its theoretical properties are simple and well-understood.

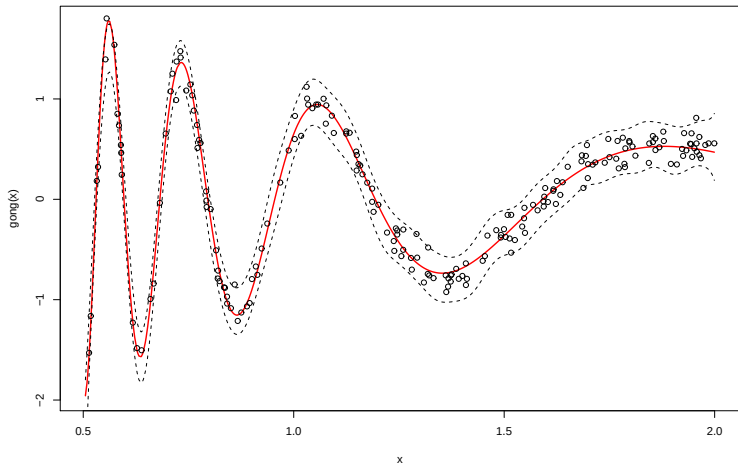

Figure 5: Mean of a Gaussian process

7 Theoretical Properties of Nonparametric Bayes

In this section we briefly discuss some theoretical properties of nonparametric Bayesian methods. We will focus on density estimation. In frequentist nonparametric inference, procedures are required to have certain guarantees such as consistency and minimaxity. Similar reasoning can be applied to Bayesian procedures. It is desirable, for example, that the posterior distribution π_n has mass that is concentrated near the true density function f.

We will be focusing on the property of consistency.

Let f_0 denote the true density. By consistency we mean that, when $f_0 \in A$, $\pi_n(A)$ should converge, in some sense, to 1. According to Doob's theorem, consistency holds under very weak conditions.

To state Doob's theorem we need some notation. The prior π and the model define a joint distribution μ_n on sequences $Y^n = (Y_1, \ldots, Y_n)$, namely, for any $B \in \mathbb{R}^n$,[1]

$$\mu_n(Y_n \in B) = \int \mathbb{P}(Y^n \in B|f)d\pi(f) = \int_B f(y_1)\cdots f(y_n)d\pi(f). \tag{24}$$

In fact, the model and prior determine a joint distribution μ on the set of infinite sequences[2] $\mathcal{Y}^\infty = \{Y^\infty = (y_1, y_2, \ldots,)\}$.

Theorem 3 (Doob 1949) *For every measurable A,*

$$\mu\left(\lim_{n\to\infty} \pi_n(A) = I(f_0 \in A)\right) = 1. \tag{25}$$

[1]More precisely, for any Borel set B.

[2]More precisely, on an appropriate σ-field over the set of infinite sequences.

By Doob's theorem, consistency holds except on a set of probability zero. This sounds good but it isn't; consider the following example.

Example 4 *Let $Y_1, \ldots, Y_n \sim N(\theta, 1)$. Let the prior π be a point mass at $\theta = 0$. Then the posterior is point mass at $\theta = 0$. This posterior is inconsistent on the set $N = \mathbb{R} - \{0\}$. This set has probability 0 under the prior so this does not contradict Doob's theorem. But clearly the posterior is useless.*

Doob's theorem is useless for our purposes because it is solopsistic. The result is with respect to the Bayesian's own distribution μ. Instead, we want to say that the posterior is consistent with respect to $\mathbb{P}_0$, the distribution generating the data.

To continue, let us define three types of neighborhoods. Let f be a density and let P_f be the corresponding probability measure. A Kullback-Leibler neighborhood around P_f is

$$B_K(p, \epsilon) = \left\{ P_g : \ \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx \leq \epsilon \right\}. \tag{26}$$

A Hellinger neighborhood around P_f is

$$B_H(p, \epsilon) = \left\{ P_g : \ \int (\sqrt{f}(x) - \sqrt{g}(x))^2 \leq \epsilon^2 \right\}. \tag{27}$$

A weak neighborhood around P_f is

$$B_W(P, \epsilon) = \left\{ Q : \ d_W(P, Q) \leq \epsilon \right\} \tag{28}$$

where d_W is the Prohorov metric

$$d_W(P, Q) = \inf\left\{ \epsilon > 0 : \ P(B) \leq Q(B^\epsilon) + \epsilon, \text{ for all } B \right\} \tag{29}$$

where $B^\epsilon = \{x : \ \inf_{y \in B} \|x - y\| \leq \epsilon\}$. Weak neighborhoods are indeed very weak: if $P_g \in B_W(P_f, \epsilon)$ it does not imply that g resembles f.

Theorem 5 (Schwartz 1963) *If*

$$\pi(B_K(f_0, \epsilon)) > 0, \quad \text{for all } \epsilon > 0 \tag{30}$$

then, for any $\delta > 0$,

$$\pi_n(B_W(P_0, \delta)) \stackrel{a.s.}{\to} 1 \tag{31}$$

with respect to P_0.

This is still unsatisfactory since weak neighborhoods are large. Let $N(\mathcal{M}, \epsilon)$ denote the smallest number of functions $f_1, \ldots, f_N$ such that, for each $f \in \mathcal{M}$, there is a f_j such that $f(x) \leq f_j(x)$ for all x and such that $\sup_x(f_j(x) - f(x)) \leq \epsilon$. Let $H(\mathcal{M}, \epsilon) = \log N(\mathcal{M}, \epsilon)$.

Theorem 6 (Barron, Schervish and Wasserman (1999) and Ghosal, Ghosh and Ramamoorth
Suppose that

$$\pi(B_K(f_0, \epsilon)) > 0, \quad \text{for all } \epsilon > 0. \tag{32}$$

Further, suppose there exists $\mathcal{M}_1, \mathcal{M}_2, \ldots$ such that $\pi(\mathcal{M}_j^c) \leq c_1 e^{-jc_2}$ and $H(\mathcal{M}_j, \delta) \leq c_3 j$ for all large j. Then, for any $\delta > 0$,

$$\pi_n(B_H(P_0, \delta)) \stackrel{a.s.}{\to} 1 \tag{33}$$

with respect to P_0.