

# Nonparametric Bayesian Methods

## 1 What is Nonparametric Bayes?

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

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

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

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

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:

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

We use the posterior to compute a point estimator such as the posterior mean of  $\theta$ . We can also summarize the posterior by drawing a large sample  $\theta_1, \dots, \theta_N$  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

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

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

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

where  $A \subset \mathcal{F}$ ,  $\mathcal{L}_n(f) = \prod_i f(Y_i)$  is the likelihood function and  $\pi$  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

$$\hat{f}(y) = \int f(y)d\pi_n(f). \quad (4)$$

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

The Bayesian model can be written as

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

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

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

where

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

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

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

$$\pi_n(A) \equiv \pi(F \in A | X_1, \dots, X_n). \quad (5)$$

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

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

### 3 Three Nonparametric Problems

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

<b>Statistical Problem</b>	<b>Frequentist Approach</b>	<b>Bayesian Approach</b>
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  $X_1, \dots, X_n$  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

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

Recall that for every  $\epsilon > 0$  and every  $F$ ,

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

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

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

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

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

The distribution  $\pi$  has two parameters,  $F_0$  and  $\alpha$  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  $\alpha$  controls how tightly concentrated the prior is around  $F_0$ . The model may be summarized as:

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

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  $s_1, s_2, \dots$  independently from  $F_0$ .
2. Draw  $V_1, V_2, \dots \sim \text{Beta}(1, \alpha)$ .
3. Let  $w_1 = V_1$  and  $w_j = V_j \prod_{i=1}^{j-1} (1 - V_i)$  for  $j = 2, 3, \dots$
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  $\delta_{s_j}$  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  $w_1, w_2, \dots$  is often called the stick breaking process. Imagine we have a stick of unit length. Then  $w_1$  is obtained by breaking the stick at 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  $w_1, w_2, \dots$ . 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. The reason is this. Recall that a random vector  $P = (P_1, \dots, P_k)$  has a Dirichlet distribution with parameters  $(\alpha, g_1, \dots, g_k)$  (with  $\sum_j g_j = 1$ ) if the distribution of  $P$  has density

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

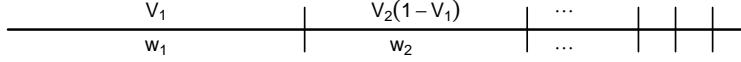


Figure 1: The stick breaking process shows how the weights  $w_1, w_2, \dots$  from the Dirichlet process are constructed. First we draw  $V_1, V_2, \dots \sim \text{Beta}(1, \alpha)$ . Then we set  $w_1 = V_1$ ,  $w_2 = V_2(1 - V_1)$ ,  $w_3 = V_3(1 - V_1)(1 - V_2), \dots$

over the simplex  $\{p = (p_1, \dots, p_k) : p_j \geq 0, \sum_j p_j = 1\}$ . Let  $(A_1, \dots, A_k)$  be any partition of  $\mathbb{R}$  and let  $F \sim \text{DP}(\alpha, F_0)$  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  $(F(A_1), \dots, F(A_k))$  has a Dirichlet distribution with parameters  $(\alpha, F_0(A_1), \dots, F_0(A_k))$ . In fact, this property characterizes the Dirichlet process.

*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  $X_1, \dots, X_n$  from  $F$ . But there is an alternative method, called the Chinese Restaurant Process or infinite Pólya urn (Blackwell 1973). The algorithm is as follows.

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

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

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

The sample  $X_1, \dots, X_n$  is likely to have ties since  $F$  is discrete. Let  $X_1^*, X_2^*, \dots$  denote the unique values of  $X_1, \dots, X_n$ . Define cluster assignment variables  $c_1, \dots, c_n$  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  $n$ th 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  $j$ th 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  $\{1, \dots, n\}$ , which corresponds to a clustering of the indices. See Figure 2.

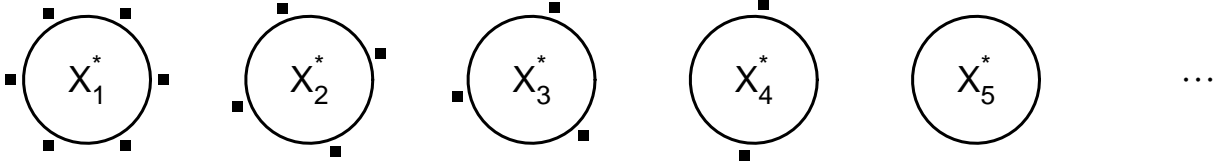


Figure 2: The Chinese restaurant process. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.

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

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

$$\bar{F}_n = \frac{n}{n + \alpha} F_n + \frac{\alpha}{n + \alpha} F_0. \quad (9)$$

Since the posterior is again a Dirichlet process, we can sample from it as we did the prior but we replace  $\alpha$  with  $\alpha + n$  and we replace  $F_0$  with  $\bar{F}_n$ . Thus the posterior mean is  $\bar{F}_n$  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  $\bar{F}_n$ .

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

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

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,  $\bar{F}_n \approx F_n$  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 specifying  $\alpha$  or  $F_0$ .

**Example 2** Figure 3 shows a simple example. The prior is  $\text{DP}(\alpha, F_0)$  with  $\alpha = 10$  and  $F_0 = N(0, 1)$ . The top left plot shows the discrete probability function resulting from a single

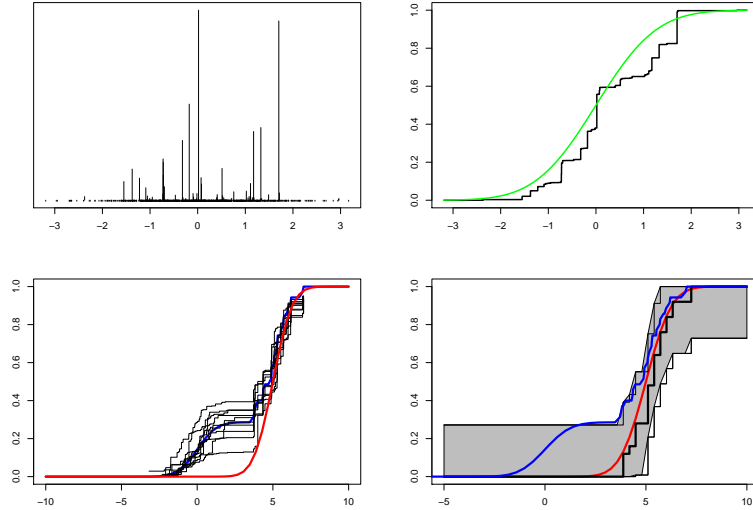


Figure 3: The top left plot shows the discrete probability function resulting from a single draw from the prior which is a  $DP(\alpha, F_0)$  with  $\alpha = 10$  and  $F_0 = N(0, 1)$ . The top right plot shows the resulting cdf along with  $F_0$ . The bottom left plot shows a few draws from the posterior 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 posterior mean (blue) and a 95 percent frequentist confidence band.

*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 posterior mean (blue) and a 95 percent frequentist confidence band.*

## 5 Density Estimation

Let  $X_1, \dots, X_n \sim F$  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.

The most common frequentist estimator is the kernel estimator

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

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, if  $f(x; \theta)$  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  $\theta_1, \dots, \theta_k$ , on  $w_1, \dots, w_k$  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  $\theta_1, \theta_2, \dots$  to be drawn from some  $F_0$  and we could take  $w_1, w_2, \dots$ , 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. This infinite mixture is the same as the random distribution  $F \sim \text{DP}(\alpha, F_0)$  which had the form  $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$  except that the point mass distributions  $\delta_{\theta_j}$  are replaced by smooth densities  $f(x|\theta_j)$ .

The model may be re-expressed as:

$$F \sim \text{DP}(\alpha, F_0) \tag{10}$$

$$\theta_1, \dots, \theta_n | F \sim F \tag{11}$$

$$X_i | \theta_i \sim f(x|\theta_i), \quad i = 1, \dots, n. \tag{12}$$

(In practice,  $F_0$  itself has free parameters which also require priors.) Note that in the DPM, *the parameters  $\theta_i$  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  $\theta_i$ 's. (Recall our earlier discussion of the Chinese Restaurant Process.) The  $k < n$  distinct values of  $\theta_i$  can be thought of as defining clusters. The beauty of this model is that the discreteness of  $F$  automatically creates a clustering of the  $\theta_j$ 's. In other words, we have implicitly created a prior on  $k$ , the number of distinct  $\theta_j$ 's.

*How to Sample From the Prior.* Draw  $\theta_1, \theta_2, \dots, F_0$  and draw  $w_1, w_2, \dots$ , from the stick 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.



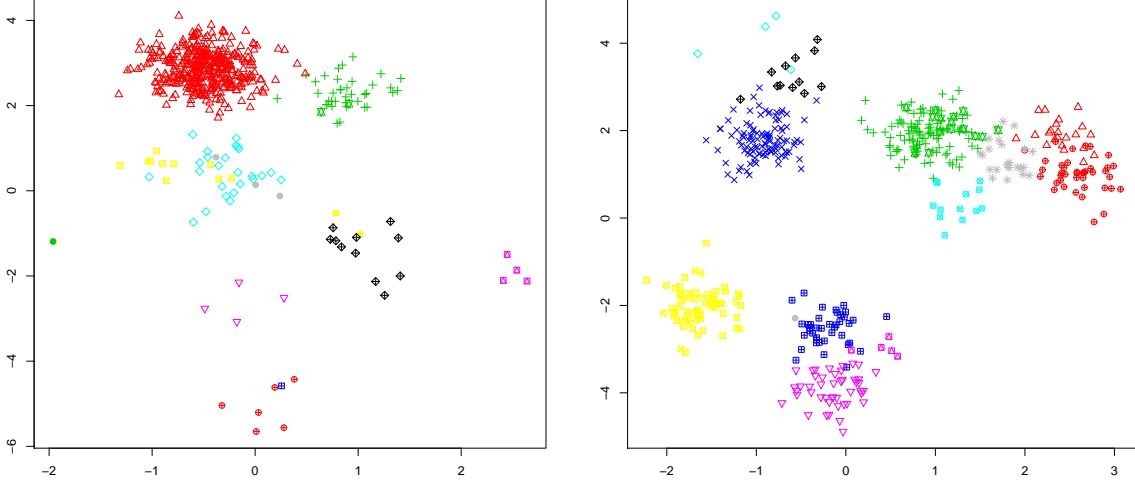


Figure 4: Samples from a Dirichlet process mixture model with Gaussian generator,  $n = 500$ .

*How to Sample From the Prior Marginal.* The prior marginal  $m$  is

$$m(x_1, x_2, \dots, x_n) = \int \prod_{i=1}^n f(x_i|F) d\pi(F) \quad (13)$$

$$= \int \prod_{i=1}^n \left( \int f(x_i|\theta) p(\theta|F) dF(\theta) \right) dP(G) \quad (14)$$

If we want to draw a sample from  $m$ , we first draw  $F$  from a Dirichlet process with parameters  $\alpha$  and  $F_0$ , and then generate  $\theta_i$  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  $\theta_j$ 's sequentially. Given  $\theta_1, \dots, \theta_{i-1}$  we draw  $\theta_j$  from

$$\alpha F_0(\cdot) + \sum_{i=1}^{n-1} \delta_{\theta_i}(\cdot). \quad (15)$$

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

$$\theta_n = \begin{cases} \theta_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1}. \end{cases} \quad (16)$$

*How to Sample From the Posterior.* We sample from the posterior by Gibbs sampling; we may discuss that later.

To understand better how to use the model, we consider how to use the DPM for estimating density using a mixture of Normals. There are numerous implementations. We consider one due to Ishwaran et al. (2002). The first step (in this particular approach) is to replace the infinite mixture with a large but finite mixture. Thus we replace the stick-breaking process with  $V_1, \dots, V_{N-1} \sim \text{Beta}(1, \alpha)$  and  $w_1 = V_1, w_2 = V_2(1 - V_1), \dots$ . This generates  $w_1, \dots, w_N$  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 that when  $n = 1,000$  it suffices to use  $N = 50$ . A full specification of the resulting model, including priors on the hyperparameters is:

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

The hyperparameters  $A, B, \gamma_1, \gamma_2, \nu_1, \nu_2$  still need to be set. Compare this to kernel density estimation which 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 times the standard deviation of the data. It is now possible to write down a Gibbs sampling algorithm for sampling from the posterior.

## 6 Nonparametric Regression

Consider the nonparametric regression model

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \dots, n \quad (17)$$

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

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\|x - X_i\|}{h}\right)}{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right)} \quad (18)$$

where  $K$  is a kernel and  $h$  is a bandwidth. The Bayesian version requires a prior  $\pi$  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  $x_1, \dots, x_n \in \mathcal{X}$  the vector  $(m(x_1), m(x_2), \dots, m(x_n))$  is Normally distributed:

$$(m(x_1), m(x_2), \dots, m(x_n)) \sim N(\mu(x), K(x)) \quad (19)$$

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

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

$$\pi(m) = (2\pi)^{-n/2} |K|^{-1/2} \exp\left(-\frac{1}{2} m^T K^{-1} m\right) \quad (20)$$

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

$$\pi(\alpha) = (2\pi)^{-n/2} |K|^{-1/2} \exp\left(-\frac{1}{2} \alpha^T K \alpha\right) \quad (21)$$

Under the additive 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

$$\log p(y|m) = -\frac{1}{2\sigma^2} \sum_i (y_i - m(x_i))^2 + \text{const} \quad (22)$$

and the log-posterior is

$$\log p(y|m) + \log \pi(m) = -\frac{1}{2\sigma^2} \|y - K\alpha\|_2^2 - \frac{1}{2} \alpha^T K \alpha + \text{const} \quad (23)$$

$$= -\frac{1}{2\sigma^2} \|y - K\alpha\|_2^2 - \frac{1}{2} \|\alpha\|_K^2 + \text{const} \quad (24)$$

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

$$\frac{1}{\lambda} = v^T K^{-1} v \quad (25)$$

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 is

$$\mathbb{E}(\alpha|Y) = (K + \sigma^2 I)^{-1} Y \quad (26)$$

and thus

$$\hat{m} = \mathbb{E}(m|Y) = K (K + \sigma^2 I)^{-1} Y. \quad (27)$$

We see that  $\hat{m}$  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 function  $K(x, y)$ . This is a delicate matter.

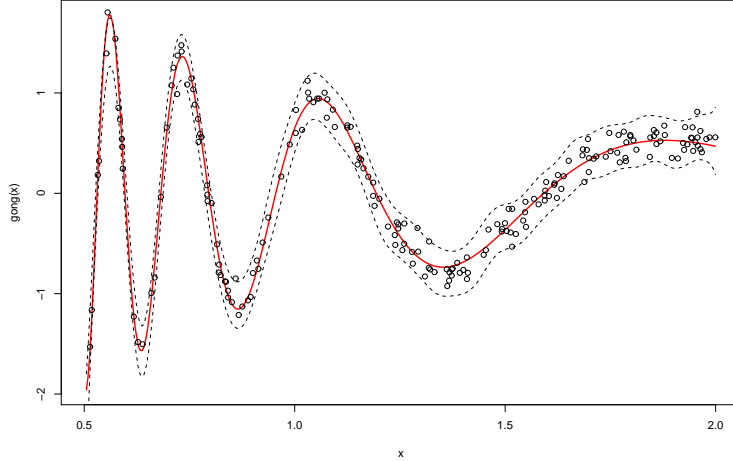


Figure 5: Mean of a Gaussian process

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

$$k = (K(x_1, x_{n+1}), \dots, K(x_n, x_{n+1})) \quad (28)$$

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

$$\begin{pmatrix} K + \sigma^2 I & k \\ k^T & k(x_{n+1}, x_{n+1}) + \sigma^2 \end{pmatrix} \quad (29)$$

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

$$Y_{n+1} | Y_{1:n}, x_{1:n} \sim N(k^T (K + \sigma^2 I)^{-1} Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T (K + \sigma^2 I)^{-1} k) \quad (30)$$

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.

## 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  $\pi_n$  has mass that is concentrated near the true density function  $f$ . More specifically, we can ask three specific questions:

1. Is the posterior consistent?
2. Does posterior concentrate at the optimal rate?
3. Does posterior have correct coverage?

## 7.1 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  $\pi$  and the model define a joint distribution  $\mu_n$  on sequences  $Y^n = (Y_1, \dots, Y_n)$ , namely, for any  $B \in \mathbb{R}^n$ ,<sup>1</sup>

$$\mu_n(Y^n \in B) = \int \mathbb{P}(Y^n \in B|f)d\pi(f) = \int_B f(y_1) \cdots f(y_n)d\pi(f). \quad (31)$$

In fact, the model and prior determine a joint distribution  $\mu$  on the set of infinite sequences<sup>2</sup>  $\mathcal{Y}^\infty = \{Y^\infty = (y_1, y_2, \dots)\}$ .

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

$$\mu \left( \lim_{n \rightarrow \infty} \pi_n(A) = I(f_0 \in A) \right) = 1. \quad (32)$$

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  $Y_1, \dots, Y_n \sim N(\theta, 1)$ . Let the prior  $\pi$  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 solopstic. The result is with respect to the Bayesian's own distribution  $\mu$ . Instead, we want to say that the posterior is consistent with respect to  $\mathbb{P}_0$ , the distribution generating the data.

---

<sup>1</sup>More precisely, for any Borel set  $B$ .

<sup>2</sup>More precisely, on an appropriate  $\sigma$ -field over the set of infinite sequences.

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

$$B_K(p, \epsilon) = \left\{ P_g : \int f(x) \log \left( \frac{f(x)}{g(x)} \right) dx \leq \epsilon \right\}. \quad (33)$$

A Hellinger neighborhood around  $P_f$  is

$$B_H(p, \epsilon) = \left\{ P_g : \int (\sqrt{f(x)} - \sqrt{g(x)})^2 \leq \epsilon^2 \right\}. \quad (34)$$

A weak neighborhood around  $P_f$  is

$$B_W(P, \epsilon) = \left\{ Q : d_W(P, Q) \leq \epsilon \right\} \quad (35)$$

where  $d_W$  is the Prohorov metric

$$d_W(P, Q) = \inf \left\{ \epsilon > 0 : P(B) \leq Q(B^\epsilon) + \epsilon, \text{ for all } B \right\} \quad (36)$$

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*

$$\pi(B_K(f_0, \epsilon)) > 0, \quad \text{for all } \epsilon > 0 \quad (37)$$

*then, for any  $\delta > 0$ ,*

$$\pi_n(B_W(P, \delta)) \xrightarrow{a.s.} 1 \quad (38)$$

*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  $f_1, \dots, f_N$  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 Ramamoorthi)** *Suppose that*

$$\pi(B_K(f_0, \epsilon)) > 0, \quad \text{for all } \epsilon > 0. \quad (39)$$

*Further, suppose there exists  $\mathcal{M}_1, \mathcal{M}_2, \dots$  such that  $\pi(\mathcal{M}_j^c) \leq c_1 e^{-j c_2}$  and  $H(\mathcal{M}_j, \delta) \leq c_3 j$  for all large  $j$ . Then, for any  $\delta > 0$ ,*

$$\pi_n(B_H(P, \delta)) \xrightarrow{a.s.} 1 \quad (40)$$

*with respect to  $P_0$ .*

**Example 7** Recall the Normal means model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \dots \quad (41)$$

where  $\epsilon_i \sim N(0, \sigma^2)$ . We want to infer  $\theta = (\theta_1, \theta_2, \dots)$ . Assume that  $\theta$  is contained in the Sobolev space

$$\theta \in \Theta = \left\{ \theta : \sum_i \theta_i^2 i^{2p} < \infty \right\}. \quad (42)$$

Recall that the estimator  $\hat{\theta}_i = b_i Y_i$  is minimax for this Sobolev space where  $b_i$  is an appropriate constant. In fact the Efromovich-Pinsker estimator is adaptive minimax over the smoothness index  $p$ . A simple Bayesian analysis is to use the prior  $\pi$  that treats each  $\theta_i$  as independent random variables and  $\theta_i \sim N(0, \tau_i^2)$  where  $\tau_i^2 = i^{-2q}$ . Have we really defined a prior on  $\Theta$ ? We need to make sure that  $\pi(\Theta) = 1$ . Fix  $K > 0$ . Then,

$$\pi\left(\sum_i \theta_i^2 i^{2p} > K\right) \leq \frac{\sum_i \mathbb{E}_\pi(\theta_i^2) i^{2p}}{K} = \frac{\sum_i \tau_i^2 i^{2p}}{K} = \frac{\sum_i \frac{1}{i^{2(q-p)}}}{K}. \quad (43)$$

The numerator is finite as long as  $q > p + (1/2)$ . Assuming  $q > p + (1/2)$  we then see that  $\pi(\sum_i \theta_i^2 i^{2p} > K) \rightarrow 0$  as  $K \rightarrow \infty$  which shows that  $\pi$  puts all its mass on  $\Theta$ . But, as we shall see below, the condition  $q > p + (1/2)$  is guaranteed to yield a posterior with a suboptimal rate of convergence.

## 7.2 Rates of Convergence

Here the situation is more complicated. Recall the Normal means model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \dots \quad (44)$$

where  $\epsilon_i \sim N(0, \sigma^2)$ . We want to infer  $\theta = (\theta_1, \theta_2, \dots) \in \Theta$  from  $Y = (Y_1, Y_2, \dots)$ . Assume that  $\theta$  is contained in the Sobolev space

$$\theta \in \Theta = \left\{ \theta : \sum_i \theta_i^2 i^{2p} < \infty \right\}. \quad (45)$$

The following results are from Zhao (2000), Shen and Wasserman (2001), and Ghosal, Ghosh and van der Vaart (2000).

**Theorem 8** Put independent Normal priors  $\theta_i \sim N(0, \tau_i^2)$  where  $\tau_i^2 = i^{-2q}$ . The Bayes estimator attains the optimal rate only when  $q = p + (1/2)$ . But then:

$$\pi(\Theta) = 0 \quad \text{and} \quad \pi(\Theta|Y) = 0. \quad (46)$$

### 7.3 Coverage

Suppose  $\pi_n(A) = 1 - \alpha$ . Does this imply that

$$\mathbb{P}_{f_0}^n(f_0 \in A) \geq 1 - \alpha? \quad (47)$$

or even

$$\liminf_{n \rightarrow \infty} \inf_{f_0} \mathbb{P}_{f_0}^n(f_0 \in A) \geq 1 - \alpha? \quad (48)$$

Recall what happens for parametric models: if  $A = (-\infty, a]$  and

$$\mathbb{P}(\theta \in A | \text{data}) = 1 - \alpha \quad (49)$$

then

$$\mathbb{P}_\theta(\theta \in A) = 1 - \alpha + O\left(\frac{1}{\sqrt{n}}\right) \quad (50)$$

and, moreover, if we use the Jeffreys' prior then

$$\mathbb{P}_\theta(\theta \in A) = 1 - \alpha + O\left(\frac{1}{n}\right). \quad (51)$$

Is the same true for nonparametric models? Unfortunately, no; counterexamples are given by Cox (1993) and Freedman (1999). In their examples, one has:

$$\pi_n(A) = 1 - \alpha \quad (52)$$

but

$$\liminf_{n \rightarrow \infty} \inf_{f_0} \mathbb{P}_{f_0}^n(f_0 \in A) = 0! \quad (53)$$