Nonparametric Bayesian Methods

김지수 (Jisu KIM)

인공지능을 위한 이론과 모델링, 2023 가을학기

The lecture note is a minor modification of the lecture notes from Prof. Larry Wasserman's "Statistical Machine Learning".

Bayesian Inference

The philosophical distinction between Bayes and frequentists is deep. The rest of the lecture will follow the frequentist framework, where, to us a probability is representing some type of long run frequency, i.e. when we say the probability that our estimator is close to some unknown "true" parameter with probability at least $1-\delta$ we are really imagining repeating this (or some other) experiment many many times and then our guarantees will be correct for at least $1-\delta$ of these experiments. Similarly, with confidence intervals, we imagine many people across the world construct confidence intervals and our guarantee is that 95% of those intervals would trap the true parameter, i.e. the goal of frequentist inference is to create procedures with long run guarantees.

Moreover, the guarantees should be uniform over θ if possible. For example, a confidence interval traps the true value of θ with probability $1-\alpha$, no matter what the true value of θ is. In frequentist inference, procedures are random while parameters are fixed, unknown quantities.

In the Bayesian approach, probability is regarded as a measure of subjective degree of belief. One can view the Bayesian approach as a way to manipulate beliefs. Beliefs are then assumed to follow the rules of normal probabilities by a notion called coherence. In this framework, everything, including parameters, is regarded as random. These procedures do not have to satisfy frequency guarantees.

A summary of the main ideas is in the table below.

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
θ	random variable	fixed
X	random variable	random variable

In frequentist inference the goal was: create procedures that have good frequency properties.

In Bayesian inference the goal is to write down a prior that captures your prior belief and compute the posterior; then you are essentially done.

Bayesian confidence sets

In frequentist inference, a $1-\alpha$ confidence set/interval is a random set C_α that captures the parameter θ with probability $1-\alpha$:

$$\mathbb{P}_{X_1,\ldots,X_n\sim P_\theta}(\theta\in C_\alpha)=1-\alpha.$$

In Bayesian inference, a $1-\alpha$ credible set/interval is a set C_α to which the posterior assigns $1-\alpha$ mass:

$$\mathbb{P}(\theta\in C_\alpha|X_1,\ldots,X_n)=1-\alpha.$$

Once again notice that the thing that is random is θ, the data is conditioned on (i.e. fixed). The set C_α is fixed (i.e. not random) here, unlike in a frequentist confidence interval. These intervals do not typically have frequency guarantees.

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:

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

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

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

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:

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

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 stull exists but cannot be obtained by simply applying Bayes' theorem. An estimate of f is the posterior mean

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

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.

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 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 $(X_1, \ldots, X_n)$,

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

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

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

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.

Three Nonparametric Problems

We will focus on two specific problems. The two 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 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

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

Then we will learn in the Concentration Inequality lecture, that for every $\epsilon > 0$ and every F,

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

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

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

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.

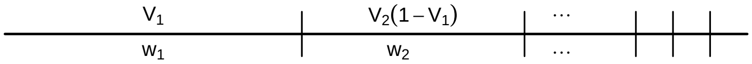

Figure 1: The stick breaking process shows how the weights $w_1, w_2, \ldots$ from the Dirichlet process are constructed. First we draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$. Then we set $w_1 = V_1,\ w_2 = V_2(1 - V_1),\ w_3 = V_3(1 - V_1)(1 - V_2), \ldots$.

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 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 $w_1, w_2, \ldots$. 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. The reason is this. 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 $\mathbb{R}$ 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 $(F(A_1), \ldots, F(A_k))$ has a Dirichlet distribution with parameters $(\alpha, F_0(A_1), \ldots, F_0(A_k))$. 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 $X_1, \ldots, X_n$ 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 $i = 2, \ldots, n$: 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 $X_1, \ldots X_{i-1}$.

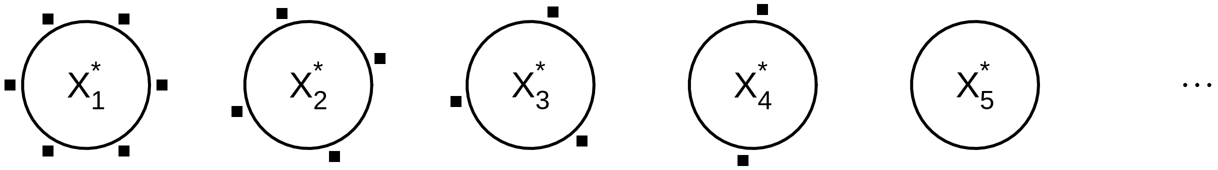

Figure 2: The Chinese restaurant process. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.

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. *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*

$$\overline{F}_n = \frac{n}{n+\alpha} F_n + \frac{\alpha}{n+\alpha} F_0. \tag{1}$$

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

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

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

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

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. 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 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.

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.

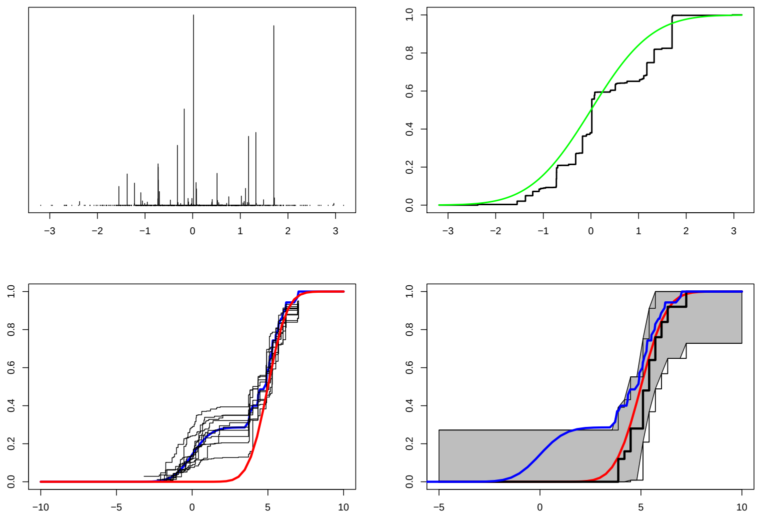

Figure 3: The top left plot shows the discrete probabilty function resulting from a single draw from the prior which is a DP(α, F_0) with $\alpha = 10$ and $F_0 = N(0, 1)$. The top right plot shows the resulting cdf along with F_0. The bottom left plot shows a few draws from the posterior based on $n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

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, 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. 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 δ_{θ_j} are replaced by smooth densities $f(x|\theta_j)$.

The model may be re-expressed as:

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

(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 erlier discussion of the Chinese Restaurant Process.) The $k < n$ distinct values of θ_i can be thought of as defining

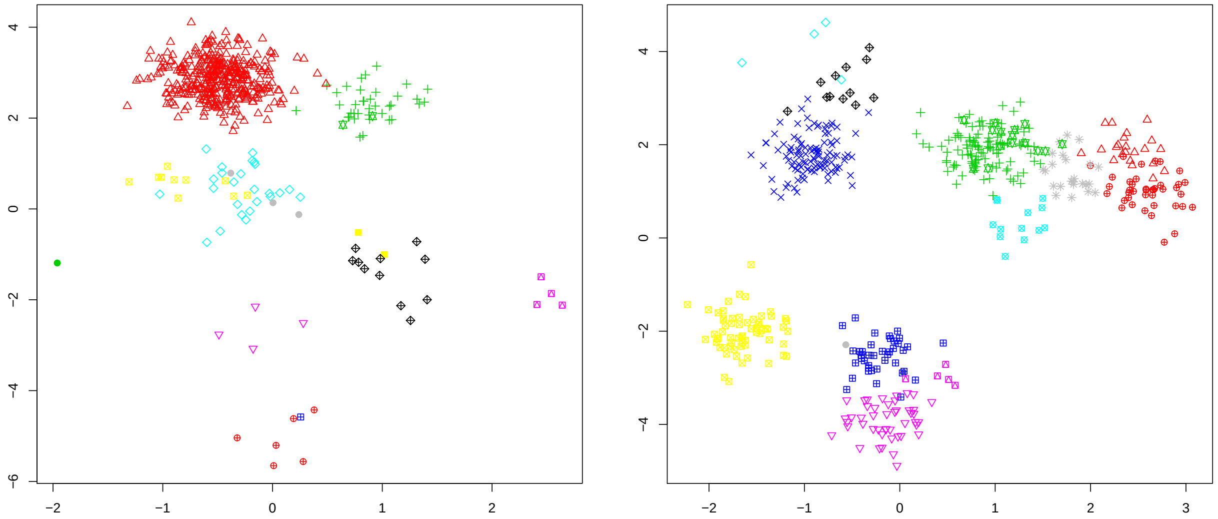

Figure 4: Samples from a Dirichlet process mixture model with Gaussian generator, $n = 500$.

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.

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.

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

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

If we want 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

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

How to Sample From the Posterior. We sample from the posterior by Gibbs sampling. We do not cover this in this class.

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. 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?

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]

$$\mu_n(Y_n \in B) = \int \mathbb{P}(Y^n \in B|f) d\pi(f) = \int_B f(y_1) \cdots f(y_n) d\pi(f). \tag{2}$$

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 (Doob 1949). *For every measurable A,*

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

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. 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.

Rates of Convergence

Example. Consider the Normal means model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \ldots,$$

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

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

Then 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 π that treats each θ_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 Θ? We need to make sure that $\pi(\Theta) = 1$. Fix $K > 0$. Then,

$$\pi\Big(\sum_i \theta_i^2 i^{2p} > K\Big) \le \frac{\sum_i \mathbb{E}_\pi(\theta_i^2) i^{2p}}{K} = \frac{\sum_i \tau_i^2 i^{2p}}{K} = \frac{\sum_i \frac{1}{i^{2(q-p)}}}{K}.$$

The numerator is finite as long as $q > p + (1/2)$. Assuming $q > p + (1/2)$ we then see that $\pi(\sum_i^2 i^{2p} > K) \to 0$ as $K \to \infty$ which shows that π puts all its mass on Θ.

But, as we see below, the condition $q > p + (1/2)$ is guaranteed to yield a posterior with a suboptimal rate of convergence. The following results are from Zhao (2000), Shen and Wasserman (2001), and Ghosal, Ghosh and van der Vaart (2000).

Theorem. *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:*

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

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

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

Coverage

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

$$\mathbb{P}^n_{f_0}(f_0 \in A) \geq 1 - \alpha?$$

or even

$$\liminf_{n\to\infty} \inf_{f_0} \mathbb{P}^n_{f_0}(f_0 \in A) \geq 1 - \alpha?$$

For parametric models: if $A = (-\infty, a]$ and

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

then

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

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

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

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

$$\pi_n(A) = 1 - \alpha,$$

but

$$\liminf_{n\to\infty} \inf_{f_0} \mathbb{P}_{f_0}(f_0 \in A) = 0.$$