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 $Y_1, \ldots, Y_n \sim f(y|\theta)$. 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

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

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:

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

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 stull 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. 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

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

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

How to Draw From 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 $V_1, V_2, \ldots \sim \mathrm{Beta}(1, \alpha)$.
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 $F \sim \mathrm{Dir}(\alpha, F_0)$ 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

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

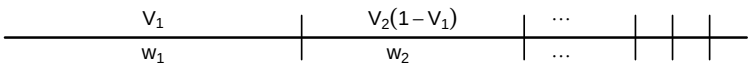


Figure 1: The stick breaking process shows how the weights $w_1, w_2, \ldots$ from the Dirichlet process are constructed. First we draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$. Then we set $w_1 = V_1,\ w_2 = V_2(1 - V_1),\ w_3 = V_3(1 - V_1)(1 - V_2), \ldots$.

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 $X_1, \ldots X_{i-1}$.

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.

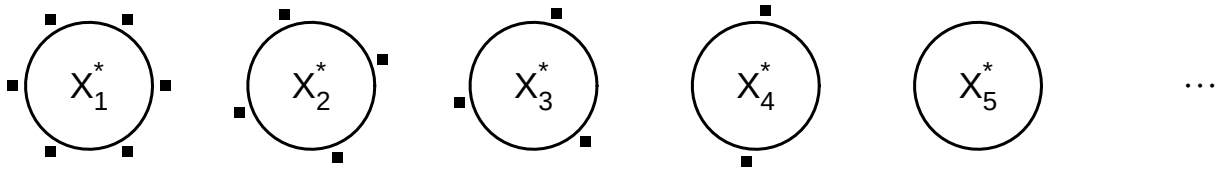


Figure 2: The Chinese restaurant process. A new person arrives and either sits at a table with people or sits at a new table. The probability of sitting at a table is proportional to the number of people at the table.

How to Sample From the Posterior. Now suppose that $X_1, \ldots, X_n \sim F$ 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* $\text{DP}(\alpha, F_0)$ *with $\alpha = 10$ and $F_0 = N(0, 1)$. The top left plot shows the discrete probabilty function resulting from a single*

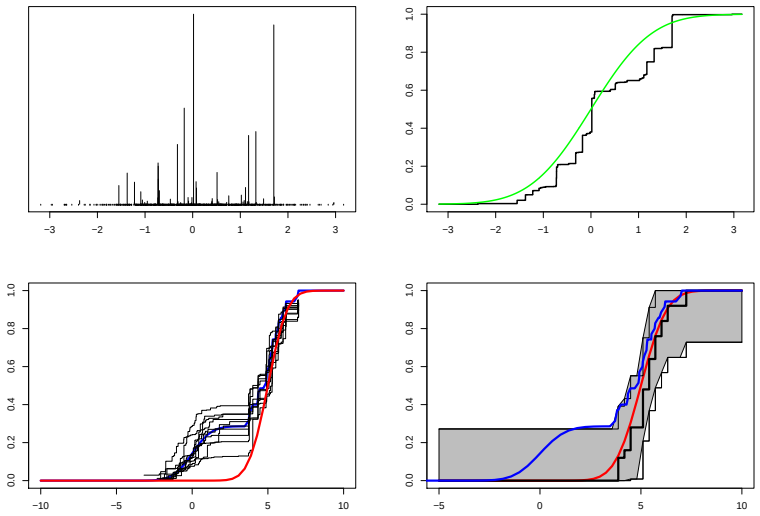


Figure 3: The top left plot shows the discrete probabilty function resulting from a single draw from the prior which is a $\mathrm{DP}(\alpha, F_0)$ with $\alpha = 10$ and $F_0 = N(0,1)$. The top right plot shows the resulting cdf along with F_0. The bottom left plot shows a few draws from the posterior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

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. 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 erlier 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.

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.

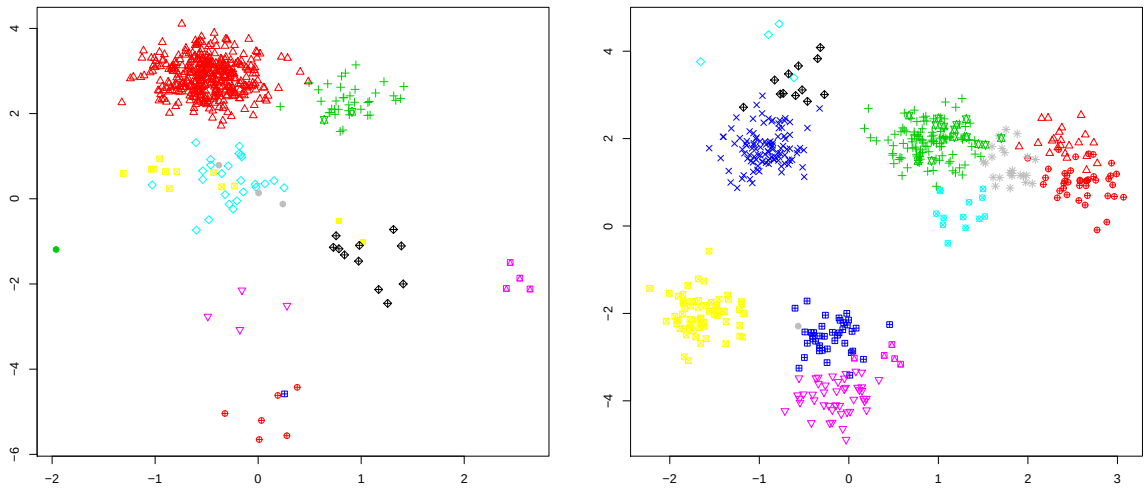


Figure 4: Samples from a Dirichlet process mixture model with Gaussian generator, $n = 500$.

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

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

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

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

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

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

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

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 $V_1, \ldots, V_{N-1} \sim \text{Beta}(1, \alpha)$ 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:

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

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

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{17}$$

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

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

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:

$$(m(x_1), m(x_2), \ldots, m(x_n)) \sim N(\mu(x), K(x)) \tag{19}$$

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

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

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

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

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

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

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

and the log-posterior is

$$\begin{aligned} \log p(y|m) + \log \pi(m) &= -\frac{1}{2\sigma^2}\|y - K\alpha\|_2^2 - \frac{1}{2}\alpha^T K \alpha + \text{const} && (23) \\ &= -\frac{1}{2\sigma^2}\|y - K\alpha\|_2^2 - \frac{1}{2}\|\alpha\|_K^2 + \text{const} && (24) \end{aligned}$$

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 λ, so that $Kv = \lambda v$. Then we have that

$$\frac{1}{\lambda} = v^T K^{-1} v \tag{25}$$

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

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

and thus

$$\widehat{m} = \mathbb{E}(m|Y) = K\left(K + \sigma^2 I\right)^{-1} Y. \tag{27}$$

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 function $K(x, y)$. This is a delicate matter.

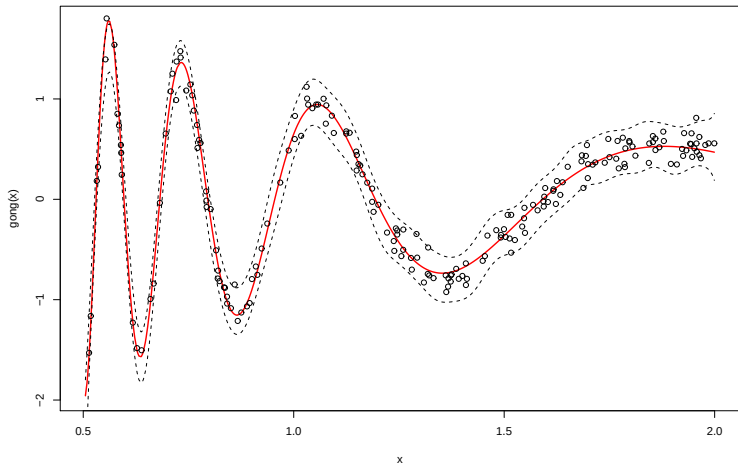


Figure 5: Mean of a Gaussian process

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

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

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

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

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

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

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-valdiation 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 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?

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 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]

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

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,*

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

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 $Y_1, \ldots, Y_n \sim N(\theta, 1)$. 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.

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

[2] More precisely, on an appropriate σ-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

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

A Hellinger neighborhood around P_f is

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

A weak neighborhood around P_f is

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

where d_W is the Prohorov metric

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

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*

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

then, for any $\delta > 0$,

$$\pi_n(B_W(P,\delta)) \stackrel{a.s.}{\to} 1 \tag{38}$$

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 $f_1,\ldots,f_N$ 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

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

Further, suppose there exists $\mathcal{M}_1, \mathcal{M}_2, \ldots$ 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 $\delta > 0$,

$$\pi_n(B_H(P,\delta)) \stackrel{a.s.}{\to} 1 \tag{40}$$

with respect to P_0.

Example 7 *Recall the Normal means model*

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \ldots \tag{41}$$

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*

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

Recall that the estimator $\widehat{\theta}_i = b_i Y_i$ *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,*

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

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 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

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}\epsilon_i, \quad i = 1, 2, \ldots \tag{44}$$

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

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

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:*

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

7.3 Coverage

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

$$\mathbb{P}^n_{f_0}(f_0 \in A) \geq 1 - \alpha? \tag{47}$$

or even

$$\liminf_{n\to\infty} \inf_{f_0} \mathbb{P}^n_{f_0}(f_0 \in A) \geq 1 - \alpha? \tag{48}$$

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

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

then

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

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

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

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

$$\pi_n(A) = 1 - \alpha \tag{52}$$

but

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

8 Bayes Versus Frequentist

People often confuse Bayesian methods and frequentist methods. Bayesian methods are designed for quantifying subjective beliefs. Frequentist methods are designed to create procedures with certain frequency gaurantees (consistency, coverage, minimaxity etc). They are two different things. We should use $F(A)$ for frequency and $B(A)$ for belief and then there would be no confusion. Unfortunately, we use the same symbol $P(A)$ for both which causes endless confusion. Let's take a closer look at the differences.

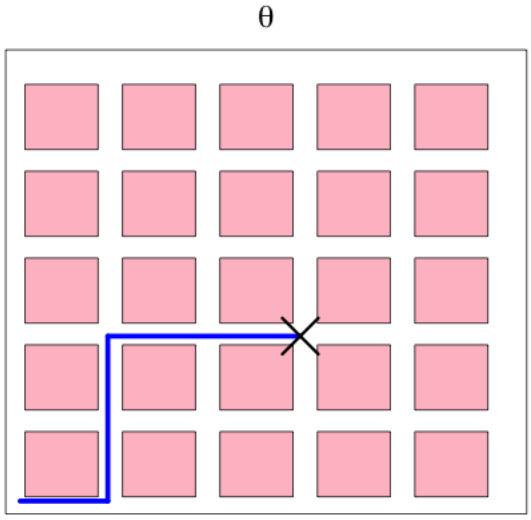
8.1 Adventures in FlatLand: Stone's Paradox

Mervyn Stone is Emeritus Professor at University College London. He is famous for his deep work on Bayesian inference as well as pioneering work on cross-validation, coordinate-free multivariate analysis, as well as many other topics.

Here we discuss a famous example of his, described in Stone (1970, 1976, 1982). In technical jargon, he shows that "a finitely additive measure on the free group with two generators is nonconglomerable." In English: even for a simple problem with a discrete parameters space, flat priors can lead to surprises.

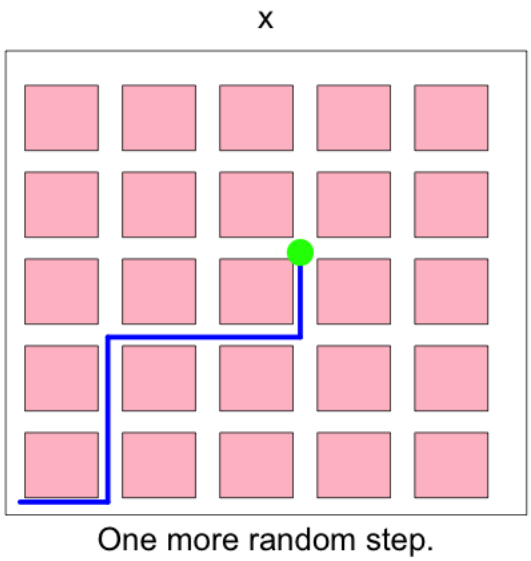
Hunting For a Treasure In Flatland. I wonder randomly in a two dimensional grid-world. I drag an elastic string with me. The string is taut: if I back up, the string leaves no slack. I can only move in four directions: North, South, West, East.

I wander around for a while then I stop and bury a treasure. Call this path θ. Here is an example:

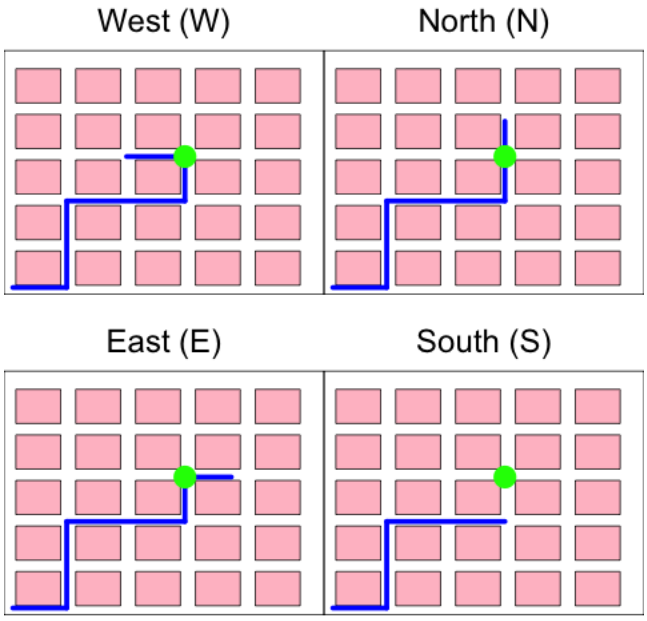


My Path. The X is the treasure.

Now I take one more random step. Each direction has equal probability. Call this path x. So it might look like this:



Two people, Bob (a Bayesian) and Carla (a classical statistician) want to find the treasure. There are only four possible paths that could have yielded x, namely:



Let us call these four paths N, S, W, E. The likelihood is the same for each of these. That is, $p(x|\theta) = 1/4$ for $\theta \in \{N, S, W, E\}$. Suppose Bob uses a flat prior. Since the likelihood is

also flat, his posterior is

$$P(\theta = N|x) = P(\theta = S|x) = P(\theta = W|x) = P(\theta = E|x) = \frac{1}{4}.$$

Let B be the three paths that extend x. In this example, $B = \{N, W, E\}$. Then $P(\theta \in B|x) = 3/4$.

Now Carla is very confident and selects a confidence set with only one path, namely, the path that shortens x. In other words, Carla's confidence set is $C = B^c$.

Notice the following strange thing: no matter what θ is, Carla gets the treasure with probability 3/4 while Bob gets the treasure with probability 1/4. That is, $P(\theta \in B|x) = 3/4$ but the coverage of B is 1/4. The coverage of C is 3/4.

Here is quote from Stone (1976): (except that I changed his B and C to Bob and Carla):

" ... it is clear that when Bob and Carla repeatedly engage in this treasure hunt, Bob will find that his posterior probability assignment becomes increasingly discrepant with his proportion of wins and that Carla is, somehow, doing better than [s]he ought. However, there is no message ... that will allow Bob to escape from his Promethean situation; he cannot learn from his experience because each hunt is independent of the other."

Stone is not criticizing Bayes (as far I can tell). He is just discussing the effect of using a flat prior.

More Trouble For Bob. Let A be the event that the final step reduced the length of the string. Using the posterior above, we see that Bob finds that $P(A|x) = 3/4$ for each x. Since this holds for each x, Bob deduces that $P(A) = 3/4$. On the other hand, Bob notes that $P(A|\theta) = 1/4$ for every θ. Hence, $P(A) = 1/4$. Bob has just proved that $3/4 = 1/4$.

The Source of The Problem. The apparent contradiction stems from the fact that the prior is improper. Technically this is an example of the non-conglomerability of finitely additive measures. For a rigorous explanation of why this happens you should read Stone's papers. Here is an abbreviated explanation, from Kass and Wasserman (1996, Section 4.2.1).

Let π denotes Bob's improper flat prior and let $\pi(\theta|x)$ denote his posterior distribution. Let π_p denote the prior that is uniform on the set of all paths of length p. This is of course a proper prior. For any fixed x, $\pi_p(A|x) \to 3/4$ as $p \to \infty$. So Bob can claim that his posterior distribution is a limit of well-defined posterior distributions. However, we need to look at this more closely. Let $m_p(x) = \sum_\theta f(x|\theta)\pi_p(\theta)$ be the marginal of x induced by π_p. Let X_p denote all x's of length p or $p+1$. When $x \in X_p$, $\pi_p(\theta|x)$ is a poor approximation to $\pi(\theta|x)$ since the former is concentrated on a single point while the latter is concentrated on four points. In fact, the total variation distance between $\pi_p(\theta|x)$ and $\pi(\theta|x)$ is 3/4 for $x \in X_p$. (Recall that the total variation distance between two probability measures P

and Q is $d(P,Q) = \sup_A |P(A) - Q(A)|$.) Furthermore, X_p is a set with high probability: $m_p(X_p) \to 2/3$ as $p \to \infty$.

While $\pi_p(\theta|x)$ converges to $\pi(\theta|x)$ as $p \to \infty$ for any fixed x, they are not close with high probability. This problem disappears if you use a proper prior. (But that still does not give frequntist coverage.)

The Four Sided Die. Here is another description of the problem. Consider a four sided die whose sides are labeled with the symbols $\{a, b, a^{-1}, b^{-1}\}$. We roll the die several times and we record the label on the lowermost face (there is a no uppermost face on a four-sided die). A typical outcome might look like this string of symbols:

$$a \quad a\ b\ a^{-1}\ b\ b^{-1}\ b\ a\ a^{-1}\ b$$

Now we apply an annihilation rule. If a and a^{-1} appear next to each other, we eliminate these two symbols. Similarly, if b and b^{-1} appear next to each other, we eliminate those two symbols. So the sequence above gets reduced to:

$$a \quad a\ b\ a^{-1}\ b\ b$$

Let us denote the resulting string of symbols, after removing annihilations, by θ. Now we toss the die one more time. We add this last symbol to θ and we apply the annihilation rule once more. This results in a string which we will denote by x.

You get to see x and you want to infer θ.

Having observed x, there are four possible values of θ and each has the same likelihood. For example, suppose $x = (a, a)$. Then θ has to be one of the following:

$$(a), \quad (a\,a\,a), \quad (a\,a\,b^{-1}), \quad (a\,a\,b)$$

The likelihood function is constant over these four values.

Suppose we use a flat prior on θ. Then the posterior is uniform on these four possibilities. Let $C = C(x)$ denote the three values of θ that are longer than x. Then the posterior satisfies

$$P(\theta \in C|x) = 3/4.$$

Thus $C(x)$ is a 75 percent posterior confidence set.

However, the frequentist coverage of $C(x)$ is 1/4. To see this, fix any θ. Now note that $C(x)$ contains θ if and only if θ concatenated with x is smaller than θ. This happens only if the last symbol is annihilated, which occurs with probability 1/4.

Likelihood. Another consequence of Stone's example is that it shows that the Likelihood Principle is bogus. According to the likelihood principle, the observed likelihood function

contains all the useful information in the data. In this example, the likelihood does not distinguish the four possible parameter values. The direction of the string from the current position — which does not affect the likelihood — has lots of information.

Proper Priors. If you want to have some fun, try coming up with proper priors on the set of paths. Then simulate the example, find the posterior and try to find the treasure. If you try this, I'd be interested to hear about the results.

Another question this example raises is: should one use improper priors? Flat priors that do not have a finite sum can be interpreted as finitely additive priors. The father of Bayesian inference, Bruno DeFinetti, was adamant in rejecting the axiom of countable additivity. He thought flat priors like Bob's were fine.

It seems to me that in modern Bayesian inference, there is not universal agreement on whether flat priors are evil or not. But in this example, I think that most statisticians would reject Bob's flat prior-based Bayesian inference.

8.2 The Robins-Ritov Example

This example is due to Robins and Ritov (1997). A simplified version appeared in Wasserman (2004) and Robins and Wasserman (2000). The example is related to ideas from the foundations of survey sampling (Basu 1969, Godambe and Thompson 1976) and also to ancillarity paradoxes (Brown 1990, Foster and George 1996).

Here is (a version of) the example. Consider iid random variables

$$(X_1, Y_1, R_1), \ldots, (X_n, Y_n, R_n).$$

The random variables take values as follows:

$$X_i \in [0,1]^d, \quad Y_i \in \{0,1\}, \quad R_i \in \{0,1\}.$$

Think of d as being very, very large. For example, $d = 100,000$ and $n = 1,000$.

The idea is this: we observe X_i. Then we flip a biased coin R_i. If $R_i = 1$ then you get to see Y_i. If $R_i = 0$ then you don't get to see Y_i. The goal is to estimate

$$\psi = P(Y_i = 1).$$

Here are the details. The distribution takes the form

$$p(x, y, r) = p_X(x) p_{Y|X}(y|x) p_{R|X}(r|x).$$

Note that Y and R are independent, given X. For simplicity, we will take $p(x)$ to be uniform on $[0,1]^d$. Next, let

$$\theta(x) \equiv p_{Y|X}(1|x) = P(Y = 1|X = x)$$

where $\theta(x)$ is a function. That is, $\theta : [0,1]^d \to [0,1]$. Of course,

$$p_{Y|X}(0|x) = P(Y = 0|X = x) = 1 - \theta(x).$$

Similarly, let

$$\pi(x) \equiv p_{R|X}(1|x) = P(R = 1|X = x)$$

where $\pi(x)$ is a function. That is, $\pi : [0,1]^d \to [0,1]$. Of course,

$$p_{R|X}(0|x) = P(R = 0|X = x) = 1 - \pi(x).$$

The function π is **known.** We construct it. Remember that $\pi(x) = P(R = 1|X = x)$ is the probability that we get to observe Y given that $X = x$. Think of Y as something that is expensive to measure. We don't always want to measure it. So we make a random decision about whether to measure it. And we let the probability of measuring Y be a function $\pi(x)$ of x. And we get to construct this function.

Let $\delta > 0$ be a known, small, positive number. We will assume that

$$\pi(x) \geq \delta$$

for all x.

The only thing in the the model we don't know is the function $\theta(x)$. Again, we will assume that

$$\delta \leq \theta(x) \leq 1 - \delta.$$

Let Θ denote all measurable functions on $[0,1]^d$ that satisfy the above conditions. The parameter space is the set of functions Θ.

Let $\mathcal{P}$ be the set of joint distributions of the form

$$p(x)\,\pi(x)^r(1-\pi(x))^{1-r}\,\theta(x)^y(1-\theta(x))^{1-y}$$

where $p(x) = 1$, and $\pi(\cdot)$ and $\theta(\cdot)$ satisfy the conditions above. So far, we are considering the sub-model $\mathcal{P}_\pi$ in which π is known.

The parameter of interest is $\psi = P(Y = 1)$. We can write this as

$$\psi = P(Y = 1) = \int_{[0,1]^d} P(Y = 1|X = x)p(x)dx = \int_{[0,1]^d} \theta(x)dx.$$

Hence, ψ is a function of θ. If we know $\theta(\cdot)$ then we can compute ψ.

The usual frequentist estimator is the Horwitz-Thompson estimator

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i R_i}{\pi(X_i)}.$$

It is easy to verify that $\widehat{\psi}$ is unbiased and consistent. Furthermore, $\widehat{\psi} - \psi = O_P(n^{-\frac{1}{2}})$. In fact, let us define

$$I_n = [\widehat{\psi} - \epsilon_n,\ \widehat{\psi} + \epsilon_n]$$

where

$$\epsilon_n = \sqrt{\frac{1}{2n\delta^2} \log\left(\frac{2}{\alpha}\right)}.$$

It follows from Hoeffding's inequality that

$$\sup_{P \in \mathcal{P}_\pi} P(\psi \in I_n) \geq 1 - \alpha$$

Thus we have a finite sample, $1 - \alpha$ confidence interval with length $O(1/\sqrt{n})$.

Remark: *We are mentioning the Horwitz-Thompson estimator because it is simple. In practice, it has three deficiencies:*

1. *It may exceed 1.*
2. *It ignores data on the multivariate vector X except for the one dimensional summary $\pi(X)$.*
3. *It can be very inefficient.*

These problems are remedied by using an improved version of the Horwitz-Thompson estimator. One choice is the so-called locally semiparametric efficient regression estimator *(Scharfstein et al., 1999):*

$$\widehat{\psi} = \int \text{expit}\left(\sum_{m=1}^{k} \widehat{\eta}_m \phi_m(x) + \frac{\widehat{\omega}}{\pi(x)}\right) dx$$

where $\text{expit}(a) = e^a/(1 + e^a)$, $\phi_m(x)$ *are basis functions, and* $\widehat{\eta}_1, \ldots, \widehat{\eta}_k, \widehat{\omega}$ *are the mle's (among subjects with* $R_i = 1$*) in the model*

$$\log\left(\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}\right) = \sum_{m=1}^{k} \eta_m \phi_m(x) + \frac{\omega}{\pi(x)}.$$

Here k can increase slowly with n. Recently even more efficient estimators have been derived. Rotnitzky et al (2012) provides a review. In the rest of this post, when we refer to the Horwitz-Thompson estimator, the reader should think "improved Horwitz-Thompson estimator."

To do a Bayesian analysis, we put some prior W on Θ. Next we compute the likelihood function. The likelihood for one observation takes the form $p(x)p(r|x)p(y|x)^r$. The reason

for having r in the exponent is that, if $r = 0$, then y is not observed so the $p(y|x)$ gets left out. The likelihood for n observations is

$$\prod_{i=1}^{n} p(X_i)p(R_i|X_i)p(Y_i|X_i)^{R_i} = \prod_i \pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}\,\theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

where we used the fact that $p(x) = 1$. But remember, $\pi(x)$ is known. In other words, $\pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}$ is known. So, the likelihood is

$$\mathcal{L}(\theta) \propto \prod_i \theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

Combining this likelihood with the prior W creates a posterior distribution on Θ which we will denote by W_n. Since the parameter of interest ψ is a function of θ, the posterior W_n for θ defines a posterior distribution for ψ.

Now comes the interesting part. The likelihood has essentially no information in it.

To see that the likelihood has no information, consider a simpler case where $\theta(x)$ is a function on $[0,1]$. Now discretize the interval into many small bins. Let B be the number of bins. We can then replace the function θ with a high-dimensional vector $\theta = (\theta_1, \ldots, \theta_B)$. With $n < B$, most bins are empty. The data contain no information for most of the θ_j's. (You might wonder about the effect of putting a smoothness assumption on $\theta(\cdot)$. We'll discuss this in Section 4.)

We should point out that if $\pi(x) = 1/2$ for all x, then Ericson (1969) showed that a certain exchangeable prior gives a posterior that, like the Horwitz-Thompson estimator, converges at rate $O(n^{-1/2})$. However we are interested in the case where $\pi(x)$ is a complex function of x; then the posterior will fail to concentrate around the true value of ψ. On the other hand, a flexible nonparametric prior will have a posterior essentially equal to the prior and, thus, not concentrate around ψ, whenever the prior W does not depend on the the known function $\pi(\cdot)$. Indeed, we have the following theorem from Robins and Ritov (1997):

Theorem. (Robins and Ritov 1997). Any estimator that is not a function of $\pi(\cdot)$ cannot be uniformly consistent.

This means that, at no finite sample size, will an estimator $\widehat{\psi}$ that is not a function of π be close to ψ for all distributions in $\mathcal{P}$. In fact, the theorem holds for a neighborhood around every pair (π, θ). Uniformity is important because it links asymptotic behavior to finite sample behavior. But when π is known and is used in the estimator (as in the Horwitz-Thompson estimator and its improved versions) we can have uniform consistency.

Note that a Bayesian will ignore π since the $\pi(X_i)'s$ are just constants in the likelihood. There is an exception: the Bayesian can make the posterior be a function of π by choosing a prior W that makes $\theta(\cdot)$ depend on $\pi(\cdot)$. But this seems very forced. Indeed, Robins and

Ritov showed that, under certain conditions, any true subjective Bayesian prior W must be independent of $\pi(\cdot)$. Specifically, they showed that once a subjective Bayesian queries the randomizer (who selected π) about the randomizer's reasoned opinions concerning $\theta(\cdot)$ (but not $\pi(\cdot)$) the Bayesian will have independent priors. We note that a Bayesian can have independent priors even when he believes with probabilty 1 that $\pi(\cdot)$ and $\theta(\cdot)$ are positively correlated as functions of x i.e. $\int \theta(x)\pi(x)\,dx > \int \theta(x)\,dx \int \pi(x)\,dx$. Having independent priors only means that learning $\pi(\cdot)$ will not change one's beliefs about $\theta(\cdot)$.

9 Freedman's Theorem

Here I will summarize an interesting result by David Freedman (Annals of Mathematical Statistics, Volume 36, Number 2 (1965), 454-456) available at projecteuclid.org.

The result gets very little attention. Most researchers in statistics and machine learning seem to be unaware of the result. The result says that, "almost all" Bayesian posterior distributions are inconsistent, in a sense we'll make precise below. The math is uncontroversial but, as you might imagine, the intepretation of the result is likely to be controversial.

Let $X_1, \ldots, X_n$ be an iid sample from a distribution P on the natural numbers $I = \{1, 2, 3, \ldots, \}$. Let $\mathcal{P}$ be the set of all such distributions. We endow $\mathcal{P}$ with the weak* topology. Hence, $P_n \to P$ iff $P_n(i) \to P(i)$ for all i.

Let μ denote a prior distribution on $\mathcal{P}$. (More precisely, a prior on an appropriate σ-field.) Let Π be all priors. Again, we endow the set with the weak* topology. Thus $\mu_n \to \mu$ iff $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous, real functions f.

Let μ_n be the posterior corresponding to the prior μ after n observations. We will say that the pair (P, μ) is consistent if

$$P^\infty(\lim_{n\to\infty} \mu_n = \delta_P) = 1$$

where P^∞ is the product measure corresponding to P and δ_P is a point mass at P.

Now we need to recall some topology. A set is nowhere dense if its closure has an empty interior. A set is meager (or first category) if it is a countable union of nowehere dense sets. Meager sets are small; think of a meager set as the topological version of a null set in measure theory.

Freedman's theorem is: the sets of consistent pairs (P, μ) is meager.

This means that, in a topological sense, consistency is rare for Bayesian procedures. From this result, it can also be shown that most pairs of priors lead to inferences that disagree. (The agreeing pairs are meager.) Or as Freedman says in his paper:

" ... it is easy to prove that for essentially any pair of Bayesians, each thinks the other is

crazy."

10 References

Basu, D. (1969). Role of the Sufficiency and Likelihood Principles in Sample Survey Theory. *Sankya*, 31, 441-454.

Brown, L.D. (1990). An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics*, 18, 471-493.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B,* 195-233.

Foster, D.P. and George, E.I. (1996). A simple ancillarity paradox. *Scandinavian journal of statistics*, 233-242.

Godambe, V. P., and Thompson, M. E. (1976), Philosophy of Survey-Sampling Practice. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, eds. W.L.Harper and A.Hooker, Dordrecht: Reidel.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.

Robins, J.M. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models. *Statistics in Medicine*, 16, 285–319.

Robins, J. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: a review of some foundational concepts. *Journal of the American Statistical Association*, 95, 1340-1346.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J.M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 439-456.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 1096-1120.

Stone, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics*, 41, 1349-1353,

Stone, M. (1976). Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71, 114-116.

Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. *Studies in Logic and the Foundations of Mathematics*, 104, 413-426.

Wasserman, L. (2004). *All of Statistics: a Concise Course in Statistical Inference*. Springer Verlag.