A Bayesian Analysis of Some Nonparametric Problems

Thomas S Ferguson

April 18, 2003

Introduction

Bayesian approach remained rather unsuccessful in treating nonparametric problems. This is primarily due to the difficulty in finding workable prior distribution on the parameter space , which in nonparametric problems is taken to be a set of probability distributions on a given sample space.

Two Desirable Properties of a Prior

1. The support of the prior should be large.
2. Posterior distribution given a sample of observation should be manageable analytically.

These properties are antagonistic : One may be obtained at the expense of other.

The Main Idea of the Paper

It presents a class of prior distributions called Dirichlet process priors, broad in the sense of (1), for which (2) is realized and for which treatment of many nonparametric problems may be carried out, yielding results that are comparable to the classical theory.

The Dirichlet Distribution

• Known to the Bayesians as the conjugate prior for the parameter of a multinomial distribution.

$\mathcal{G}(\alpha,\beta)$ denote the gamma distribution with shape parameter $\alpha \geq 0$, scale parameter $\beta > 0$.
$\alpha = 0$ it is degenerate at 0.

Let $z_j \overset{iid}{\sim} \mathcal{G}(\alpha, 1), j = 1, \ldots, k$.
Where $\alpha_j \geq 0\ \forall j$ and $\alpha_j > 0$ for some j.

The Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_k)$ denoted by $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ is defined as the distribution of $(Y_1, \ldots, Y_k)$, where

$$Y_j = \frac{Z_j}{\sum_{i=1}^{k} Z_i}, j = 1, \ldots, k. \tag{1}$$

Note:
• Since $Y_1 + \ldots + Y_k = 1$, the distribution is always singular.
• If any $\alpha_j = 0$, the corresponding Y_j is degenerate at 0.
• If $\alpha_j > 0, \forall j$ distribution of $(Y_1, \ldots, Y_k)$ is continuous with density

$$\begin{aligned} f &\quad (y_1, \ldots, y_{k-1}) \qquad (2) \\ &= \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} (\prod_{j=1}^{k-1} y_j^{\alpha_j - 1})(1 - \sum_{j=1}^{k-1} y_j)^{\alpha_{k-1} - 1} \\ &\times \quad I_S(y_1, \ldots, y_{k-1}) \end{aligned}$$

$$S = \{(y_1, \ldots, y_{k-1}) : y_j \geq 0, \sum_{j=1}^{k-1} y_j \leq 1\}$$

for $k = 2$, (2) reduces to Beta distribution.

Properties of Dirichlet Distribution

(i) If $(Y_1, \dots, Y_k) \in \mathcal{D}(\alpha_1, \dots, \alpha_k)$. Let $r_1, \dots, r_l$ are integers , $\ni 0 < r_1 < \dots < r_l = k$, then
$(\sum_1^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i, \dots, \sum_{r_l-1}^{r_l} Y_i)$
$\in \mathcal{D}(\sum_1^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, \dots, \sum_{r_l-1}^{r_l} \alpha_i)$.

(ii) Marginal distribution of Y_j is Beta.
$Y_j \in \mathcal{B}e(\alpha_j, (\sum_1^k) - \alpha_j)$

(iii) $E(Y_i) = \frac{\alpha_i}{\alpha}$
$E(Y_i^2) = \frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)}$
$E(Y_iY_j) = \frac{\alpha_i\alpha_j}{\alpha(\alpha+1)}, i \neq j, \alpha = \sum_1^k \alpha_i$

The Dirichlet Process

$\mathcal{X}$: a set.
$\mathcal{A}$: σ-field of subsets of $\mathcal{X}$.
α : finite non-null measure on $(\mathcal{X}, \mathcal{A})$.

Define a random probability P by defining the joint distribution of $(P(B_1), \dots, P(B_k)), \forall k$ and all measurable partitions $(B_1, \dots, B_k)$ of $\mathcal{X}$.

We say P is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α if for everly $k = 1, 2, \dots,$ and measurable partition $(B_1, \dots, B_k)$ of $\mathcal{X}$, the distribution of $(P(B_1), \dots, P(B_k))$ is Dirichlet, $\mathcal{D}(\alpha(B_1), \dots, \alpha(B_k))$.

Proposition 1. Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α and let $A \in \mathcal{A}$. If $\alpha(A) = 0$, then $P(A) = 0$ with probability one. If $\alpha(A) > 0$ then $P(A) > 0$ with probability one. Furthermore, $E(P(A)) = \frac{\alpha(A)}{\alpha(\mathcal{X})}$.

Proposition 2. Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α, and let Q be a fixed probability measure on $(\mathcal{X}, \mathcal{A})$ with $Q \ll \alpha$. Then, for any positive integer m and measurable sets $A_1, \ldots, A_m$, and $\epsilon > 0$, $\mathcal{P}\{|P(A_i) - Q(A_i)| < \epsilon \text{ for } i = 1, \ldots, m\} > 0$.

Proposition 3. Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α and let X be a sample of size 1 from P. Then for $A \in \mathcal{A}$, $\mathcal{P}(X \in A) = \alpha(A)/\alpha(\mathcal{X})$.

Theorem 1 : Let P be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α , and let $X_1, \ldots, X_n$ be a sample of size n from P. Then conditional distribution of P given $X_1, \ldots, X_n$, is a Dirichlet process with parameter $\alpha + \sum_1^n \delta_{X_i}$.

Theorem 2 : Let P be the Dirichlet process and let Z be a measurable real valued function defined on $(\mathcal{X}, \mathcal{A})$. If $\int |Z| d\alpha < \infty$, then $\int |Z| dP$ with probability one, and $E(\int Z dP) = \int Z dE(P) = \alpha(\mathcal{X})^{-1} \int Z d\alpha$.

Application

- $P \in \mathcal{D}(\alpha)$ says "P is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α."

- $\mathcal{R}$: The real line.

- $\mathcal{B}$: The σ-field of Borel sets.

- $(\mathcal{X}, \mathcal{A}) = (\mathcal{R}, \mathcal{B})$

Nonparametric Statistical Decision Problem

- The parameter space is the set of all probability measures P on $(\mathcal{X}, \mathcal{A})$.

- Statistician is to choose an action a in some space ,thereby incurring a loss , $L(P, a)$.

- A sample $X_1, \ldots, X_n$ from P are available to the statistician, upon which he may base his choice of action.

What a Bayesian does ?

- He seeks a Bayes rule with respect to the prior distribution, $P \in \mathcal{D}(\alpha)$.

- The posterior distribution of P given the observations is $\mathcal{D}(\alpha + \sum_1^n \delta_{X_i})$.

- Find a Bayes rule for no sample problem (n = 0).

- Bayes rule for the general problem is found by replacing α by $\alpha + \sum_1^n \delta_{X_i}$.

(a) **Estimation of a distribution function**

Let $(\mathcal{X}, \mathcal{A}) = (\mathcal{R}, \mathcal{B})$, and the space of action is the space of all distribution function on $\mathcal{R}$.

Loss function : $L(P, \hat{F}) = \int (F(t) - \hat{F}(t))^2 dW(t)$
W : a given finite measure on $(\mathcal{R}, \mathcal{B})$
(a weight function).

$F(t) = P((-\infty, t])$

$P \in \mathcal{D}(\alpha) \Rightarrow F(t) \in \mathcal{B}e(\alpha((-\infty, t]), \alpha((t, \infty]))$ for each t.

Bayes risk for no sample problem ,
$E(L(P, \hat{F})) = \int E(F(t) - \hat{F}(t))^2 dW(t)$

Minimized by choosing $\hat{F}(t) = E(F(t))$.

Thus the bayes rule for the no-sample problem is $\hat{F}(t) = E(F(t)) = F_0(t)$
Where, $F_0(t) = \alpha((-\infty, t])/\alpha(\mathcal{R})$ For a sample of size n the bayes rule is ,

$$\begin{aligned} \hat{F}_n & \quad (t|X_1, \dots, X_n) \\ = & \quad \frac{\alpha((-\infty, t]) + \sum_1^n \delta_{X_i}((-\infty, t])}{\alpha(\mathcal{R}) + n} \\ = & \quad p_n F_0(t) + (1 - p_n) F_n(t|X_1, \dots, X_n) \end{aligned}$$

$p_n = \alpha(\mathcal{R})/(\alpha(\mathcal{R}) + n)$
$F_n(t|X_1, \dots, X_n) = \frac{1}{n} \sum_1^n \delta_{X_i}((-\infty, t])$

The Bayes rule is a mixture of our prior guess at F and the empirical distribution.

$\alpha(\mathcal{R})$ can be interpreted as a measure of faith in the prior guess at F.

(b) **Estimation of the mean**

Let $(\mathcal{X},\mathcal{A}) = (\mathcal{R},\mathcal{B})$.
Loss function : $L(P,\hat{\mu}) = (\mu - \hat{\mu})^2$
$\mu = \int x dP(x)$

Assume $P \in \mathcal{D}$, where α has finite first moment.The mean of the corresponding probability measure $\alpha(.)/\alpha(\mathcal{R})$ is ,
$\mu_0 = \int x d\alpha(x)/\alpha(\mathcal{R})$

By Theorem 2, the r.v μ exists. The Bayes rule for the no-sample problem is the mean of μ , $\hat{\mu} = \mu_0$, by Theorem 2 . For a sample of size n, the Bayes rule is,

$$\begin{aligned} \hat{\mu}_n &\quad (X_1,\ldots,X_n) \\ &= (\alpha(\mathcal{R}) + n)^{-1} \int x d(\alpha(x) + \sum_1^n \delta_{X_i}(x)) \\ &= p_n\mu_0 + (1-p_n)\bar{X}_n \end{aligned}$$

$$p_n \text{ same as before and } \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \ .$$

• The Bayes estimate is between the prior guess at μ and the sample mean.

• As $\alpha(\mathcal{R}) \to 0, \hat{\mu}_n$ converges to $\bar{X}_n$.

• As $n \to \infty, p_n \to 0$, i.e the Bayes estimate is strongly consistent within the class of distributions with finite first moments.

(c) **Estimation of the median**

Let $(\mathcal{X}, \mathcal{A}) = (\mathcal{R}, \mathcal{B})$.
We need to estimate the median,
$m =$ med P ,of an unknown probability measure P on $(\mathcal{R}, \mathcal{B})$.

For $P \in \mathcal{D}(\alpha), m$ is a random variable.

Loss function : $L(P, \hat{m}) = |m - \hat{m}|$
(absolute error loss).

Any median of the distribution of m is a Bayes estimate of m. For the Dirichlet process prior, any median of the distribution of m is a median of the expectation of P :
med(dist. med P) $=$ med $E(P)$

A number t is a median of the distribution of m iff $\mathcal{P}\{m < t\} \leq \frac{1}{2} \leq \mathcal{P}\{m \leq t\}$.
$\mathcal{P}\{m \leq t\} = \mathcal{P}\{F(t) > \frac{1}{2}\}$

$F(t)$ has a $\mathcal{B}e(\alpha((-\infty, t]), \alpha((t, \infty]))$. Its median is a nondecreasing function of t, with value half iff the 2 parameters are equal.

Hence t satisfies the above
iff $\frac{\alpha((-\infty,t))}{\alpha(\mathcal{R})} \leq \frac{1}{2} \leq \frac{\alpha((-\infty,t])}{\alpha(\mathcal{R})}$.
$\Rightarrow$ such t are the medians of $E(P)$.

Thus any number t satisfying the above is a Bayes estimate of m for prior $\mathcal{D}(\alpha)$ and absolute error loss. For F_0 defined as in (a) ,
$\hat{m} =$ median of F_0.

For a sample size n, the Bayes estimate is , $\hat{m}_n(X_1, \ldots, X_n) =$ median of $\hat{F}_n$.
$\hat{F}_n$ is the Bayes estimate of F as obtained from (a).

(d) **Estimation of $\int FdG$ for a two-sample problem**

F and G are 2 distribution functions on the real line. Let $X_1, \ldots, X_m$ be a sample from F and $Y_1, \ldots, Y_n$ a sample from G.

Estimation of $\Delta = P(X_1 \leq Y_1)$,
$\Delta = \int FdG$
Loss function: Squared error.

Priors : F is the d.f. of $P_1 \in \mathcal{D}(\alpha_1)$
G is the d.f. of $P_2 \in \mathcal{D}(\alpha_2)$
P_1 and P_2 are independent.

The Bayes rule for the no-sample problem is, $\Delta_0 = E(\Delta) = \int F_0 dG_0$
$F_0 = E(F)$; $G_0 = E(G)$.

Given 2 samples, Bayes rule is
$\hat{\Delta}(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \int \hat{F}_m d\hat{G}_n$
$\hat{F}_m$ and $\hat{G}_n$ are the respective Bayes estimate of F and G as in application (a).

It can be written as,

$$\begin{aligned}
\hat{\Delta} & \ (X_1, \ldots, X_m, Y_1, \ldots, Y_n) \\
= & \ p_{1,m} p_{1,n} \Delta_0 + p_{1,m}(1 - p_{2,n}) \frac{1}{n} \sum_1^n F_0(Y_j) \\
+ & \ (1 - p_{1,m}) p_{2,n} \frac{1}{n} \sum_1^m (1 - G_0(X_i^-)) \\
+ & \ (1 - p_{1,m})(1 - p_{2,n}) \frac{1}{mn} U
\end{aligned}$$

$p_{1,m} = \frac{\alpha_1(\mathcal{R})}{\alpha_1(\mathcal{R}) + m}$, $p_{2,m} = \frac{\alpha_2(\mathcal{R})}{\alpha_2(\mathcal{R}) + n}$

$U = \sum_i \sum_j I_{(-\infty, Y_j)}(X_i)$ is the Mann-Whitney statistic.