# A Bayesian Analysis of Some Nonparametric Problems

Thomas S. Ferguson, 1973

# Roadmap

1. Brief Motivation
2. Dirichlet Distribution
3. Construction of the Dirichlet Process
4. Properties of the Dirichlet Process: large support and conjugacy
5. Alternative construction of the Dirichlet Process and consequences
6. Applications: distribution function estimation, mean estimation

# Motivation

- We will build and analyze the properties of the Dirichlet process.
- A good nonparametric prior should have two features:
  1. it should have a large support.
  2. It should be analytically tractable.
- Clearly, these two features are antithetical: easy to get one if we sacrifice the other, but we want both.

# Dirichlet Distribution I

- Let $Z_1, ..., Z_n$ be independent random variables with distribution Gamma with non negative shape parameter $\alpha_1, ..., \alpha_n$ and scale parameter $1$.
- Then, the vector $Y = (Y_1, ..., Y_n)$ has **Dirichlet distribution** with parameters $(\alpha_1, ..., \alpha_n)$, denoted as $Y \in \mathfrak{D}(\alpha_1, ..., \alpha_n)$ if each $Y_i$ is defined as
$$Y_i := \frac{Z_i}{\sum_{j=1}^{n} Z_j}$$
- The marginal distribution of each $Y_i$ is a beta with parameters $(\alpha_i, \sum_{j \neq i} \alpha_j)$

# Dirichlet Distribution II

- The Dirichlet distribution is singular with respect to the n-dimensional Lebesgue measure. However, the joint distribution of $Y_1, ..., Y_{n-1}$ has density

$$f(y_1, ..., y_{n-1}) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n-1} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} y_i\right)^{\alpha_n - 1} \mathbf{1}_{\mathbb{S}}(y)$$

  Here, $\mathbb{S} := \{y \in \mathbb{R}^{n-1} : y \geq 0, \sum y_i \leq 1\}$ .

# Construction of the Dirichlet Process

- Let $\mathcal{X}$ be a set endowed with a $\sigma$-field $\mathcal{A}$.
- We want to define a random probability $P$ by defining the joint distribution of random variables $(P(A_1), ..., P(A_n))$ for each finite sequence of measurable sets. $(A_i)_{i=1}^n$.
- To do so, we fix the distribution of $(P(B_1), ..., P(B_k))$ for every $k$, where $B_1, ..., B_k$ is a measurable partition of $\mathcal{X}$.

# Construction of the Dirichlet Process I

- Take a finite number of arbitrary measurable sets $A_1, ..., A_m$. If each $\nu_j = 0$ or $1$, we can define $B_{\nu_1,...,\nu_m}$ as

$$B_{\nu_1,...,\nu_m} := \bigcap_{j=1}^{m} A_j^{\nu_j}$$

  Where $A_j^0 = A_j^c$, $A_j^1 = A_j$. Then, $\{B_{\nu_1,...,\nu_m}\}_{\nu \in \{0,1\}^n}$ is a measurable partition of $\mathcal{X}$.

- Then, we can define

$$P(A_i) = \sum_{(\nu_j)_{j=1}^m :\ \nu_i = 1} P(B_{\nu_1,...,\nu_m})$$

- If $(A_1, ..., A_m)$ is a partition to begin with, no contradiction as long as we assume that $P(\emptyset)$ is degenerate at 0.

# Construction of the Dirichlet Process II: consistency

## Condition C

If $(B'_1, ..., B'_k)$ and $(B_1, ..., B_n)$ are measurable partitions, and if $(B'_1, ..., B'_k)$ is a refinement of $(B_1, ..., B_n)$, so that we can find $r_1, ..., r_{n-1}$ such that for each $i$, it holds

$$B_i = \bigcup_{j=r_{i-1}+1}^{r_i} B'_j$$

then the distribution of $\left(\sum_{i=1}^{r_1} P(B'_i),\ \ldots\ , \sum_{i=r_{n-1}+1}^{k} P(B'_i)\right)$ is identical to the distribution of $((P(B_1)..., P(B_n))$.

- Notice that, assuming that $P$ is an additive measure, $P(B_i) = \sum_{j=r_{i-1}+1}^{r_i} P(B'_j)$: we are asking that the distribution of $P(B_i)$ is thus consistent with the distribution of the $P(B'_j)$s.

# Construction of the Dirichlet Process

## Lemma 1: Existence

If a system of joint distributions of $(P(B_1), ..., P(B_k))$ for all $k$ and all measurable partitions $(B_i)_{i=1}^{k}$ is defined satistying condition C and if for arbitrary measurable sets $(A_1, ..., A_m)$ the distribution of $(P(A_1), ..., P(A_m))$ is defined as above, then there exists a probability measure $\mathscr{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{B}([0,1]^{\mathcal{A}}))$ yielding these distributions.

- Application of Kolmogorov's existance theorem.
- We will call $P$ a **random measure** if condition C is satisfied, $P(A)$ only takes values in $[0,1]$ and if $P(\mathcal{X}) = 1$ with probability 1.

# Construction of the Dirichlet Process III

## Definition

Let $\alpha$ be a non-null finite measure on $(\mathcal{X}, \mathcal{A})$. Then, we say that $P$ is a **Dirichlet Process** on $(\mathcal{X}, \mathcal{A})$ with parameter $\alpha$ (denoted as $P \in \mathfrak{D}(\alpha)$) if for every $k$ and every measurable partition $(B_i)_{i=1}^k$ of $\mathcal{X}$ we have $(P(B_1), ..., P(B_k)) \in \mathfrak{D}(\alpha(B_1), ..., \alpha(B_k))$.

- By Lemma 1 it follows that $P$ defined as above is a well defined random process.
- $(P(\mathcal{X}), P(\emptyset)) \in \mathfrak{D}(\alpha(\mathcal{X}), 0)$, which implies that $P(\mathcal{X})$ is degenerate at one. Then $P$ is a proper random measure.

# Key Properties of the Dirichlet Process I: Large support

**Proposition 3:**

Let $P$ be a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter $\alpha$, and let $Q$ be a fixed probability measure on $(\mathcal{X}, \mathcal{A})$, with $Q \ll \alpha$. Then, for every $m \in \mathbb{N}$, $\varepsilon > 0$ and measurable sets $A_1, ..., A_m$ we have

$$\mathscr{P}\left(\{|P(A_i) - Q(A_i)| < \varepsilon \ \forall i = 1, ..., m\}\right) > 0$$

- If we endow $[0,1]^{\mathcal{A}}$ with the topology of weak convergence, the support of the Dirichlet process is the set of all probability measures whose support is contained in the support of $\alpha$.

# Key Properties of the Dirichlet Process: Tractability

**Theorem 1**

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

- The Dirichlet process is a non parametric conjugate prior.
- Easy posterior computation.

# Alternative Construction

- Let $\alpha$ be a non-null finite measure, and let $N(x) := -\alpha(\mathcal{X}) \int_x^\infty \frac{e^{-y}}{y} dy$, where $x > 0$.
- We define the distribution of random variables $J_1, ..., J_n$ as

$$\mathscr{P}(J_1 \leq x_1) = e^{N(x_1)} \quad x_1 > 0$$

and for $j = 1, 2, 3...$ and $0 < x_i < x_{i-1}$,

$$\mathscr{P}(J_i \leq x_i | J_{i-1} = x_{i-1}, ..., J_1 = x_1) = e^{N(x_i) - N(x_{i-1})}$$

- It can be shown that $Z_1 := \sum_{i=1}^\infty J_i$ converges with probability one and that $Z_1$ is a Gamma with parameters $\alpha(\mathcal{X})$ and 1.
- Define $Q(A) = \frac{\alpha(A)}{\alpha(\mathcal{X})}$, let $V_j : (\mathcal{X}^\infty, \mathcal{A}^\infty, Q^\infty) \to (\mathcal{X}, \mathcal{A})$ be the random variable $(x_i)_{i \in \mathbb{N}} \mapsto x_j$.

# Alternative construction II

**Theorem 2**

The random probability measure defined by

$$P(A) := \frac{1}{Z_1} \sum_{n \in \mathbb{N}} J_n \delta_{V_n}(A)$$

is a Dirichlet process with parameter $\alpha$.

- Intuition: constructed similarly to Dirichlet distribution (with ratios of gamma distributions)
- The Dirichlet process can be written as a series of point masses.
- Important implication: the realization of the Dirichlet process are discrete with probability 1.

# Key Properties of the Dirichlet Process: Tractability

**Theorem 3**

Let $P$ be a Dirichlet process with parameter $\alpha$, and let $Z$ be a measurable real valued function defined on $(\mathcal{X}, \mathcal{A})$. If $\int |Z| d\alpha < \infty$, then $\mathscr{P}(\{\int |Z| dP < \infty\}) = 1$ and

$$\mathbb{E}\left(\int Z dP\right) = \int Z d\mathbb{E}(P) = \frac{1}{\alpha(\mathcal{X})} \int Z d\alpha.$$

- In particular, if $\mathcal{X} = \mathbb{R}$ and if $\alpha$ has a finite n-th moment, with probability 1 $P$ has a finite n-th moment.

# Application: distribution function estimation

- We will take $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose that $W$ is a finite measure; the stastician seeks a rule with respect to the prior distribution, $P \in \mathfrak{D}(\alpha)$.

- If we can find a rule for the no sample problem, we immediately obtain a rule for the $n$ sample problem by replacing $\alpha$ with $\alpha + \sum_{i=1}^{n} \delta_{X_i}$

- The statistician will want to minimize a loss function defined as

$$L(P, \hat{F}) := \int_{\mathbb{R}} \left(F(x) - \hat{F}(x)\right)^2 dW(x)$$

- The empirical risk for the no-sample problem is

$$\mathbb{E}(L(P, \hat{F})) = \int \mathbb{E}\left(\left(F(x) - \hat{F}(x)\right)^2\right) dW(x)$$

# Application: distribution function estimation II

- The minimizer for the non-sample problem is

$$F_0(x) := \mathbb{E}(F(x)) = \mathbb{E}(P((-\infty, x))) = \frac{\alpha((-\infty, x)}{\alpha(\mathbb{R})}$$

.

- For a sample of size $n$, we replace $\alpha$ with $\alpha + \sum_{i=1}^{n} \delta_{X_i}$ to obtain

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

where $p_n := \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R})+n}$ and $F_n(t) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}((-\infty, t]))$

- By Glivenko Cantelli, the Bayes estimates converges to the true distribution function uniformly.

# Application: mean estimation

- Assume $P \in \mathfrak{D}(\alpha)$, where $\alpha$ has finite first moment. Suppose that the statistician wants to minimize $L(P, \hat{\mu}) = (\mu - \hat{\mu})^2$ where $\mu := \int x dP(x)$ (well defined by theorem 3).

- Denote by $\mu_0 := \frac{\int x d\alpha(x)}{\alpha(\mathbb{R})}$. By theorem 3, the Bayes rule for the no-sample problem is $\hat{\mu} = \mu_0$.

- For a sample of size $n$, we instead obtain the Bayes rule

$$\begin{aligned}\hat{\mu}_n(X_1, ..., X_n) &= \frac{\int x d(\alpha + \sum_{i=1}^n \delta_{X_i})}{\alpha(\mathbb{R}) + n} \\ &= p_n \mu_0 + (1 - p_n)\bar{X}_n\end{aligned}$$

Where $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$.

# Conclusion

- We have defined the Dirichlet Process and pointed out its main properties.
- The Dirichlet process enjoys a large support and is tractable, satisfying our desiderata.
- Downside: the realizations of the process are discrete with probability 1.