The origins of the stick breaking representation for Dirichlet priors

Jayaram Sethuraman
Department of Statistics
Florida State University
Tallahassee, FL 32306

sethu@stat.fsu.edu

December 28, 2017

Figure : Born: April 30, 1924, Delhi, India. Died: June 7, 1997, Chicago,IL

Abstract

- The stick breaking construction of the Dirichlet process has a nearly appeared as early as Ferguson (1973)!

Abstract

- The stick breaking construction of the Dirichlet process has a nearly appeared as early as Ferguson (1973)!
- For a special case, I saw this construction from studying the Blackwell and MacQueen (1973) paper.

Abstract

- The stick breaking construction of the Dirichlet process has a nearly appeared as early as Ferguson (1973)!
- For a special case, I saw this construction from studying the Blackwell and MacQueen (1973) paper.
- It appeared in a published form in Sethuraman (1994),

Abstract

- The stick breaking construction of the Dirichlet process has a nearly appeared as early as Ferguson (1973)!
- For a special case, I saw this construction from studying the Blackwell and MacQueen (1973) paper.
- It appeared in a published form in Sethuraman (1994), where I incorrectly said that it was discovered when I was teaching a seminar course on Dirichlet processes in Spring 1979.

Abstract

- The stick breaking construction of the Dirichlet process has a nearly appeared as early as Ferguson (1973)!
- For a special case, I saw this construction from studying the Blackwell and MacQueen (1973) paper.
- It appeared in a published form in Sethuraman (1994), where I incorrectly said that it was discovered when I was teaching a seminar course on Dirichlet processes in Spring 1979.

  Jim Lynch has jagged my memory and it was in Fall 1978.

What is the Dirichlet process?

The Dirichlet process or Dirichlet is prior is the distribution of a random probability measure P on R_1 which can serve as a prior distribution for the standard nonparametric problem – $X_1, X_2, \ldots, X_n$ are i.i.d. P.

So it is also a probability measure on the space of all probability distributions $\mathcal{P}$ on R_1.

Probability measures

What is a probability measure on the real line $(\mathcal{X}, \mathcal{B})$ with its Borel σ-field?

It is a function P on $\mathcal{B}$ such that

$$P(\mathcal{X}) = 1, \quad 0 \leq P(A) \leq 1 \text{ for all } A \in \mathcal{B}, \text{ and}$$

$$P(\cup_1^\infty A_i) = \sum_1^\infty P(A_i)$$

for each collection of disjoint subsets $\{A_1, A_2, \dots\}$ in $\mathcal{B}$.

Probability measures

What is a probability measure on the real line $(\mathcal{X}, \mathcal{B})$ with its Borel σ-field?

It is a function P on $\mathcal{B}$ such that

$$P(\mathcal{X}) = 1, \;\; 0 \leq P(A) \leq 1 \text{ for all } A \in \mathcal{B}, \text{ and}$$

$$P(\cup_1^\infty A_i) = \sum_1^\infty P(A_i)$$

for each collection of disjoint subsets $\{A_1, A_2, \dots\}$ in $\mathcal{B}$.

Can we verify this for the normal distribution? How many times do we have to verify this? Can a carefully chosen number of countable verifications do?

Examples of probability measures

Let $x_1, x_2, \ldots$ be distinct points in $\mathcal{X}$. Then $P = \delta_{x_1},\ P = \delta_{x_2}, \ldots$ are examples of probability measures (degenerate) and are points in $\mathcal{P}$.

$P = p_1\delta_{x_1} + p_2\delta_{x_2} + \ldots$ is also a (discrete) probability measure, if . . . .

What is the class $\mathcal{P}$ of all probability measures?

Examples of probability measures

Alternatively, we can define probability measures by defining real random variables and looking at their distributions (which will arise from some grand daddy space).
Thus $X(\omega) \equiv x_1, X(\omega) \equiv x_2, \ldots$ have degenerate probability distributions.
How do we define random variables to get other discrete distributions? Random variables require a grand daddy space to begin with.

Random probability measures

For the nonparametric Bayes problem we should be considering measures, Q, which are random probability measures, that is, probability measures Q on $(\mathcal{P}, \mathcal{C})$, the space of probability measures on $(\mathcal{X}, \mathcal{B})$. (Define $\mathcal{C}$ suitably.)

Random probability measures

For the nonparametric Bayes problem we should be considering measures, Q, which are random probability measures, that is, probability measures Q on $(\mathcal{P}, \mathcal{C})$, the space of probability measures on $(\mathcal{X}, \mathcal{B})$. (Define $\mathcal{C}$ suitably.)

Or, we can just consider a random variable (defined on some grand daddy space) $P = P(\omega)$, also called random probability measure, on $(\mathcal{X}, \mathcal{B})$; then its distribution Q will be a nonparametric prior.

Random probability measures

Let $P_1, \ P_2, \ldots$ be probability measures, that is, points in $\mathcal{P}$.
Then $P \equiv \delta_{P_1}, \ P \equiv \delta_{P_2}, \ldots$ are random probability measures (discrete).
Simpler still, $P = \delta_{\delta_{x_1}} = \delta_{x_1}$ (for short) is a discrete random probability measure.
$P = \sum_1^\infty p_i \delta_{x_i}$ is also a discrete random probability measure.
$P = \sum_1^\infty p_i(\omega) \delta_{X_i(\omega)}$ is a random probability measure and its distribution Q is a nonparametric prior.
This will give only a small class of nonparametric priors.
Fortunately, the Dirichlet prior is in this class.
However, what is the class of all nonparametric priors?

Assertions concerning Dirichlet priors

There is a random probability measure P with distribution $\mathcal{D}(\alpha, \beta(\cdot))$ called the Dirichlet prior (process) with parameters α and $\beta(\cdot)$.

Its main properties are

1. Under P, the distribution of $(P(A_1), \ldots, P(A_k))$ is the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$ for measurable partitions $(A_1, \ldots, A_k)$ of R_1.

Assertions concerning Dirichlet priors

There is a random probability measure P with distribution $\mathcal{D}(\alpha, \beta(\cdot))$ called the Dirichlet prior (process) with parameters α and $\beta(\cdot)$.

Its main properties are

1. Under P, the distribution of $(P(A_1), \ldots, P(A_k))$ is the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$ for measurable partitions $(A_1, \ldots, A_k)$ of R_1.
2. The posterior distribution of P given X_1 is $\mathcal{D}((\alpha + 1), \frac{\beta(\cdot)+\delta_{X_1}(\cdot)}{\alpha+1})$.

Assertions concerning Dirichlet priors

There is a random probability measure P with distribution $\mathcal{D}(\alpha, \beta(\cdot))$ called the Dirichlet prior (process) with parameters α and $\beta(\cdot)$.

Its main properties are

1. Under P, the distribution of $(P(A_1), \ldots, P(A_k))$ is the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$ for measurable partitions $(A_1, \ldots, A_k)$ of R_1.
2. The posterior distribution of P given X_1 is $\mathcal{D}((\alpha+1), \frac{\beta(\cdot)+\delta_{X_1}(\cdot)}{\alpha+1})$.
3. The random probability measure P is a discrete probability measure.

Assertions concerning Dirichlet priors

There is a random probability measure P with distribution $\mathcal{D}(\alpha, \beta(\cdot))$ called the Dirichlet prior (process) with parameters α and $\beta(\cdot)$.

Its main properties are

1. Under P, the distribution of $(P(A_1), \ldots, P(A_k))$ is the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$ for measurable partitions $(A_1, \ldots, A_k)$ of R_1.
2. The posterior distribution of P given X_1 is $\mathcal{D}((\alpha+1), \frac{\beta(\cdot)+\delta_{X_1}(\cdot)}{\alpha+1})$.
3. The random probability measure P is a discrete probability measure.

It first appeared in three papers in 1973 - Ferguson, Blackwell, and Blackwell-MacQueen.

Summary

- What is the stick breaking construction?
- Details from Ferguson (1973)
  - First definition of a DP
  - Alternate definition of DP
- As an aside "What about Blackwell (1973)?"
- Details from Blackwell and MacQueen (1973)
  - Nonparametric priors and exchangeable random variables; Pólya urn sequences
  - The stick breaking construction when β is non-atomic
- Sethuraman construction of Dirichlet priors
- Misconceptions about the stick breaking construction
- Some properties of Dirichlet priors

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables.

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

This has been called "stick breaking".

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. *Beta*$(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

This has been called "stick breaking". It was known in the literature much long ago as the "RAM" model or as the model with $V_1, V_2, \dots$ as (discrete) failure rates.

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

This has been called "stick breaking". It was known in the literature much long ago as the "RAM" model or as the model with $V_1, V_2, \dots$ as (discrete) failure rates.

The distribution of the random discrete distribution $\mathbf{p} = (p_1, p_2, \dots)$ is also known as the GEM(α) or GEM($\mathbf{V}$) (Griffith-Engen-McCloskey) distribution.

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

This has been called "stick breaking". It was known in the literature much long ago as the "RAM" model or as the model with $V_1, V_2, \dots$ as (discrete) failure rates.

The distribution of the random discrete distribution $\mathbf{p} = (p_1, p_2, \dots)$ is also known as the GEM(α) or GEM($\mathbf{V}$) (Griffith-Engen-McCloskey) distribution.

The distribution of $(p_1, p_2, \dots, p_n, (1 - p_1 - \dots - p_n))$ is not any simple finite dimensional Dirichlet distribution – its pdf is proportional to

$$\frac{(1 - p_1 - \dots - p_n)^{1-\alpha}}{(1 - p_1)(1 - p_1 - p_2) \dots (1 - p_1 - \dots - p_n)}.$$

The stick breaking construction - I

Let $\mathbf{V} = (V_1, V_2, \dots)$ be i.i.d. $Beta(1, \alpha)$ random variables. Define $p_1 = V_1, p_2 = (1 - V_1)V_2, p_3 = (1 - V_1)(1 - V_2)V_3, \dots$.

This has been called "stick breaking". It was known in the literature much long ago as the "RAM" model or as the model with $V_1, V_2, \dots$ as (discrete) failure rates.

The distribution of the random discrete distribution $\mathbf{p} = (p_1, p_2, \dots)$ is also known as the GEM(α) or GEM($\mathbf{V}$) (Griffith-Engen-McCloskey) distribution.

The distribution of $(p_1, p_2, \dots, p_n, (1 - p_1 - \dots - p_n))$ is not any simple finite dimensional Dirichlet distribution – its pdf is proportional to

$$\frac{(1 - p_1 - \dots - p_n)^{1-\alpha}}{(1 - p_1)(1 - p_1 - p_2) \dots (1 - p_1 - \dots - p_n)}.$$

Connor and Mosimann (1969).

The stick breaking construction - II

Let $\mathbf{Z} = Z_1, Z_2, \ldots$ be i.i.d. $\beta(\cdot)$. For measurable sets A, define

$$P(A) = P(\mathbf{p}, \mathbf{Z})(A) = \sum p_j \; I(Z_j \in A) = \sum p_j \; \delta_{Z_j}(A).$$

The stick breaking construction - II

Let $\mathbf{Z} = Z_1, Z_2, \ldots$ be i.i.d. $\beta(\cdot)$. For measurable sets A, define

$$P(A) = P(\mathbf{p}, \mathbf{Z})(A) = \sum p_j \; I(Z_j \in A) = \sum p_j \; \delta_{Z_j}(A).$$

This is the stick breaking construction of a random probability measure $P(\cdot)$ whose distribution is $\mathcal{D}(\alpha, \beta(\cdot))$.

Ferguson

The Ferguson paper

Ferguson (1973) – I

The Annals of Statistics of 1973, Issue 2 contains the famous paper of Ferguson. It also contains two other famous papers, one by Blackwell and another by Blackwell and MacQueen - all dealing with Dirichlet processes.

Ferguson (1973) – I

The Annals of Statistics of 1973, Issue 2 contains the famous paper of Ferguson. It also contains two other famous papers, one by Blackwell and another by Blackwell and MacQueen - all dealing with Dirichlet processes.

In the first three sections of his paper, Ferguson defined the Dirichlet process $\mathcal{D}(\alpha, \beta(\cdot))$ as the distribution of a random probability measure P for which

$$(P(A_1), \ldots, P(A_k)) \sim \mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$$

for all finite measurable partitions $(A_1, \ldots, A_k)$.

Ferguson (1973) – I

The Annals of Statistics of 1973, Issue 2 contains the famous paper of Ferguson. It also contains two other famous papers, one by Blackwell and another by Blackwell and MacQueen - all dealing with Dirichlet processes.

In the first three sections of his paper, Ferguson defined the Dirichlet process $\mathcal{D}(\alpha, \beta(\cdot))$ as the distribution of a random probability measure P for which

$$(P(A_1), \ldots, P(A_k)) \sim \mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$$

for all finite measurable partitions $(A_1, \ldots, A_k)$.

Do you know such a random probability measure P exists before positing some of its distributional properties as its definition?

Ferguson (1973) – II

Ferguson showed that the posterior distribution given an observation X from P is $\mathcal{D}(\alpha + 1, \frac{\beta(\cdot)+\delta_X(\cdot)+1}{\alpha+1})$.

Ferguson (1973) – II

Ferguson showed that the posterior distribution given an observation X from P is $\mathcal{D}(\alpha + 1, \frac{\beta(\cdot)+\delta_X(\cdot)+1}{\alpha+1})$.

Ferguson used a peculiar definition of what it means to say that X is an observation from P.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

A process $\{X(t), t \in [0,1]\}$ is a Gamma process with parameter α if it has independent increments and the distribution of $X(t)$ is $Gamma(\alpha t)$.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

A process $\{X(t), t \in [0, 1]\}$ is a Gamma process with parameter α if it has independent increments and the distribution of $X(t)$ is $Gamma(\alpha t)$. It will follow that $X(0) = 0$ and $X(1) \sim Gamma(\alpha)$.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

A process $\{X(t), t \in [0, 1]\}$ is a Gamma process with parameter α if it has independent increments and the distribution of $X(t)$ is $Gamma(\alpha t)$. It will follow that $X(0) = 0$ and $X(1) \sim Gamma(\alpha)$.

Let $J_1 \geq J_2 \geq J_3 \cdots$ be the ordered jumps of this Gamma process.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

A process $\{X(t), t \in [0,1]\}$ is a Gamma process with parameter α if it has independent increments and the distribution of $X(t)$ is $Gamma(\alpha t)$. It will follow that $X(0) = 0$ and $X(1) \sim Gamma(\alpha)$.

Let $J_1 \geq J_2 \geq J_3 \cdots$ be the ordered jumps of this Gamma process.

The $J = \sum J_i = X(1)$ is finite and has distribution Gamma(α).
Let $\pi_1 = J_1/J, \pi_2 = J_2/J, \ldots$.

Ferguson (1973) – III

In Section 4 of his paper, Ferguson presents an alternative definition of the DP.

A process $\{X(t), t \in [0,1]\}$ is a Gamma process with parameter α if it has independent increments and the distribution of $X(t)$ is $Gamma(\alpha t)$. It will follow that $X(0) = 0$ and $X(1) \sim Gamma(\alpha)$.

Let $J_1 \geq J_2 \geq J_3 \cdots$ be the ordered jumps of this Gamma process.

The $J = \sum J_i = X(1)$ is finite and has distribution Gamma(α).
Let $\pi_1 = J_1/J, \pi_2 = J_2/J, \ldots$.

Then $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ is a random discrete probability measure and is called the Poisson-Dirichlet distribution.

Ferguson (1973) – IV

Let $\mathbf{W} = (W_1, W_2, \dots)$ be i.i.d. $\beta(\cdot)$ and independent of $\boldsymbol{\pi}$.
For measurable sets A, define

$$P(A) = \sum \pi_j \; I(W_j \in A) = \sum \pi_j \; \delta_{W_j}(A).$$

Ferguson (1973) – IV

Let $\mathbf{W} = (W_1, W_2, \dots)$ be i.i.d. $\beta(\cdot)$ and independent of $\boldsymbol{\pi}$. For measurable sets A, define

$$P(A) = \sum \pi_j \; I(W_j \in A) = \sum \pi_j \; \delta_{W_j}(A).$$

As an aside, note that $P(A)$ will be the same if the terms in this summation are permuted, even if the permutation is random and depends on $\boldsymbol{\pi}$ alone.

Ferguson (1973) – IV

Let $\mathbf{W} = (W_1, W_2, \dots)$ be i.i.d. $\beta(\cdot)$ and independent of $\boldsymbol{\pi}$. For measurable sets A, define

$$P(A) = \sum \pi_j \; I(W_j \in A) = \sum \pi_j \; \delta_{W_j}(A).$$

As an aside, note that $P(A)$ will be the same if the terms in this summation are permuted, even if the permutation is random and depends on $\boldsymbol{\pi}$ alone.

Ferguson showed that this random probability measure P has the DP distribution $\mathcal{D}(\alpha, \beta(\cdot))$.

Ferguson (1973) – IV

Let $\mathbf{W} = (W_1, W_2, \dots)$ be i.i.d. $\beta(\cdot)$ and independent of $\boldsymbol{\pi}$.
For measurable sets A, define

$$P(A) = \sum \pi_j \; I(W_j \in A) = \sum \pi_j \; \delta_{W_j}(A).$$

As an aside, note that $P(A)$ will be the same if the terms in this summation are permuted, even if the permutation is random and depends on $\boldsymbol{\pi}$ alone.

Ferguson showed that this random probability measure P has the DP distribution $\mathcal{D}(\alpha, \beta(\cdot))$.

This looks like the stick breaking definition but the stick is very sticky.

The Blackwell paper

The Blackwell paper

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

It shows that a random probability measure P can be described through a collection of independent r.v.'s $(U_1, U_2, \dots)$ in $[0, 1]$.

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

It shows that a random probability measure P can be described through a collection of independent r.v.'s $(U_1, U_2, \dots)$ in $[0, 1]$.

The ideas of the proof can be used to construct random probability measures that sit on the subset of continuous probability measures.

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

It shows that a random probability measure P can be described through a collection of independent r.v.'s $(U_1, U_2, \dots)$ in $[0, 1]$.

The ideas of the proof can be used to construct random probability measures that sit on the subset of continuous probability measures.

We can state the posterior distribution of $(U_1, U_2, \dots)$, (and thus of P also), given an an observation X.

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

It shows that a random probability measure P can be described through a collection of independent r.v.'s $(U_1, U_2, \dots)$ in $[0, 1]$.

The ideas of the proof can be used to construct random probability measures that sit on the subset of continuous probability measures.

We can state the posterior distribution of $(U_1, U_2, \dots)$, (and thus of P also), given an an observation X.

It does not give any hints for a stick breaking construction.

The Blackwell paper

The beautiful paper of Blackwell, *Discreteness of Ferguson Selections*, in the same 1973 issue of the Annals of Statistics as the classical Ferguson paper defines the Dirichlet process, the posterior distribution and establishes that the corresponding random probability measure is discrete.

It shows that a random probability measure P can be described through a collection of independent r.v.'s $(U_1, U_2, \dots)$ in $[0, 1]$.

The ideas of the proof can be used to construct random probability measures that sit on the subset of continuous probability measures.

We can state the posterior distribution of $(U_1, U_2, \dots)$, (and thus of P also), given an an observation X.

It does not give any hints for a stick breaking construction.

This paper also contains all the ideas of random probability measures using Polyá trees – see Mauldin, Sudderth, Williams (1992).

The Blackwell and MacQueen's paper

Blackwell and MacQueen's paper

This paper gives a definition of the DP in terms of Ployá sequences.

Blackwell and MacQueen's paper

This paper gives a definition of the DP in terms of Ployá sequences.

A Polyá sequence is exchangeable sequence of random variables. These authors re-establish de Finetti's theorem for Polyá sequences in a novel way and give more insights.

Blackwell and MacQueen's paper

This paper gives a definition of the DP in terms of Ployá sequences.

A Polyá sequence is exchangeable sequence of random variables. These authors re-establish de Finetti's theorem for Polyá sequences in a novel way and give more insights.

We will now give an expansive alternate treatment of the results of this paper which will lead us to the stick breaking representation for the case $\beta(\cdot)$ is non-atomic,

Blackwell and MacQueen's paper

This paper gives a definition of the DP in terms of Ployá sequences.

A Polyá sequence is exchangeable sequence of random variables. These authors re-establish de Finetti's theorem for Polyá sequences in a novel way and give more insights.

We will now give an expansive alternate treatment of the results of this paper which will lead us to the stick breaking representation for the case $\beta(\cdot)$ is non-atomic, short of the full stick breaking construction.

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

This follows from an examination of De Finetti's theorem (1931), Blackwell and MacQueen (1973) as explained below. See also Hewitt and Savage (1955), Kingman (1978).

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

This follows from an examination of De Finetti's theorem (1931), Blackwell and MacQueen (1973) as explained below. See also Hewitt and Savage (1955), Kingman (1978).

Let $X_1, X_2, \ldots$ be an infinite sequence of exchangeable (def?) sequence of random variables with a joint distribution Q.

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

This follows from an examination of De Finetti's theorem (1931), Blackwell and MacQueen (1973) as explained below. See also Hewitt and Savage (1955), Kingman (1978).

Let $X_1, X_2, \ldots$ be an infinite sequence of exchangeable (def?) sequence of random variables with a joint distribution Q.

Then, from De Finetti's theorem (or reversed martingale theorem)

1. The empirical distribution functions $F_n(x) \to F(x)$ with probability 1 for all x.

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

This follows from an examination of De Finetti's theorem (1931), Blackwell and MacQueen (1973) as explained below. See also Hewitt and Savage (1955), Kingman (1978).

Let $X_1, X_2, \ldots$ be an infinite sequence of exchangeable (def?) sequence of random variables with a joint distribution Q.

Then, from De Finetti's theorem (or reversed martingale theorem)

1. The empirical distribution functions $F_n(x) \to F(x)$ with probability 1 for all x. In fact, $\sup_x |F_n(x) - F(x)| \to 0$ with probability 1.

Re-reading Blackwell and MacQueen (1973) – I

The class of all nonparametric priors are the same as the class of all exchangeable sequences of random variables!

This follows from an examination of De Finetti's theorem (1931), Blackwell and MacQueen (1973) as explained below. See also Hewitt and Savage (1955), Kingman (1978).

Let $X_1, X_2, \ldots$ be an infinite sequence of exchangeable (def?) sequence of random variables with a joint distribution Q.

Then, from De Finetti's theorem (or reversed martingale theorem)

1. The empirical distribution functions $F_n(x) \to F(x)$ with probability 1 for all x. In fact, $\sup_x |F_n(x) - F(x)| \to 0$ with probability 1.
   (Note that $F(x)$ is a random distribution function.)

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.
3. Given P, $X_1, X_2, \ldots$ are i.i.d. P.

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.
3. Given P, $X_1, X_2, \ldots$ are i.i.d. P.
4. Let us denote the distribution of P under Q by ν^Q. This ν^Q is a nonparametric prior – it is a pm on the space of pm's on R_1.

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.
3. Given P, $X_1, X_2, \ldots$ are i.i.d. P.
4. Let us denote the distribution of P under Q by ν^Q. This ν^Q is a nonparametric prior – it is a pm on the space of pm's on R_1.
5. The class of all nonparametric priors arises in this fashion.

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.
3. Given P, $X_1, X_2, \ldots$ are i.i.d. P.
4. Let us denote the distribution of P under Q by ν^Q. This ν^Q is a nonparametric prior – it is a pm on the space of pm's on R_1.
5. The class of all nonparametric priors arises in this fashion.
6. The distribution of $X_2, X_3, \ldots,$ given X_1 is also exchangeable; denote it by Q_{X_1}.
7. The limit P of the empirical probability measures of $X_1, X_2, \ldots$ is also the limit of the empirical probability measures of $X_2, X_3, \ldots$.

Re-reading Blackwell and MacQueen (1973) – II

2. The empirical probability measures P_n converge to a random probability measure P weakly with probability 1.
3. Given P, $X_1, X_2, \ldots$ are i.i.d. P.
4. Let us denote the distribution of P under Q by ν^Q. This ν^Q is a nonparametric prior – it is a pm on the space of pm's on R_1.
5. The class of all nonparametric priors arises in this fashion.
6. The distribution of $X_2, X_3, \ldots,$ given X_1 is also exchangeable; denote it by Q_{X_1}.
7. The limit P of the empirical probability measures of $X_1, X_2, \ldots$ is also the limit of the empirical probability measures of $X_2, X_3, \ldots$. Thus the distribution of P given X_1 (the posterior distribution) is the distribution of P under Q_{X_1} and, by mere notation, is $\nu^{Q_{X_1}}$.

Dirichlet prior based on a Pólya urn sequences

The Pólya urn sequence is an example of an infinite exchangeable random variables.

Let β be a pm on R_1 and let $\alpha > 0$. Define the joint distribution $Pol(\alpha, \beta)$ of $X_1, X_2, \ldots$ through

$$X_1 \sim \beta(\cdot),\ X_2 | X_1 \sim \frac{\alpha\beta(\cdot) + \delta_{X_1}(\cdot)}{\alpha + 1}$$

$$X_n | (X_1, \ldots, X_{n-1}) \sim \frac{\alpha\beta(\cdot) + \sum_1^{n-1} \delta_{X_i}(\cdot)}{\alpha + n - 1}, n = 3, 4, \ldots$$

This defines $Pol(\alpha, \beta)$ as an exchangeable probability measure. (It takes just some effort to establish this.)

Dirichlet prior based on a Pólya urn sequences

We gave the posterior distribution even before obtaining a full description of the prior.

Dirichlet prior based on a Pólya urn sequences

We gave the posterior distribution even before obtaining a full description of the prior.

Blackwell show that under $\nu^{Pol(\alpha,\beta)}$, the distribution of $(P(A_1), \ldots, P(A_k))$ is $\mathcal{D}(\alpha\beta(A_1), \ldots, \alpha\beta(A_k))$ for any partition $(A_1, \ldots, A_k)$ (by comparing moments).

That is, $\nu^{Pol(\alpha,\beta)} = \mathcal{D}(\alpha, \beta(\cdot))$.

Dirichlet prior based on a Pólya urn sequences

We gave the posterior distribution even before obtaining a full description of the prior.

Blackwell show that under $\nu^{Pol(\alpha,\beta)}$, the distribution of $(P(A_1),\ldots,P(A_k))$ is $\mathcal{D}(\alpha\beta(A_1),\ldots,\alpha\beta(A_k))$ for any partition $(A_1,\ldots,A_k)$ (by comparing moments).

That is, $\nu^{Pol(\alpha,\beta)} = \mathcal{D}(\alpha,\beta(\cdot))$.
In particular, for any A, $P(A) \sim Beta(\alpha\beta(A),\alpha\beta(A^c))$.

Dirichlet prior based on a Pólya urn sequences

We gave the posterior distribution even before obtaining a full description of the prior.

Blackwell show that under $\nu^{Pol(\alpha,\beta)}$, the distribution of $(P(A_1),\ldots,P(A_k))$ is $\mathcal{D}(\alpha\beta(A_1),\ldots,\alpha\beta(A_k))$ for any partition $(A_1,\ldots,A_k)$ (by comparing moments).

That is, $\nu^{Pol(\alpha,\beta)} = \mathcal{D}(\alpha,\beta(\cdot))$.
In particular, for any A, $P(A) \sim Beta(\alpha\beta(A),\alpha\beta(A^c))$. Can we allow $A=\{X_1\}$ in the above?

Dirichlet prior based on a Pólya urn sequences

- The conditional distribution of $(X_2, X_3, \dots)$ given X_1 is $Pol(\alpha + 1, \frac{\alpha\beta + \delta_{X_1}}{\alpha+1})$.

Dirichlet prior based on a Pólya urn sequences

- The conditional distribution of $(X_2, X_3, \dots)$ given X_1 is $Pol(\alpha + 1, \frac{\alpha\beta + \delta_{X_1}}{\alpha+1})$.
- Thus posterior distribution of P given X_1 is $\nu^{Pol(\alpha+1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})}$

Dirichlet prior based on a Pólya urn sequences

- The conditional distribution of $(X_2, X_3, \dots)$ given X_1 is $Pol(\alpha+1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})$.
- Thus posterior distribution of P given X_1 is $\nu^{Pol(\alpha+1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})}$which is equal to $\mathcal{D}(\alpha+1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

Dirichlet prior based on a Pólya urn sequences

- The conditional distribution of $(X_2, X_3, \dots)$ given X_1 is $Pol(\alpha + 1, \frac{\alpha\beta + \delta_{X_1}}{\alpha+1})$.
- Thus posterior distribution of P given X_1 is $\nu^{Pol(\alpha+1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})}$which is equal to $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.
- Though each P_n is a discrete rpm and the limit P in general will be just a rpm.

Dirichlet prior based on a Pólya urn sequences

- The conditional distribution of $(X_2, X_3, \dots)$ given X_1 is $Pol(\alpha + 1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})$.
- Thus posterior distribution of P given X_1 is $\nu^{Pol(\alpha+1, \frac{\alpha\beta+\delta_{X_1}}{\alpha+1})}$ which is equal to $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.
- Though each P_n is a discrete rpm and the limit P in general will be just a rpm.
- For the present case of a Pólya urn sequence, Blackwell and MacQueen (1973) show that $P(\{X_1, \dots, X_n\}) \to 1$ with probability 1 and thus P is a discrete rpm. (A little tricky. We will show some details.)

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

The conditional distribution of $P(\{X_1\})$ given X_1 is

$$B(\alpha\beta(\{X_1\}) + 1), \alpha\beta(R_1 \setminus \{X_1\})).$$

This is tricky. Is $P(\{X_1\})$ measurable to begin with?

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

The conditional distribution of $P(\{X_1\})$ given X_1 is

$$B(\alpha\beta(\{X_1\}) + 1), \alpha\beta(R_1 \setminus \{X_1\})).$$

This is tricky. Is $P(\{X_1\})$ measurable to begin with?

The conditional distribution of $P(\{X_1, \ldots, X_n\})$ given $(X_1, \ldots, X_n)$ is $Beta(\alpha\beta(\{X_1, \ldots, X_n\}) + n, \alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\}))$
and
$E(P(\{X_1, \ldots, X_n\}^c | X_1, \ldots, X_n)) = \frac{\alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\})}{\alpha+n}$

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha+1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

The conditional distribution of $P(\{X_1\})$ given X_1 is

$$B(\alpha\beta(\{X_1\})+1), \alpha\beta(R_1 \setminus \{X_1\})).$$

This is tricky. Is $P(\{X_1\})$ measurable to begin with?

The conditional distribution of $P(\{X_1, \ldots, X_n\})$ given $(X_1, \ldots, X_n)$ is $Beta(\alpha\beta(\{X_1, \ldots, X_n\}) + n, \alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\}))$
and
$E(P(\{X_1, \ldots, X_n\}^c | X_1, \ldots, X_n)) = \frac{\alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\})}{\alpha+n} \le \frac{\alpha}{\alpha+n}$

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha+1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

The conditional distribution of $P(\{X_1\})$ given X_1 is

$$B(\alpha\beta(\{X_1\})+1), \alpha\beta(R_1 \setminus \{X_1\})).$$

This is tricky. Is $P(\{X_1\})$ measurable to begin with?

The conditional distribution of $P(\{X_1,\ldots,X_n\})$ given $(X_1,\ldots,X_n)$ is $Beta(\alpha\beta(\{X_1,\ldots,X_n\})+n, \alpha\beta(R_1 \setminus \{X_1,\ldots,X_n\}))$
and
$E(P(\{X_1,\ldots,X_n\}^c|X_1,\ldots,X_n)) = \frac{\alpha\beta(R_1\setminus\{X_1,\ldots,X_n\})}{\alpha+n} \le \frac{\alpha}{\alpha+n} \to 0.$

Dirichlet prior based on a Pólya urn sequences

The conditional distribution of P given X_1 is $\mathcal{D}(\alpha + 1, \frac{\beta+\delta_{X_1}}{\alpha+1})$.

The conditional distribution of $P(\{X_1\})$ given X_1 is

$$B(\alpha\beta(\{X_1\}) + 1), \alpha\beta(R_1 \setminus \{X_1\})).$$

This is tricky. Is $P(\{X_1\})$ measurable to begin with?

The conditional distribution of $P(\{X_1, \ldots, X_n\})$ given $(X_1, \ldots, X_n)$ is $Beta(\alpha\beta(\{X_1, \ldots, X_n\}) + n, \alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\}))$
and
$E(P(\{X_1, \ldots, X_n\}^c | X_1, \ldots, X_n)) = \frac{\alpha\beta(R_1 \setminus \{X_1, \ldots, X_n\})}{\alpha+n} \leq \frac{\alpha}{\alpha+n} \to 0.$
This means that P is a discrete random probability measure.

Dirichlet prior based on a Pólya urn sequences

This already gives a sticky stick representation. The random probability measure P is discrete and sits on $\{X_1, X_2 \dots\} = \{Y_1, Y_2, \dots\}$ where $Y_1, Y_2, \dots$ are the distinct observations.

Thus

$$P(A) = \sum_1^\infty P(\{Y_i\})\delta_{Y_i}(A).$$

However, we do not know the joint distribution of $(P(\{Y_1\}), Y_1, \dots)$.

Dirichlet prior based on a Pólya urn sequences

Let the probability masses of the random probability measure P be $\pi_1, \pi_2, \ldots$ written in some order.

Dirichlet prior based on a Pólya urn sequences

Let the probability masses of the random probability measure P be $\pi_1, \pi_2, \ldots$ written in some order.

Given P, the probability mass $P(\{X_1\}) = P(\{Y_1\})$ arises by picking an r with probability π_r and setting $P(\{Y_1\}) = \pi_r$.

Dirichlet prior based on a Pólya urn sequences

Let the probability masses of the random probability measure P be $\pi_1, \pi_2, \dots$ written in some order.

Given P, the probability mass $P(\{X_1\}) = P(\{Y_1\})$ arises by picking an r with probability π_r and setting $P(\{Y_1\}) = \pi_r$.

Similarly, $P(\{Y_2\})$ arises by picking an $s \neq r$ with probability $\frac{\pi_s}{(1-\pi_r)}$ and setting $P(\{Y_2\}) = \pi_s$ and so on.

Dirichlet prior based on a Pólya urn sequences

Let the probability masses of the random probability measure P be $\pi_1, \pi_2, \ldots$ written in some order.

Given P, the probability mass $P(\{X_1\}) = P(\{Y_1\})$ arises by picking an r with probability π_r and setting $P(\{Y_1\}) = \pi_r$.

Similarly, $P(\{Y_2\})$ arises by picking an $s \neq r$ with probability $\frac{\pi_s}{(1-\pi_r)}$ and setting $P(\{Y_2\}) = \pi_s$ and so on.

That is $(P(\{Y_1\}), P(\{Y_2\}), \ldots)$ is a size biased permutation of $(\pi_1, \pi_2 \ldots)$, and hence, is invariant under size biased permutation.

Dirichlet prior based on a Pólya urn sequences

From now on, assume that β is non-atomic.

Dirichlet prior based on a Pólya urn sequences

From now on, assume that β is non-atomic.

The conditional distribution of $P(\{X_1\})$ given X_1 is $B(\alpha\beta(\{X_1\})+1), \alpha\beta(R_1 \setminus \{X_1\}) = B(1, \alpha)$ and does not depend on X_1

Dirichlet prior based on a Pólya urn sequences

From now on, assume that β is non-atomic.

The conditional distribution of $P(\{X_1\})$ given X_1 is $B(\alpha\beta(\{X_1\})+1), \alpha\beta(R_1 \setminus \{X_1\}) = B(1,\alpha)$ and does not depend on X_1 and thus X_1 and $P(\{X_1\})$ are independent.

Dirichlet prior based on a Pólya urn sequences

From now on, assume that β is non-atomic.

The conditional distribution of $P(\{X_1\})$ given X_1 is $B(\alpha\beta(\{X_1\})+1), \alpha\beta(R_1 \setminus \{X_1\}) = B(1, \alpha)$ and does not depend on X_1 and thus X_1 and $P(\{X_1\})$ are independent.

The distribution of X_1 is β from the definition of the Polya sequence.

Dirichlet prior based on a Pólya urn sequences

Let $Y_1, Y_2, \dots$ be the distinct values among $X_1, X_2, \dots$ listed in the order of their appearance.

Then $Y_1 = X_1$,

$Y_1, P(\{Y_1\})$ are independent

Dirichlet prior based on a Pólya urn sequences

Let $Y_1, Y_2, \ldots$ be the distinct values among $X_1, X_2, \ldots$ listed in the order of their appearance.

Then $Y_1 = X_1$,

$Y_1, P(\{Y_1\})$ are independent and $Y_1 \sim \beta, P(\{Y_1\}) \sim B(1, \alpha)$.

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \ldots$ and remove all occurrences of X_1 which is the same as Y_1.

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \ldots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1.

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \dots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1. Its first element is Y_2.

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \ldots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1. Its first element is Y_2.

As before, Y_2 and $\frac{P(\{Y_2\})}{1-P(\{Y_1\})}$ are independent, $Y_2 \sim \beta, \frac{P(\{Y_2\})}{1-P(\{Y_1\})} \sim B(1, \alpha)$.

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \dots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1. Its first element is Y_2.

As before, Y_2 and $\frac{P(\{Y_2\})}{1-P(\{Y_1\})}$ are independent,
$Y_2 \sim \beta, \frac{P(\{Y_2\})}{1-P(\{Y_1\})} \sim B(1, \alpha)$.

Thus $P(\{Y_1\}), \frac{P(\{Y_2\})}{1-P(\{Y_1\})}, \frac{P(\{Y_3\})}{1-P(\{Y_1\})-P(\{Y_2\})}, \dots$ are i.i.d. $B(1, \alpha)$, i.e. GEM(α)

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \dots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1. Its first element is Y_2.

As before, Y_2 and $\frac{P(\{Y_2\})}{1-P(\{Y_1\})}$ are independent, $Y_2 \sim \beta$, $\frac{P(\{Y_2\})}{1-P(\{Y_1\})} \sim B(1, \alpha)$.

Thus $P(\{Y_1\}), \frac{P(\{Y_2\})}{1-P(\{Y_1\})}, \frac{P(\{Y_3\})}{1-P(\{Y_1\})-P(\{Y_2\})}, \dots$ are i.i.d. $B(1, \alpha)$, i.e. GEM(α) (i.e. stick breaking)

Dirichlet prior based on a Pólya urn sequences

Consider the sequence $X_2, X_3, \dots$ and remove all occurrences of X_1 which is the same as Y_1. This reduced sequence is the Pólya urn sequence $Pol(\alpha, \beta)$ and independent of Y_1. Its first element is Y_2.

As before, Y_2 and $\frac{P(\{Y_2\})}{1-P(\{Y_1\})}$ are independent,
$Y_2 \sim \beta$, $\frac{P(\{Y_2\})}{1-P(\{Y_1\})} \sim B(1, \alpha)$.

Thus $P(\{Y_1\}), \frac{P(\{Y_2\})}{1-P(\{Y_1\})}, \frac{P(\{Y_3\})}{1-P(\{Y_1\})-P(\{Y_2\})}, \dots$ are i.i.d. $B(1, \alpha)$, i.e. GEM(α) (i.e. stick breaking)

and all these are independent of $Y_1, Y_2, Y_3 \dots$ which are i.i.d. β.

Dirichlet prior based on a Pólya urn sequences

We already saw that $P = \sum_1^\infty P(\{Y_i\})\delta_{Y_1}$.
Put $p_i = P(Y_i), i = 1, 2, \ldots$. Then $P = \sum_1^\infty p_i \delta_{Y_i}$; i.e. we have the Sethuraman stick breaking construction of the Dirichlet prior (if β is non-atomic).

Dirichlet prior based on a Pólya urn sequences

We already saw that $P = \sum_1^\infty P(\{Y_i\})\delta_{Y_1}$.
Put $p_i = P(Y_i), i = 1, 2, \ldots$. Then $P = \sum_1^\infty p_i \delta_{Y_i}$; i.e. we have the Sethuraman stick breaking construction of the Dirichlet prior (if β is non-atomic).

This is how we can turn around the article by Blackwell and MacQueen (1973) to obtain the stick breaking result when β is non-atomic.

Dirichlet prior based on a Pólya urn sequences

We already saw that $P = \sum_1^\infty P(\{Y_i\})\delta_{Y_1}$.
Put $p_i = P(Y_i), i = 1, 2, \ldots$. Then $P = \sum_1^\infty p_i \delta_{Y_i}$; i.e. we have the Sethuraman stick breaking construction of the Dirichlet prior (if β is non-atomic).

This is how we can turn around the article by Blackwell and MacQueen (1973) to obtain the stick breaking result when β is non-atomic.

Note that the statement of the stick breaking construction does not to specify any properties of β!

Sethuraman construction of Dirichlet priors

Sethuraman (1994)

Sethuraman construction of Dirichlet priors

Let $\alpha > 0$ and let $\beta(\cdot)$ be a pm on $\mathcal{X}$.

We do not assume that β is non-atomic. Further more, restrictions like $\mathcal{X} = R_1$ do not have to made.

Let $V_1, V_2, \ldots,$ be i.i.d. $B(1, \alpha)$ and let $Z_1, Z_2, \ldots$ be independent of $V_1, V_2, \ldots$ and be i.i.d. $\beta(\cdot)$ and let $\mathbf{p}$ =GEM($\mathbf{V}$).

Sethuraman construction of Dirichlet priors

The stick breaking construction is

$$P(\cdot) = P(\mathbf{p}, \mathbf{Z})(\cdot) = \sum_{1}^{\infty} p_i \delta_{Z_i}(\cdot)$$

Sethuraman construction of Dirichlet priors

The stick breaking construction is

$$P(\cdot) = P(\mathbf{p}, \mathbf{Z})(\cdot) = \sum_{1}^{\infty} p_i \delta_{Z_i}(\cdot)$$

It is clearly a discrete random probability measure.

Sethuraman construction of Dirichlet priors

The stick breaking construction is

$$P(\cdot) = P(\mathbf{p}, \mathbf{Z})(\cdot) = \sum_1^\infty p_i \delta_{Z_i}(\cdot)$$

It is clearly a discrete random probability measure.
We have the canonical identity

$$P = p_1\delta_{Z_1} + (1-p_1)\sum_2^\infty \frac{p_i}{1-p_1}\delta_{Z_i} = p_1\delta_{Z_1} + (1-p_1)P(\mathbf{p}^{-1}/(1-p_1), \mathbf{Z}^{-1})$$

where $\mathbf{p}^{-1}, \mathbf{Z}^{-1}$ have the obvious meanings.

Sethuraman construction of Dirichlet priors

The stick breaking construction is

$$P(\cdot) = P(\mathbf{p}, \mathbf{Z})(\cdot) = \sum_{1}^{\infty} p_i \delta_{Z_i}(\cdot)$$

It is clearly a discrete random probability measure.
We have the canonical identity

$$P = p_1\delta_{Z_1} + (1-p_1)\sum_{2}^{\infty} \frac{p_i}{1-p_1}\delta_{Z_i} = p_1\delta_{Z_1} + (1-p_1)P(\mathbf{p}^{-1}/(1-p_1), \mathbf{Z}^{-1})$$

where $\mathbf{p}^{-1}, \mathbf{Z}^{-1}$ have the obvious meanings.
The canonical identity shows that

$$P = p_1\delta_{Z_1} + (1-p_1)P^*$$

where all the random variables are independent,
$p_1 \sim B(1,\alpha), Z_1 \sim \beta$ and the two rpm's P, P^* have the same distribution.

Sethuraman construction of Dirichlet priors

That is, we have a distributional equation for the distribution of P:

$$P \stackrel{d}{=} p_1 \delta_{Z_1} + (1 - p_1) P.$$

In Sethuraman (1994) we show that $\mathcal{D}(\alpha\beta)$ is a solution to this equation, and also that, if there is a solution then it is unique.

Sethuraman construction of Dirichlet priors

That is, we have a distributional equation for the distribution of P:

$$P \stackrel{d}{=} p_1 \delta_{Z_1} + (1 - p_1) P.$$

In Sethuraman (1994) we show that $\mathcal{D}(\alpha\beta)$ is a solution to this equation, and also that, if there is a solution then it is unique.

In the canonical identity, we could have split with index R, (even a random index R) instead of the index 1.

Sethuraman construction of Dirichlet priors

That is, we have a distributional equation for the distribution of P:

$$P \stackrel{d}{=} p_1 \delta_{Z_1} + (1 - p_1) P.$$

In Sethuraman (1994) we show that $\mathcal{D}(\alpha\beta)$ is a solution to this equation, and also that, if there is a solution then it is unique.

In the canonical identity, we could have split with index R, (even a random index R) instead of the index 1.

We will use this to obtain the posterior distribution.

Sethuraman construction of Dirichlet priors

What about the posterior distribution?

Sethuraman construction of Dirichlet priors

What about the posterior distribution?

Let R be a random variable such $Q(R = r|\mathbf{p}) = p_r, r = 1, 2, \ldots$ and let $Y = Z_R$. Then

$$\begin{aligned} Q(Y \in A|P) &= Q(Y \in A|(\mathbf{p}, \mathbf{Z})) \\ &= \sum_r Q(Y \in A, R = r|(\mathbf{p}, \mathbf{Z})) \\ &= \sum_r Q(Z_r \in A) p_r = P(A) \end{aligned}$$

Thus Y is a like an observation from P and we need the distribution of P given Y.

Sethuraman construction of Dirichlet priors

The canonical identity gives

$$\begin{aligned} P &= p_R\delta_Y + (1-p_R)P(\mathbf{p}^{-R}/(1-p_R), \mathbf{Z}^{-R}) \\ &= p_R\delta_Y + (1-p_R)P^* \end{aligned}$$

Sethuraman construction of Dirichlet priors

The canonical identity gives

$$\begin{aligned} P &= p_R\delta_Y + (1-p_R)P(\mathbf{p}^{-R}/(1-p_R), \mathbf{Z}^{-R}) \\ &= p_R\delta_Y + (1-p_R)P^* \end{aligned}$$

where the conditional distribution of P^* given (R, Y) is $\mathcal{D}(\alpha\beta)$.

Sethuraman construction of Dirichlet priors

The canonical identity gives

$$\begin{aligned} P &= p_R\delta_Y + (1-p_R)P(\mathbf{p}^{-R}/(1-p_R), \mathbf{Z}^{-R}) \\ &= p_R\delta_Y + (1-p_R)P^* \end{aligned}$$

where the conditional distribution of P^* given (R, Y) is $\mathcal{D}(\alpha\beta)$.Conditional on Y, the distribution of P is that of

$$p_R\delta_Y + (1-p_R)P^*$$

which is $\mathcal{D}(\alpha+1, \frac{\alpha\beta+\delta_Y}{\alpha+1})$, from standard identities of Dirichlet distributions.

Misconceptions on the stick breaking construction

It is amply clear that Sethuraman (1994) did not impose any conditions on the base measure $\beta(\cdot)$ that it should be non-atomic.

Many papers continue to assert that Sethuraman (1994) assumes that $\beta(\cdot)$ should be non-atomic.

Misconceptions on the stick breaking construction

It is amply clear that Sethuraman (1994) did not impose any conditions on the base measure $\beta(\cdot)$ that it should be non-atomic.

Many papers continue to assert that Sethuraman (1994) assumes that $\beta(\cdot)$ should be non-atomic.

Paisley (2010) says "We use a little-known property of the constructive definition in (Sethuraman, 1994)" following my personal assurance to him that he can use the stick breaking construction to generate the $Beta(a, b)$ distribution.

Misconceptions on the stick breaking construction

It is amply clear that Sethuraman (1994) did not impose any conditions on the base measure $\beta(\cdot)$ that it should be non-atomic.

Many papers continue to assert that Sethuraman (1994) assumes that $\beta(\cdot)$ should be non-atomic.

Paisley (2010) says "We use a little-known property of the constructive definition in (Sethuraman, 1994)" following my personal assurance to him that he can use the stick breaking construction to generate the $Beta(a, b)$ distribution.

Let $Z_1, Z_2, \ldots$ be i.i.d. with $Q(Z_1 = 1) = 1 - Q(Z_1 = 0) = \frac{a}{a+b}$ and $(p_1, p_2, \ldots)$ be GEM$(a + b)$.

Misconceptions on the stick breaking construction

It is amply clear that Sethuraman (1994) did not impose any conditions on the base measure $\beta(\cdot)$ that it should be non-atomic.

Many papers continue to assert that Sethuraman (1994) assumes that $\beta(\cdot)$ should be non-atomic.

Paisley (2010) says "We use a little-known property of the constructive definition in (Sethuraman, 1994)" following my personal assurance to him that he can use the stick breaking construction to generate the $Beta(a, b)$ distribution.

Let $Z_1, Z_2, \ldots$ be i.i.d. with $Q(Z_1 = 1) = 1 - Q(Z_1 = 0) = \frac{a}{a+b}$ and $(p_1, p_2, \ldots)$ be $\mathrm{GEM}(a + b)$.

$$P = \sum p_i \; I(Z_1 = 1) \sim Beta(a, b)$$

Misconceptions on the stick breaking construction

Ferguson showed that the support of the $\mathcal{D}(\alpha\beta)$ is the collection of probability measures in $\mathcal{P}$ whose support is contained in the support of β.

If the support of β is R_1 then the support of $\mathcal{D}_{\alpha\beta}$ is $\mathcal{P}$.

Misconceptions on the stick breaking construction

Ferguson showed that the support of the $\mathcal{D}(\alpha\beta)$ is the collection of probability measures in $\mathcal{P}$ whose support is contained in the support of β.

If the support of β is R_1 then the support of $\mathcal{D}_{\alpha\beta}$ is $\mathcal{P}$.

We already saw that $\mathcal{D}(\alpha\beta)$ gives probability 1 to the class of discrete pm's.

Misconceptions on the stick breaking construction

Ferguson showed that the support of the $\mathcal{D}(\alpha\beta)$ is the collection of probability measures in $\mathcal{P}$ whose support is contained in the support of β.

If the support of β is R_1 then the support of $\mathcal{D}_{\alpha\beta}$ is $\mathcal{P}$.

We already saw that $\mathcal{D}(\alpha\beta)$ gives probability 1 to the class of discrete pm's.

$\mathcal{D}(\alpha\beta)$ is not itself a discrete probability measure.

Some properties of Dirichlet priors

A simple problem is the estimation of the "true mean", i.e. $\int xdP(x)$ from data $X_1, X_2, \ldots, X_n$ which are i.i.d. P.

In the Bayesian nonparametric problem, P has a prior distribution $\mathcal{D}(\alpha\beta)$ and given P, the data $X_1, \ldots, X_n$ are i.i.d. P.

The Bayesian estimate (under squared error loss function) of $\int xdP(x)$ is its mean under the posterior distribution, which is

$$\frac{\alpha \int xd\beta(x) + n\bar{X}_n}{\alpha + n}.$$

Some properties of Dirichlet priors

A simple problem is the estimation of the "true mean", i.e. $\int x dP(x)$ from data $X_1, X_2, \ldots, X_n$ which are i.i.d. P.

In the Bayesian nonparametric problem, P has a prior distribution $\mathcal{D}(\alpha\beta)$ and given P, the data $X_1, \ldots, X_n$ are i.i.d. P.

The Bayesian estimate (under squared error loss function) of $\int x dP(x)$ is its mean under the posterior distribution, which is

$$\frac{\alpha \int x d\beta(x) + n\bar{X}_n}{\alpha + n}.$$

For this we need to assume that $\int |x| d\beta(x) < \infty$

Some properties of Dirichlet priors

A simple problem is the estimation of the "true mean", i.e. $\int x dP(x)$ from data $X_1, X_2, \ldots, X_n$ which are i.i.d. P.

In the Bayesian nonparametric problem, P has a prior distribution $\mathcal{D}(\alpha\beta)$ and given P, the data $X_1, \ldots, X_n$ are i.i.d. P.

The Bayesian estimate (under squared error loss function) of $\int x dP(x)$ is its mean under the posterior distribution, which is

$$\frac{\alpha \int x d\beta(x) + n\bar{X}_n}{\alpha + n}.$$

For this we need to assume that $\int |x| d\beta(x) < \infty$ and $\int x^2 d\beta(x) < \infty$.

Some properties of Dirichlet priors

However $\int x dP(x)$ may be a well defined even when $\int |x| d\beta(x) = \infty$!

Some properties of Dirichlet priors

However $\int x dP(x)$ may be a well defined even when $\int |x| d\beta(x) = \infty$!

Feigin and Tweedie (1989), and others later, gave necessary and sufficient conditions for $\int x dP(x)$ may be a well defined, namely $\int \log(1 + |x|)) d\beta(x) < \infty$.

From our constructive definition,

$$\int |x| dP(x) = \sum_{1}^{\infty} p_1 |Z_i|.$$

The Kolmogorov three series theorem gives a simple direct proof of this result. Sethuraman (2010).

Some properties of Dirichlet priors

The actual distribution of $\int x dP(x)$ under $\mathcal{D}(\alpha\beta)$ is a vexing problem. Regazzini, Lijoi and Prünster (2003), Lijoi and Prünster (2009) have the best results.

When β is the Cauchy distribution, it is easy from the constructive definition that

$$\int x dP(x) = \sum_{1}^{\infty} p_i Z_i$$

where $Z_1, Z_2, \ldots$ are i.i.d. Cauchy, and hence $\int x Pd(x)$ is Cauchy. One does not need the GEM property of $(p_1, p_2, \ldots)$ for this; it is enough for it to be independent of $(Z_1, Z_2, \ldots)$. Yamato (1984) was the first to prove this.

Some properties of Dirichlet priors

The constructive definition

$$P(\cdot) = \sum_{1}^{\infty} p_i \delta_{Z_i}(\cdot)$$

leads to the inequality

$$||P - \sum_{1}^{M} p_i \delta_{Z_i}|| \leq \prod_{1}^{M} (1 - p_i).$$

So one can allow for several kinds of random stopping to stay within chosen errors. One can also stop at nonrandom times and have probability bounds for errors. Mulliere and Tardella (1998) has several results of this type.

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Instead of the posterior distribution of P given X, we could consider the posterior distribution of $(\mathbf{V}, \mathbf{Z})$ given X.

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Instead of the posterior distribution of P given X, we could consider the posterior distribution of $(\mathbf{V}, \mathbf{Z})$ given X.

This posterior distribution of P turns out to be another stick breaking version

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Instead of the posterior distribution of P given X, we could consider the posterior distribution of $(\mathbf{V}, \mathbf{Z})$ given X.

This posterior distribution of P turns out to be another stick breaking version where $\mathbf{V}$ and $\mathbf{Z}$ with $(V_1, V_2, \dots)$ independent and $(Z_1, Z_2, \dots)$ independent;

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Instead of the posterior distribution of P given X, we could consider the posterior distribution of $(\mathbf{V}, \mathbf{Z})$ given X.

This posterior distribution of P turns out to be another stick breaking version where $\mathbf{V}$ and $\mathbf{Z}$ with $(V_1, V_2, \dots)$ independent and $(Z_1, Z_2, \dots)$ independent; but not i.i.d.

Some properties of Dirichlet priors

The stick breaking construction of the random probability measure P is replaced by to sequences of r.v.'s $\mathbf{V}$ and $\mathbf{Z}$.

Instead of the posterior distribution of P given X, we could consider the posterior distribution of $(\mathbf{V}, \mathbf{Z})$ given X.

This posterior distribution of P turns out to be another stick breaking version where $\mathbf{V}$ and $\mathbf{Z}$ with $(V_1, V_2, \dots)$ independent and $(Z_1, Z_2, \dots)$ independent; but not i.i.d.

This is the main virtue of the stick breaking construction.

Some properties of Dirichlet priors

Current Bayes applications use the Dirichlet prior not for the distribution F of the observed random variables but for the distribution of latent variables that are used to model F. This leads to a host of applications in very diverse fields.

THANK YOU