Research & Meeting Note

Bayesian Nonparametric Mixture Model for Multistate Processes

Supervisor: Liqun Diao

Latex by Justin Li

Contents

1. Bayesian Statistics Fundamental

1.1 Bayesian Statistics Fundamental

Definition 1.1.1 — Bayes' Rule.
The starting point for Bayesian inference is **Bayes' Rule**, the simplest form is

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\overline{A})P(B \mid \overline{A})}$$

where $P(B) > 0$

Posterior Probablity: $P(\text{A before B is known to have occurred}) = P(A \mid B) = P(A) \cdot \dfrac{P(B \mid A)}{P(B)}$

Prior Probablity: $P(\text{A after B is known to have occurred}) = P(A \mid B)$

More generally, if we have a sequence of events $A_1, ..., A_k$ form a partition of A such that with $B \subseteq A$, then for all $i = 1, 2, .., k$:

$$P(A_i \mid B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j)P(B \mid A_j)}$$

■ **Example 1.1**
Medical Testing: Let A be the event that the person has disease and B be the event that has test positive. Assume the test for disease is 90% accurate and $P(A) = 0.01$. Find the probability that the person actually has the disease given that person has test positive.

Note that we want to find $P(A \mid B)$, first we have

$$P(B \mid A) = 0.9 \quad \textbf{and} \quad P(B \mid \overline{A}) = 0.1$$

then

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\overline{A})P(B \mid \overline{A})} = \frac{0.01 \cdot 0.9}{0.01 \cdot 0.9 + (1 - 0.01) \cdot 0.1} \approx 0.08333$$

■

Definition 1.1.2 — Bayes Factor.
Suppose that on the basis of an observed event D (standing for data), wish to test the **null hypothesis** $H_0 : E_0$ versus an alternative hypothesis $H_1 : E_1$, where E_0 and E_1 are two events (not necessarily mutually exclusive or even exhaustive of the event space), then we define

$\pi_0 = P(E_0) =$ the prior probability of the null hypothesis
$\pi_1 = P(E_1) =$ the prior probability of the alternative hypothesis
$\mathbf{PRO} = \dfrac{\pi_0}{\pi_1} =$ the prior odds in favour of the null hypothesis
$p_0 = P(E_0 \mid D) =$ the posterior (data is given) probability of the null hypothesis
$p_1 = P(E_1 \mid D) =$ the posterior (data is given) probability of the alternative hypothesis
$\mathbf{POO} = \dfrac{p_0}{p_1} =$ the posterior odds in favour of the null hypothesis

The **Bayes Factor** is defined as

$$\mathbf{BF} = \frac{POO}{PRO} = \frac{p_0\pi_1}{p_1\pi_0} = P(D) \cdot \frac{p_0}{\pi_0} \cdot \frac{1}{P(D)} \frac{\pi_1}{p_1} = \left[P(D) \cdot \frac{P(E_0 \mid D)}{P(E_0)} \right] \cdot \left[\frac{1}{P(D)} \frac{P(E_1)}{P(E_1 \mid D)} \right] = \frac{P(D \mid E_0)}{P(D \mid E_1)}$$

If $\mathbf{BF} > 1$, then data has increased the relative likelihood of the null; if $\mathbf{BF} < 1$, the data has decreased the relative likelihood of the null. The magnitude of **BF** tells us how much effect the data has had on relative likelihood.

Definition 1.1.3 — Bayesian Models.
A **Baysian Model** has the following basic components:

y : data
θ : model parameter
$f(y \mid \theta)$ or $F(y \mid \theta)$: model distribution
$f(\theta)$: prior distribution

Definition 1.1.4 — Prior and Posterior Distribution.
The posterior distribution of θ has pdf:

$$f(\theta \mid y) = \frac{f(\theta)f(y \mid \theta)}{f(y)}$$

and the prior (unconditional) distribution of y is:

$$f(y) = \int f(y \mid \theta) dF(\theta) = \begin{cases} \int f(\theta) f(y \mid \theta) d\theta & \text{if } \theta \text{ is continuous} \\ \sum_{\theta} f(\theta) f(y \mid \theta) & \text{if } \theta \text{ is discrete} \end{cases}$$

■ Example 1.2
Consider the biased dices, A has 0.1 probability of coming up 6, B, C have 0.2 probability of coming up 6 and D, E, F have 0.3 probability of coming up 6. Now given a die rolled twice, and both have 6 comes up. What is the posterior probability distribution of θ, the probability of 6 comes up on the given die.

Note that

$$f(\theta) = \begin{cases} \frac{1}{6} & \text{If } \theta = 0.1 \\ \frac{1}{3} & \text{If } \theta = 0.2 \\ \frac{1}{2} & \text{If } \theta = 0.3 \end{cases} \qquad \text{and} \qquad (y \mid \theta) \sim Bin(2, \theta)$$

then

$$f(y = 2 \mid \theta) = \binom{2}{y} \theta^y (1-\theta)^{2-y} = \theta^2 \qquad \text{and} \qquad f(y) = \sum_{\theta} f(\theta) f(y \mid \theta) = 0.06$$

Then the posterior probability distribution of θ is

$$f(\theta \mid y) = \frac{f(\theta) f(y \mid \theta)}{f(y)} = \begin{cases} 0.02778 & \text{If } \theta = 0.1 \\ 0.22222 & \text{If } \theta = 0.2 \\ 0.75 & \text{If } \theta = 0.3 \end{cases}$$

This result means that if the chosen die were to be tossed again a large number of times then there is a 75% chance that 6 would come up about 30% of the time, a 22.2% chance that 6 would come up about 20% of the time, and a 2.8% chance that 6 would come up about 10% of the time. ■

Proposition 1.1.1 — Proportionality Formula.
If $f(y)$ is a constant with respect to θ, then we can write

$$f(\theta \mid y) = \frac{f(\theta)(f(y \mid \theta)}{f(y)} = c \cdot f(\theta) f(y \mid \theta) \qquad \text{where} \qquad c = \frac{1}{f(y)}$$

That can be represent as $f(\theta \mid y) \propto_{\theta} f(\theta) f(y \mid \theta)$ or $f(\theta \mid y) \propto_{\theta} f(\theta) L(\theta \mid y)$ where $L(y \mid \theta)$ is the likelihood function.

Definition 1.1.5 — Conjugate Pair.
When the prior and posterior distributions are members of the **same class of distributions**, we say that they form a **conjugate pair**.

■ **Example 1.3**
Consider the binomial-beta model:

$$(y \mid \theta) \sim BIN(n, \theta)$$
$$\theta \sim Beta(\alpha, \beta) \quad \textbf{prior}$$
$$(\theta \mid y) \sim Beta(\alpha + y, \beta + n - y) \quad \textbf{posterior}$$

Both prior and posterior are beta distribution (same class of distribution), then they form a **conjugate pair** (prior is conjugate) ■

1.2 Bayesian Estimation

Definition 1.2.1 — Bayesian Point Estimation.
When $f(\theta \mid y)$ is given, the **Bayesian Point Estimation** of the model parameter θ can be calculated, it usually called **best estimate**. The following are three common point estimation function:

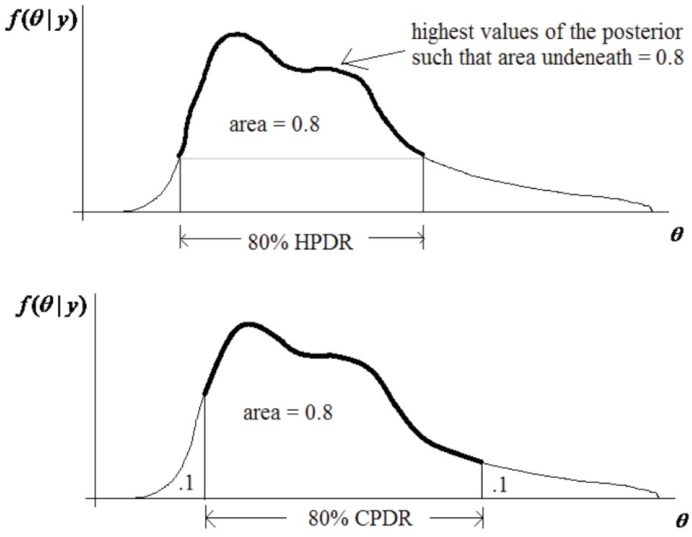
$$E(\theta \mid y) = \int \theta dF(\theta \mid y)$$
$$Mode(\theta \mid y) = \max_{\theta} f(\theta \mid y)$$
$$Median(\theta \mid y) = \text{the value of } \lambda \text{ such that } P(\theta \leq \lambda \mid y) \geq \frac{1}{2} \text{ and } P(\theta \geq \lambda \mid y) \geq \frac{1}{2}$$

Definition 1.2.2 — Bayesian Interval Estimation (HPDR and CPDR).
The $100(1-\alpha)\%$ **HPDR (highest posterior density region)** for θ is the smallest S s.t.

$$P(\theta \in S \mid y) \geq 1 - \alpha \quad \text{and} \quad f(\theta_1 \mid y) \geq f(\theta_2 \mid y) \quad \text{if } \theta_1 \in S \text{ and } \theta_2 \notin S$$

The $100(1-\alpha)\%$ **CPDR (central posterior density region)** for θ is the smallest interval $[a, b]$ such that

$$P(\theta < a \mid y) \leq \frac{\alpha}{2} \quad \text{and} \quad P(\theta > b \mid y) \geq \frac{\alpha}{2}$$



■ **Example 1.4**
Given that

$$f(\theta \mid y) = \begin{cases} 0.1 & \theta = 1 \\ 0.4 & \theta = 2 \\ 0.5 & \theta = 3 \end{cases}$$

Find 40% HPDR.

Only for $S = \{2\}$ or $\{3\}$ we have $P(\theta \in S \mid y) \geq 0.4$, but we also need $f(\theta_1 \mid y) \geq f(\theta_2 \mid y)$ if $\theta_1 \in S$ and $\theta_2 \notin S$, then we can only take $S = \{3\}$ ■

Proposition 1.2.1 — Inference on Function.
Let's define $\lambda = g(\theta)$ for some function g strictly increasing (decreasing), then by one-to-one transformation we have

$$f(\lambda \mid y) = f(\theta \mid y)\left|\frac{d\theta}{d\lambda}\right|$$

Then the estimated mean is

$$E[\lambda \mid y] = \int \lambda f(\lambda \mid y) d\lambda = \int g(\theta) f(\theta \mid y) d\theta = E[g(\theta) \mid y]$$

Definition 1.2.3 — Credibility Estimation.
Credibility Estimation is the one can be expressed in a weighted average form:

$C = (1-k)A + kB$

A: subjective estimate (collateral data estimate. ex. expected value)

B: objective estimate (direct data estimate, ex. MLE)

k: credibility factor with range $[0,1]$, the weight of B

Proposition 1.2.2 — Frequentist Characteristics of Bayesian Estimators.
Consider $(y_1, ..., y_n \mid u) \sim N(\mu, \sigma^2)$ are i,i.d and $\mu \sim N(\mu_0, \sigma_0^2)$. This leads to the point estimate $\hat{\mu} = \bar{y}$ and interval estimate $I = \bar{y} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. These estimates are exactly the same as the usual estimates used in the context of the corresponding classical model

$$y_1, ..., y_n \sim N(\mu, \sigma^2)$$

are i.i.d and μ is unknown constant and σ^2 is given. Therefore, Bayesian estimation is considered as a **proxy for classical estimates** and the frequentist operating characteristics of the Bayesian estimates are immediately known.

Frequentist: bias of $\hat{\mu}$ is 0 and coverage probability is exactly $1-\alpha$

Baysian: expected value of $(\bar{y} \mid \mu)$ is μ for all value of μ, $P(\mu \in I) = 1-\alpha$ for all value of μ

Definition 1.2.4 — Frequentist Relative Bias.

The **Frequentist Relative Bias** of Baysian point estimate is defined as

$$R_\theta = \frac{B_\theta}{\theta} \quad \text{where} \quad B_\theta = E[\hat{\theta} - \theta \mid \theta]$$

Definition 1.2.5 — Frequentist Coverage Probability.
The **Frequentist Coverage Probability (FCP)** of a Baysian interval estimate is

$$C_\theta = P(\theta \in I \mid I) \quad \text{where} \quad I = I(y) = [L(y), R(y)]$$

Definition 1.2.6 — Mixture Distribution.
A random variable X with **mixture distribution** has a distribution of the form

$$f(x) = \sum_{i=1}^{n} c_i \cdot f_i(x) \quad \text{where} \quad \sum_{i=1}^{n} c_i = 1 \ \text{ and } \ c_i \geq 0 \ \text{ and } \ f_i(x) \ \text{ is proper density for some distribution}$$

If our prior beliefs for θ do not follow any single well-know distribution, it can conveniently approximated to any degree precision by the **mixture prior distribution** from the above formula.

Remark: It can be shown if each $f_i(\theta)$ is conjugate then $f(\theta)$ is also conjugate.

Definition 1.2.7 — Priori Ignorance.
Priori Ignorance means there is no prior information at all.

■ **Example 1.5**
The normal-normal model $(y_1, ..y_n \mid \mu) \sim N(\mu, \sigma^2)$ and $\mu \sim N(\mu_0, \sigma_0^2)$, a **uninformative** prior is given by $\sigma_0 = \infty$, this gives $f(\mu) \propto 1$ for all μ

The normal-gamma $(y_1, ..y_n \mid \mu) \sim N(\mu, \frac{1}{\lambda})$ and $\lambda \sim Gamma(\alpha, \beta)$, a uninformative prior is given by $\alpha = \beta = 0$, this gives $f(\lambda) \propto \frac{1}{\lambda}$ ■

Definition 1.2.8 — Jeffreys Prior.
The **Jeffreys Prior** is given by the following

$$f(\theta) \propto \sqrt{I(\theta)} \quad \text{where} \quad I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \log f(y \mid \theta)\right)^2 \mid \theta\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(y \mid \theta) \mid \theta\right] \ \text{Fisher Information}$$

This is a prior which is invariant under reparameterisation, the following is the proof.
Let a prior be $f(\theta) \propto \sqrt{I(\theta)}$ and transformated parameter $\phi = g(\theta)$ with g is strightly increasing, then

$$\begin{aligned} f(\phi) \propto f(\theta)\left|\frac{\partial \theta}{\partial \phi}\right| \propto \sqrt{I(\theta)\left(\frac{\partial \theta}{\partial \phi}\right)^2} &= \sqrt{E\left[\left(\frac{\partial}{\partial \theta} \log f(y \mid \theta) \frac{\partial \theta}{\partial \phi}\right)^2 \mid \theta\right]} \\ &= \sqrt{E\left[\left(\frac{\partial}{\partial \phi} \log f(y \mid \phi)\right)^2 \mid \phi\right]} \\ &= \sqrt{I(\phi)} \end{aligned}$$

That is if we have $f(\theta) \propto \sqrt{I(\theta)}$, then for other parameter $\phi = g(\theta)$, then $f(\phi) \propto \sqrt{I(\phi)}$

1.3 Bayesian Decision Theory

Definition 1.3.1 — Loss Function.
The **loss function** L represents the cost incurred when the true value θ is estimated by $\hat{\theta}$ and usually satisfies the property $L(\theta, \hat{\theta}) = 0$

■ **Example 1.6**

Absolute Error Loss: $L(\hat{\theta}, \theta) = \left|\hat{\theta} - \theta\right|$
Quadratic Error Loss: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
Indicator Error Loss: $L(\hat{\theta}, \theta) = I(\hat{\theta} \neq \theta)$

■

Definition 1.3.2 — Risk Functions.
The **risk function** of θ is defined as:

$$R(\theta) = E[L(\hat{\theta}, \theta) \mid \theta] = \int L(\hat{\theta}(y), \theta) f(y \mid \theta) dy$$

which provides the idea of expected loss given any particular value of θ

If we want to obtain the expected loss, we need to find the overall expected loss, which is called **Bayes Risk** for L is defined as

$$r = EL(\hat{\theta}, \theta) = E[E[L(\hat{\theta}(y), \theta) \mid \theta]] = E[R(\theta)] = \int R(\theta) f(\theta) d\theta$$

Definition 1.3.3 — Posterior Expected Loss.
The **Posterior Expected Loss (PEL)** is the expectation of the loss function given the data, which is defined by

$$PEL(y) = E[L(\hat{\theta}, \theta) \mid y] = \int L(\hat{\theta}(y), \theta) f(\theta \mid y) d\theta$$

Then the **Bayes Risk** for **PEL** is

$$r = EL(\hat{\theta}, \theta) = E[E[L(\hat{\theta}(y), \theta) \mid y]] = E[PEL(y)] = \int PEL(y) f(y) dy$$

Definition 1.3.4 — Bayesian Estimator.
The **Bayesian Estimator** is defined as the choice of function $\hat{\theta} = \hat{\theta}(y)$ for which Bayes risk $r = EL(\hat{\theta}, \theta)$ of $PEL(y)$ is minimized.

Remark: The estimator that minimizes PEL(y) for all y can also minimizes the Bayes risk, that is because it's the weighted average of PEL(y)

■ **Example 1.7**
Find the Beayesian estimator for quadratic error loss function.

Note that

$$PEL(y) = E[(\hat{\theta} - \theta)^2 \mid y] = \hat{\theta}^2 - 2\hat{\theta}E[\theta \mid y] + E[\theta^2 \mid y] = \left[\hat{\theta} - E[\theta \mid y]\right]^2 - E^2[\theta \mid y] + E[\theta^2 \mid y]$$

We can see that $\hat{\theta} = E[\theta \mid y]$ minimzed the PEL(y) for all y, so it's the Bayes estimator. ■

■ **Example 1.8**
Find the Beayesian estimator for absolute error loss function

Let $t = \hat{\theta}$, first we note that

$$PEL(y) = \int_{-\infty}^{\infty} |t - \theta| f(\theta \mid y) d\theta = \int_{-\infty}^{t} (t - \theta) f(\theta \mid y) d\theta + \int_{t}^{\infty} (\theta - t) f(\theta \mid y) d\theta$$

then by **Leibniz's rule for differentiation**:

$$\frac{\partial}{\partial t} PEL(y) = \int_{-\infty}^{t} f(\theta \mid y) d\theta + \int_{t}^{\infty} (-1) f(\theta \mid y) d\theta = P(\theta < t \mid y) - P(\theta > t \mid y)$$

Setting this to zero get when $P(\theta < t \mid y) = P(\theta > t \mid y)$ gives us the $t = \hat{\theta}$ is the posterior median minimized the PEL(y), so it's the Bayes estimator. ■

■ **Example 1.9**
Find the Beayesian estimator for indicator error loss function.

Let $t = \hat{\theta}$, then

$$PEL(y) = E[L(t, \theta) \mid y] = E[1 - I(t = \theta) \mid y] = 1 - E[I(t = \theta) \mid y] = 1 - P(t = \theta \mid y) = 1 - f(\theta = t \mid y)$$

we can see PEL(y) is minimized when $t = \hat{\theta}$ maximizes the posterior density $f(\theta \mid y)$, which is posterior mode $Mode(\theta \mid y)$ if θ is discrete. ■

1.4 Bayesian Inference

■ **Remark 1.1 — Inference Given Function of Data.**
When we observe a **function** of data instead the data itself, the function typically **degrades** the information available in some way. For example, the **censoring** and **rounding**.

■ **Example 1.10 — Censoring.**
Each bulb of certain type has life is conditionally exponential with mean $m = \frac{1}{c}$, where c follow the standard exponential distribution. We observe $n = 5$ bulbs for 6 units of time, their life time are

$$D = \{y_1, y_2, y_3, y_4, y_5\} = \{2.6, 3.2, ?, 1.2, ?\}$$

where ? indicates a censored value, which is larger than 6. Find the posterior distribution and mean of the average light bulb lifetime m
First the probability of censoring is:

$$P(y > 6 \mid c) = \int_{6}^{\infty} ce^{-cy} dy = e^{-6c}$$

Then the posterior density of c:

$$\begin{aligned} f(c \mid D) \propto f(c)f(D \mid c) &\propto f(c) \cdot \prod_{i=1}^{5} f(y_i \mid c) \\ &= e^{-c} \cdot (ce^{-cy_1})(ce^{-cy_2})(e^{-6c})(ce^{-cy_4})(e^{-6c}) \\ &= c^{4-1}e^{-20c} \end{aligned}$$

so $(c \mid D) \sim G(4,20)$, $(m \mid D) \sim IG(4,20)$, so $E(m \mid D) = \frac{20}{3}$, which is higher than $\frac{1}{3} \cdot (2.6+3.2+6+1.2+6) = 3.8$ ∎

■ Example 1.11
Given $(y \mid \theta) \sim U(0,\theta)$ and $\theta \sim U(0,2)$ where $x = g(y)$ with rounding function g. Find the posterior density and mean of θ if $x = 1$.

Note that

$$P(x=1 \mid \theta) = P\left(\frac{1}{2} < y < \frac{3}{2} \mid \theta\right) = \begin{cases} \frac{\theta - \frac{1}{2}}{\theta} & \text{if } \frac{1}{2} < \theta < \frac{3}{2} \\ \frac{1}{\theta} & \text{if } \frac{3}{2} < \theta < 2 \end{cases}$$

then

$$f(\theta \mid x=1) \propto f(\theta)f(x \mid \theta) \propto g(\theta) = \begin{cases} \frac{\theta - \frac{1}{2}}{\theta} & \text{if } \frac{1}{2} < \theta < \frac{3}{2} \\ \frac{1}{\theta} & \text{if } \frac{3}{2} < \theta < 2 \end{cases}$$

Let's define $S = \int g(\theta)d\theta$, so we have the density $f(\theta \mid x=1) = \frac{g(\theta)}{S}$ and the expectation $E[\theta \mid x=1] = \int \theta f(\theta \mid x=1)d\theta = \frac{1}{S}$ ∎

Definition 1.4.1 — Posterior Predictive Distribution.
Given $f(y \mid \theta)$ and $f(\theta)$, consider any other quantity x whose distribution is defined by a density of the form $f(x \mid y, \theta)$. The **Posterior Predictive Distribution** of x is given by the **Posterior Predictive Density** $f(x \mid y)$. This can typically derived using the following equation:

$$f(x \mid y) = \int f(x, \theta \mid y)d\theta = \int f(x \mid y, \theta)f(\theta \mid y)d\theta$$

Moreover, we denote the $\hat{x} = E[x \mid y]$ as the **predictive mean** (posterior mean), and it can be written as

$$\hat{x} = E[x \mid y] = \int xf(x \mid y)dx \quad \text{or} \quad \hat{x} = E[x \mid y] = E[E[x \mid y, \theta] \mid y] = \int E[x \mid y, \theta]f(\theta \mid y)d\theta$$

The **predictive variance** is defined as

$$Var[x \mid y] = E[Var[x \mid y, \theta] \mid y] + Var[E[x \mid y, \theta) \mid y]$$

There is a special case in Bayesian predictive inference, where the quantity of interest x is an **independent future replicate** of y. This mean $(x \mid y, \theta)$ has exactly the same distribution as $(y \mid \theta)$, it can write it as $(x \mid y, \theta)$. Then in this case we can write $f(x \mid y, \theta)$ as $f(x \mid \theta)$, then

$$f(x \mid y) = \int f(x \mid \theta)f(\theta \mid y)d\theta$$

Definition 1.4.2 — Posterior predictive p-values.
In a single Bayesian model with data y and parameter θ, the theory of **Posterior predictive p-values** involves the followings steps:

1. Define a suitable **discrepancy measure** (test statistic), denoted as $T(y,\theta)$
2. Define x as **independent future replicate** of data y
3. Calculate the **Posterior predictive p-values (ppp-value)**:

$$p = P(T(x,\theta) \geq T(y,\theta) \mid y)$$

■ **Example 1.12**
Given $(y \mid \lambda) \sim Poi(\lambda)$ and $f(\lambda) = e^{-\lambda}$ with $\lambda > 0$, we observe $y = 3$.

1. Find a suitable ppp-value for testing:

$$H_0 : \lambda = 1 \quad \text{vs} \quad H_1 : \lambda > 2$$

2. Find a suitable ppp-value for testing:

$$H_0 : \lambda \in \{1,2\} \quad \text{vs} \quad H_1 : \lambda > 2$$

1. Define x as independent future replicate of data y to get $(x \mid y,\theta) \sim Poi(\lambda)$ and $T(y,\lambda) = y$, Then we have

$$p = P(x \geq y \mid y, \lambda = 1) = 1 - F_{Poi(1)}(y-1) = 1 - F_{Poi(1)}(2)$$

2. Note that

$$f(\lambda \mid y, H_0) \propto f(\lambda \mid H_0) \cdot f(y \mid H_0, \lambda) = \frac{e^{-\lambda}}{e^{-1}+e^{-2}} \cdot \frac{e^{-\lambda}\lambda^y}{y!}$$

Then we have

$$\begin{aligned} p = P(x \geq y \mid y, H_0) = P(x \geq 3 \mid H_0) &= P(x \geq 3 \mid \lambda = 1)P(\lambda = 1 \mid H_0) + P(x \geq 3 \mid \lambda = 2)P(\lambda = 2 \mid H_0) \\ &= (1 - F_{Poi(1)}(2)) \cdot P(\lambda = 1 \mid H_0) + (1 - F_{Poi(2)}(2)) \cdot P(\lambda = 2 \mid H_0) \end{aligned}$$

as desired. ■

Definition 1.4.3 — Multi-Parameters Bayesian Model.
The Bayesian model with parameter $\theta = (\theta_1, ..., \theta_n)$ called **Multi-Parameters Bayesian Model**. The joint prior density $f(\theta)$ can be written as unconditional prior multiplied by a conditional prior. For example, let $\theta = (\theta_1, \theta_2)$ then

$$f(\theta_1, \theta_2) = f(\theta_1) \cdot f(\theta_2 \mid \theta_1)$$

The **marginal posterior density** can be written as

$$f(\theta_1 \mid y) = \int f(\theta \mid y) d\theta_2 \qquad f(\theta_2 \mid y) = \int f(\theta \mid y) d\theta_1$$

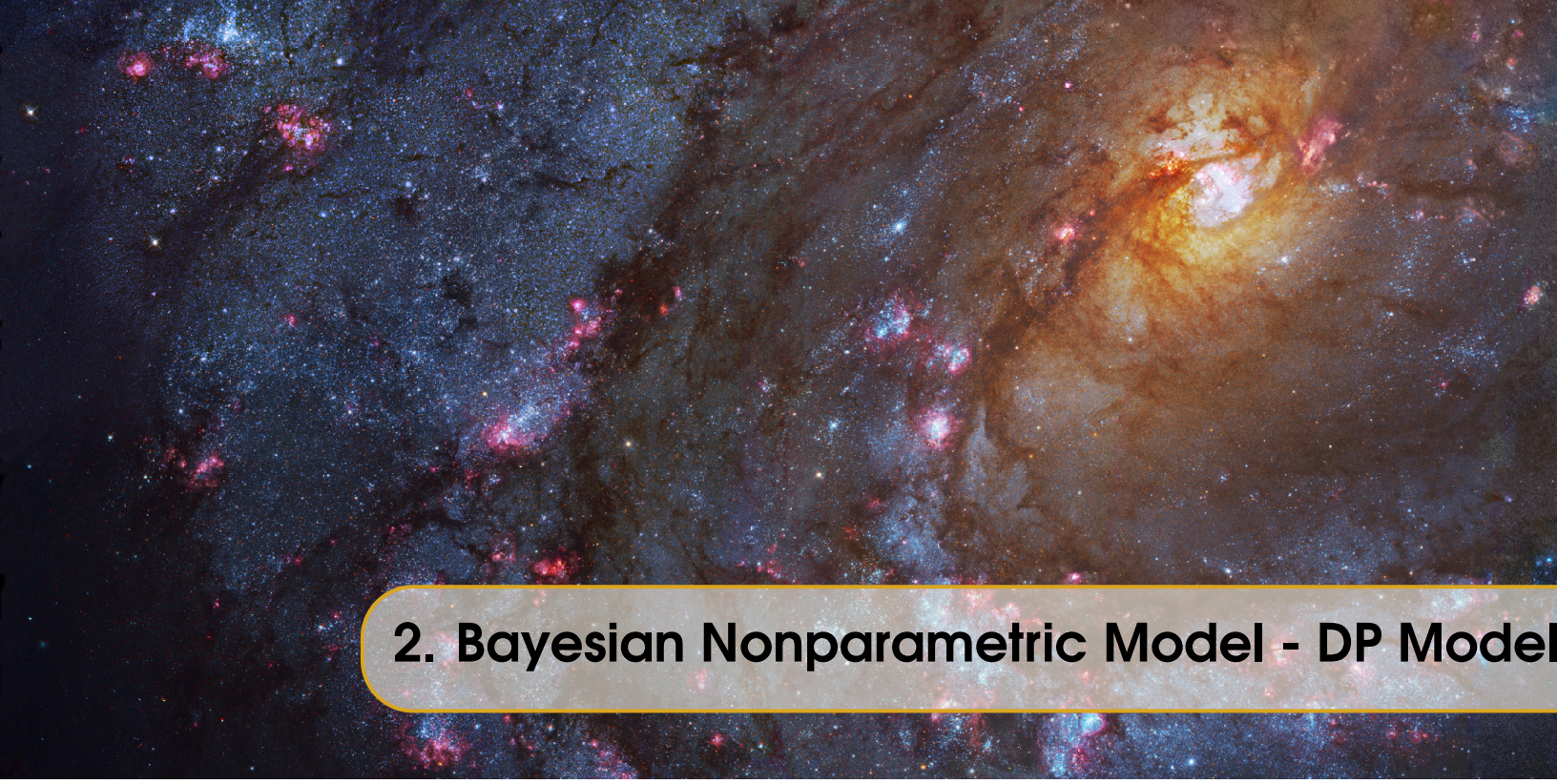
Then the **marginal posterior mean** of θ_1 is

$$\hat{\theta} = E[\theta_1 \mid y] = \int \theta_1 f(\theta_1 \mid y) d\theta_1 \quad \text{or} \quad \hat{\theta} = E[\theta_1 \mid y] = E[E[\theta_1 \mid y, \theta_2] \mid y] = \int E[\theta_1 \mid y, \theta_2] f(\theta_2 \mid y) d\theta_2$$

The **conditional posterior mean** of θ_1 is

$$E[\theta_1 \mid y, \theta_2] = \int \theta_1 f(\theta_1 \mid y, \theta_2) d\theta_1$$

and the **conditional posterior density** $f(\theta_1 \mid y, \theta_2)$ and $f(\theta_1 \mid y, \theta_2) \propto f(\theta_1, \theta_2 \mid y)$



2. Bayesian Nonparametric Model - DP Model

2.1 DP Models

Definition 2.1.1 — Dirichlet Distribution.
The **Dirichlet Distribution** of order $K \geq 2$ (number of category) with parameters $\alpha_1, ..., \alpha_K > 0$ has a probability density function with respect to Lebesgue Measure on the Euclidean space $\mathbb{R}^{K-1}$ given by

$$\mathbf{Dir}(\alpha_1, ..., \alpha_K) \quad : f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\sum_{i=1}^{K} x_i = 1$ and $x_i \in [0,1]$. The **normalizing constant** is the multivariate beta function, which can be express in terms of the gamma function:

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)} \qquad \alpha = (\alpha_1, ...\alpha_K) \qquad \alpha_0 = \sum_{i=1}^{K} \alpha_i$$

Mean : $E[X_i] = \frac{\alpha_i}{\alpha_0}$

Variance: $Var[X_i] = \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\alpha_0 + 1}$ where $\tilde{\alpha}_i = \frac{\alpha_i}{\alpha_0}$

Covariance: $Cov(X_i, X_j) = \frac{\delta_{ij}\tilde{\alpha}_i - \tilde{\alpha}_i\tilde{\alpha}_j}{\alpha_0 + 1}$ where $\delta_{ij} = 1$ iff $i = j$

Mode: $x_i = \frac{\alpha_i - 1}{\alpha_0 - K}$ $\alpha_i > 1$

Definition 2.1.2 — Dirichlet Process (DP).
Let $M > 0$ and G_0 be a probability measure defined on S. A **Dirichlet Process (DP)** with parameters (M, G_0) is a random probability measure G defined on S which assigns probability $G(B)$ to every measureable set B such that for each measureable finite partition $\{B_1, ..., B_k\}$ of S, the joint distribution of the vector

$$(G(B_1), ..., G(B_k)) \sim \mathbf{Dir}(MG_0(B_1), ..., MG_0(B_k))$$

where **Dir** denote the Dirichlet Distribution. The Dirichlet Process (DP) is usually denoted as $DP(M, G_0)$.

M : precision or total mass parameter

G_0 : centering measure

$\alpha \equiv MG_0$: base measure of the DP

Remark: G is a discrete measure, which can be written as a weighted sum of point masses

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$$

where $w_1, w_2....$ are probability weight and $\delta_x(\cdot)$ is the **Dirac measure** at x, i.e. $\mathbb{I}_A(x)$. Moreover, the **DP** has large weak support, it means under mild conditions, any distribution with the same support as G_0 can be approximated weakly by a DP random probability measure. If M is large, G is highly concentrated about G_0

Definition 2.1.3 — Constructive definition of DP.
For the process

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$$

the locations m_h are i.i.d draws from the centering measure G_0, and each weight w_h is defined as a fraction of $(1 - \sum_{l<h} w_l)$. That is a fraction of what is left after preceding $h-1$ point masses, what is

$$w_h = v_h \cdot \prod_{l<h} (1 - v_l)$$

with $v_h \sim Beta(1, M)$ i.i.d and $m_h \sim G_0$ i.i.d where v_h and m_h are independent. Then

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot)$$

defines a $DP(M, G_0)$ random probability measure. If $G \sim DP(M, G_0)$, $m \sim G_0$ and $W \sim Beta(1, M)$ are independent, then $W\delta_m(\cdot) + (1 - W)G(\cdot) \sim DP(M, G_0)$

■ **Remark 2.1**
If $G_0(A) > 0$, then the restriction of G to A is defined by $G\,|_A\,(B) = G(B \mid A) = \frac{G(A \cap B)}{G(A)}$ is also a DP with parameter M and $G_0\,|_A$ and independent of $G(A)$

Proposition 2.1.1 — Posterior Updating.
The DP is conjugate with respect to i.i.d sampling. That is under the sampling model with a DP on G, the posterior distribution for G is again a DP. The base measure of the posterior DP adds a point mass to the prior base measure at each observed data point y_i:

$$(y_1,...,y_n)\,|\,G \sim G \;\text{ and }\; G \sim DP(M,G_0) \quad\Longrightarrow\quad G\,|\,(y_1,...,y_n) \sim DP\left(M+n, \frac{M}{M+n}G_0 + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{y_i}\right)$$

That is the posterior DP centering measure is weighted average of G_0 and empirical distribution

$$\hat{f}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}(\cdot)$$

and posterior total mass parameter is increased to $M+n$

Algorithm 1: DP_Posterior_Update($G\_DP, \vec{y}$)

$n, Dirac \leftarrow Length(\vec{y}), 0$
for $d \leftarrow 1$ *to* n **do**
  $Dirac \leftarrow Dirac + \delta_{y[i]}$
$G\_DP.G_0 \leftarrow \frac{G\_DP.M}{G\_DP.M+n}\cdot G_0 + \frac{1}{G\_DP.M+n}\cdot Dirac$
$G\_DP.M \leftarrow M+n$
return

■ **Example 2.1 — Marginal Distribution.**
Consider random samples $y_i \,|\, G \sim G$ is i.i.d for $i = 1,..,n$. The discreteness of G implies a positive probability of ties among the y_i. That is at the heart of the Polya urn representation for marginal distribution

$$p(y_1,...,y_n) = \int \prod_{i=1}^{n} G(y_i)\, d\pi(G)$$

The Polya urn specifies the marginal distribution as a product of a sequence of increasing conditional

$$p(y_1,...,y_n) = p(y_1)\prod_{i=2}^{n} p(y_i \,|\, y_1,...,y_{i-1}) \quad \text{with} \quad p(y_i \,|\, y_1,...,y_{i-1}) = \frac{1}{M+i-1}\sum_{h=1}^{i-1}\delta_{y_h}(y_i) + \frac{M}{M+i-1}G_0(y_i)$$

for $i = 2,3,..$ and $y_1 \sim G_0$. Since y_i are i.i.d given G the marginal joint distribution of $(y_1,...,y_n)$ is exchangeable, that is the probabilities remain unchanged under any permutation of the indices. The complete conditional $p(y_i \,|\, y_h, h \neq i)$ has the same form as above for y_n, an important special case is the posterior predictive for a future observation y_{n+1} given $y_1,...,y_n$. It takes the form for $i = n+1$. ■

2.2 Dirichlet Process Mixture

The DP generates distributions that are discrete with probability one, making it awkward for continuous density estimation. This limitation can be fixed by convolving its trajectories with some continuous kernel, or more generally, by using a DP random measure as the mixing measure in a mixture over some simple parametric forms.

Definition 2.2.1 — Dirichlet Process Mixture (DPM).
Let Θ be finite-dimensional parameter space, for each $\theta \in \Theta$ let f_θ be a continuous p.d.f (kernel). Given a probability distribution G defined in Θ, a mixture of f_θ with respect to G has the p.d.f:

$$f_G(y) = \int f_\theta(y) dG(\theta)$$

The such mixtures can form a very rich family. Then, a prior on densities may be induced by putting a DP prior on the mixing distribution G, which is called **Dirichlet Process Mixture Models (DPM)**

■ **Remark 2.2**
The mixture model together with a DP prior on the mixing measure G can equivalently be written as a hierarchical model. Suppose $y_i \mid G \sim F_G$ is i.i.d, an equivalent hierarchical model is

$$\begin{aligned} y_i \mid \theta_i &\sim f_{\theta_i} \\ \theta_i \mid G &\sim G \end{aligned}$$

and $G \sim DP(M, G_0)$. The hierarchical model introduces new latent variables θ_i specific to each experimental unit. Just integrate with respect to θ_i to marginalize the hierarchical model with respect to θ_i. The result is exactly DPM. Under this hierarchical model, the posterior distribution on G is a mixture of DP's. That means $p(G \mid y_1, .., y_n)$ is a mixture of DP models, mixing with respect to the latent θ_i

Proposition 2.2.1
If $y_i \mid \theta_i \sim f_{\theta_i}$ is i.i.d for $i = 1, ..., n$ and $\theta_i \sim G$ is also i.i.d and $G \sim DP(\alpha) = DP(MG_0) = DP(M, G_0)$, then

$$G \mid \mathbf{y} \sim \int DP\left(\alpha + \sum_{i=1}^{n} \delta_{\theta_i}\right) dp(\theta \mid \mathbf{y})$$

where $\theta = (\theta_1, ..., \theta_n)$ and $\mathbf{y} = (y_1, ..., y_n)$

Definition 2.2.2 — Mixture of DPM.
The generalization of DPM model arises when the base measure of the DP prior includes unknow hyper-parameters η and the model is extended with a hyper-prior for η. Similarly the model could include a hyper-prior on a unknown total mass parameter M. The complete model is:

$$\begin{aligned} y_i \mid \theta_i &\sim f_{\theta_i} \\ \theta_i \mid G &\sim G \quad \text{i.i.d} \\ G \mid \eta, M &\sim DP(M, G_\eta) \\ (n, M) &\sim \pi \end{aligned}$$

For example, $G_\eta = N(m, s)$ with $\eta = (m, s)$ and a normal/inverse gamma prior on (m, s). The posterior characterization remains valid, now conditional on η and M.

2.3 Clustering Under the DPM

An important implication of is the fact that the DPM model induces a probability model on clusters, in the following sense. The discrete nature of the DP implies a positive probability for **ties** among the latent θ_i. Let's $\theta_j^\star$ for $j = 1,...,k$ where $k \leq n$ are the unique value from the n samples (**the j^{th} unique θ_i has lowest index i**). Define

$$S_j = \{i : \theta_i = \theta_j^\star\} \quad \text{and} \quad n_j = |S_j|$$

Then the multiset $\rho_n = \{S_1,...,S_k\}$ forms a partition of the set of experimental units $\{1,...,n\}$. Since θ_i is random, the sets S_j are random. That means the DPM implies a model on a random partition of the experimental units. The model $p(\rho_n)$ is also know as the Polya urn. The posterior model $p(\rho_n \mid \mathbf{y})$ reports posterior inference on clustering of the data.

Proposition 2.3.1 — Posterior simulation for DPM models.
Define $s_i = j$ if $i \in S_j$, then by definition:

$$\begin{aligned} &s_1 = 1 \\ &s_{i_2} = 2 \text{ for the lowest } i_2 > 1 \text{ with } \theta_{i_2} \neq \theta_1 \\ &s_{i_3} = 3 \text{ for the lowest } i_3 > i_2 \text{ with } \theta_{i_3} \notin \{\theta_1, \theta_{i_2}\} \\ &\vdots \end{aligned}$$

Let k_i denote the number of unique θ_ℓ among $\{\theta_1,....,\theta_i\}$ and let $n_{i,j}$ denote the multiplicity of the j^{th} of these unique values. Since we have

$$\sum_{j=1}^{k_i} n_{ij} = i$$

Then the following equation:

$$p(y_i \mid y_1,...,y_{i-1}) = \frac{1}{M+i-1}\sum_{h=1}^{i-1}\delta_{y_h}(y_i) + \frac{M}{M+i-1}G_0(y_i)$$

implies the density of $p(s_i \mid s_1,...,s_{i-1})$:

$$p(s_i = j \mid s_1,...,s_{i-1}) = \begin{cases} \dfrac{n_{i-1,j}}{M+i-1} & \text{for } j = 1,2,..,k_{i-1} \\ \dfrac{M}{M+i-1} & j = k_{i-1}+1 \end{cases}$$

Let $\mathbf{s}_{-i} = (s_1,...,s_{i-1},s_{i+1},..,s_n)$ By exchangeability of θ_i the conditional (prior) probability $p(s_i = j \mid \mathbf{s}_{-i})$ takes the same form as the density of $p(s_i \mid s_1,...,s_{i-1})$ from above for $i = n$. Also we can now read off the prior $p(\rho_n)$ as

$$p(s) = \prod_{i=2}^{n} p(s_i \mid s_1,...,s_{i-1}) = \frac{M^{k-1}\prod_{j=1}^{k}(n_j-1)!}{(M+1)\dots(M+n-1)}$$

Let's say implied conditional distribution for θ_i that follows the $p(s_i \mid s_1,...,s_{i-1})$, we define $\theta_{i,j}^\star$ denote the j^{th} unique value among $\{\theta_1,,,.,\theta_i\}$. Noting that $s_i = j$ implies $\theta_i = \theta_{i-1,j}^\star$ and $s_i = k_{i-1}+1$ implies that

$\theta_i \sim G_0$ by the density of $p(s_i \mid s_1, ..., s_{i-1})$. Then we have

$$p(\theta_i \mid \theta_1, ..., \theta_{i-1}) \propto \sum_{j=1}^{k_{i-1}} n_{i-1,j} \delta_{\theta^\star_{i-1,j}}(\theta_i) + MG_0(\theta_i)$$

Note that the DPM model is exchangeable in $\theta_1, ..., \theta_n$, then

$$p(\theta_i \mid \theta_{-i}) \propto \sum_{j=1}^{k^-} n_j^- \delta_{\theta_j^{\star -}}(\theta_i) + MG_0(\theta_i)$$

where k^- is the number of unique values in θ_{-i} and $\theta_j^{\star -}$ is the j^{th} unique element in it.

Algorithm 2: PosteriorSimulation

1 TBD

2.4 Posterior Simulation for DPM Models

2.5 Generalizations of the Dirichlet Processes