Frequentist vs Bayesian

Statistical methods, such as hypothesis testing (p-values), maximum likelihood estimates (MLE), and confidence intervals (CI), are known as frequentist methods. In the frequentist framework,

- probabilities refer to long run frequencies and are objective quantities;
- parameters are fixed but unknown constants;
- statistical procedures should have well-defined long run frequency properties (e.g. 95% CI).

There is another approach to inference known as Bayesian inference. In the Bayesian framework,

- probabilities reflect (subjective) personal belief;
- unknown parameters are treated as random variables;
- we make inferences about a parameter θ by producing a probability distribution for θ.

Bayesian Analysis

The Bayesian inference is carried out in the following way:

1. Choose a statistical model $p(x|\theta)$, i.e., the likelihood, same as in frequentist approach;

2. Choose a prior distribution $\pi(\theta)$;

3. Calculate the posterior distribution $\pi(\theta|x)$.

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}$$

Alternatively, we can write the posterior as

$$\begin{aligned} \pi(\theta|x) &= \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta} \\ &\propto p(x|\theta)\pi(\theta) \end{aligned}$$

where we drop the scale factor $\left[\int p(x|\theta)\pi(\theta)d\theta\right]^{-1}$ since it is a constant not depending on θ.

For example, if $\pi(\theta|x) \propto e^{-\theta^2+2\theta}$, then we know that $\theta|x \sim \mathsf{N}(1, 1/2)$.

Now any inference on the parameter θ can be obtained from the posterior distribution $\pi(\theta|x)$. For example, if one wants a

- point estimate of θ, we can report the mean ($\mathbb{E}(\theta \mid x)$), the median (median$(\theta \mid x)$), or the mode of the posterior distribution ($\max_\theta \pi(\theta \mid x)$);
- an interval estimate of θ, we can report the 95% credible interval, which is a region with 0.95 posterior probability.
- The mode estimate is often referred to as the MAP (maximum a posteriori) estimate. It maximizes

$$\log \pi(\theta|x) = \log p(x|\theta) + \log \pi(\theta),$$

which can be seen as a regularization of the MLE.

- Your first resistance to Bayesian inference may be the priors. Where does one find priors?
- Priors, like the likelihood, is part of your assumption: it's one's initial guess of the parameter; after observing the data which carry information about the parameter, one updates his/her prior to the posterior. Priors matter and do not matter.
- Next I'll introduce some default prior choices. Of course the sensitivity of prior choices–how different priors affect the final result–should always be examined in practice.

A Bernoulli Example

- Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ denotes the outcomes from n coin-tossings, 1 means a head and 0 means a tail. They are iid samples from a Bernoulli distribution with parameter θ, where θ is the probability of getting a head.
- Without any information about the coin, we can put a uniform prior on θ, that is,

$$\pi(\theta) = 1, \quad 0 \leq \theta \leq 1.$$

Next we calculate the posterior distribution of θ given $\mathbf{X}$.

$$\begin{aligned} \pi(\theta|\mathbf{X}) &\propto \prod_{i=1}^{n} p(X_i|\theta) \\ &\propto \theta^s(1-\theta)^{n-s}, \quad s = \sum X_i, \end{aligned}$$

which implies that $\theta|\mathbf{X} \sim \text{Beta}(s+1, n+1-s)$. The corresponding posterior mean can be used as a point estimate for θ,

$$\hat{\theta} = \frac{s+1}{n+2}. \tag{1}$$

$$\text{Post-mean} = \frac{s+1}{n+2}, \quad \text{MLE} = \frac{s}{n}.$$

MLE is equal to the observed frequency of heads among n experiments; the Bayes estimator is the frequency of heads among $(n+2)$ experiments in which there are two "prior" experiments, one is a head and the other one is a tail. Without the data, one just looks at the prior experiments and a reasonable guess for θ is $1/2$. After observing the data, the final estimate (1) is some number between $1/2$ and s/n as a compromise between the prior information and the MLE.

$$\text{Post-mean} = \frac{s+1}{n+2}, \quad \text{MLE} = \frac{s}{n}.$$

The extra counts–one for head and one for tail–are often called the pseudo-counts. Having pseudo-counts is appealing in cases where θ is likely to take extreme values close to 1 or 0. For example, to estimate θ for a rare event, it is likely to observe $X_i = 0$ for all $i = 1, \ldots, n$, but it may be dangerous to conclude $\hat{\theta} = 0$.

Beta distributions are often used as a prior on θ; in fact, uniform is a special case of Beta. Suppose the prior on θ is Beta(α, β),

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)},$$

then

$$\pi(\theta|\mathbf{X}) \sim \text{Beta}(s+\alpha, n-s+\beta). \tag{2}$$

We call Beta distributions the conjugate family for Bernoulli models since both the prior and the posterior distributions belong to the same family. The posterior mean of (2) is equal to $(\alpha+s)/(n+\alpha+\beta)$. So Beta priors can be viewed as having $(\alpha+\beta)$ prior experiments in which we have α heads and β tails.

A Multinomial Example

- Suppose you randomly draw a card from an ordinary deck of playing cards, and then put it back in the deck. Repeat this exercise five times (i.e., sampling with replacement).

- Let (N_1, N_2, N_3, N_4) denote the number of spades, hearts, diamonds, and clubs among the five cards. We say (N_1, N_2, N_3, N_4) follow a multinomial distribution, with a probability distribution function given by

$$\begin{aligned} & P(N_1 = n_1, \ldots, N_4 = n_4 \mid n, \theta_1, \ldots, \theta_4) \\ = & \frac{(n)!}{(n_1)!(n_2)!(n_3)!(n_4)!} \theta_1^{n_1} \cdots \theta_4^{n_4}, \end{aligned}$$

  where $n_1 + \cdots n_4 = n = 5$ and $\theta_1 = \cdots = \theta_4 = 1/4$.

- In the Bernoulli example, we conduct n independent trials and each trial results in one of two possible outcomes, e.g., head or tail, with probabilities θ and $(1-\theta)$, respectively.
- In the multinomial example, we conduct n independent trials and each trial results in one of k possible outcomes, e.g., $k = 4$ in the aforementioned card example, with probabilities $\theta_1, \ldots, \theta_k$, respectively.

- For multinomial distributions, the parameter of interest is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ which lies in a simplex of $\mathbb{R}^k$,

$$\mathcal{S} = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k); \sum_i \theta_i = 1, \theta_i \geq 0\}.$$

- A Dirichlet distribution on $\mathcal{S}$, $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, is an extension of the Beta distribution, with a density function given by

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}.$$

  The Dirichlet distributions are conjugate priors for Multinomial models.

A Normal Example

Assume $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The parameters here are (θ, σ^2) and we would like to get the posterior distribution $\pi(\theta, \sigma^2|\mathbf{X})$. If using Gibbs sampler which will be introduced later, we only need to know the posterior distribution of $\pi(\theta|\sigma^2, \mathbf{X})$ and $\pi(\sigma^2|\theta, \mathbf{X})$.

For the location parameter θ, the conjugate prior is normal. Suppose $\pi(\theta) = N(\mu_0, \tau_0^2)$, then $\theta|\sigma^2, \mathbf{X} \sim N(\mu, \tau^2)$ where

$$\mu = w\bar{X} + (1-w)\mu_0, \quad w = \frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma^2/n} + \frac{1}{\tau_0^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\sigma^2/n} + \frac{1}{\tau_0^2}.$$

For the scale parameter σ^2, the conjugate prior is Invse Gamma, that is, the prior on $1/\sigma^2$ is Gamma. Suppose $\pi(\sigma^2) = \text{InvGa}(\alpha, \beta)$, then

$$\begin{aligned}\pi(\sigma^2|\mu, \mathbf{X}) &\propto \Big(\frac{1}{\sigma^2}\Big)^{n/2} \exp\{-\frac{\sum(X_i-\mu)^2}{2\sigma^2}\}\Big(\frac{1}{\sigma^2}\Big)^{\alpha-1} e^{-\frac{\beta}{\sigma^2}} \\ &\sim \text{InvGa}\big(\frac{n}{2}+\alpha, \frac{\sum(X_i-\mu)^2}{2}+\beta\big).\end{aligned}$$

In practice, we often specify $\pi(\sigma^2)$ using Invχ^2 distributions which are special cases of InvGa,

$$\text{Inv}\chi^2(v_0, s_0^2) = \text{InvGa}(\frac{v_0}{2}, \frac{v_0}{2}s_0^2).$$

With prior Inv$\chi^2(v_0, s_0^2)$ the posterior distribution is also Inv$\chi^2(v_n, s_n^2)$ where

$$v_n = v_0 + n, \quad v_n s_n^2 = v_0 s_0^2 + \sum(X_i-\mu)^2.$$

Gibbs Samplers for Posterior Inference

Suppose the random variables X and Y have a joint probability density function $p(x, y)$.

Sometimes it is not easy to simulate directly from the joint distribution.

Instead, suppose it is possible to simulate from the individual conditional distributions $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$.

Then the Gibbs sampler draws $(X_1, Y_1), \ldots, (X_T, Y_T)$ as follows:

1. Initialization: let (X_0, Y_0) be some starting values; set $n = 0$.
2. draw $X_{n+1} \sim p_{X|Y}(x|Y_n)$
3. draw $Y_{n+1} \sim p_{Y|X}(y|X_{n+1})$
4. Go to step 2 and repeat.

Gibbs samplers are MCMC algorithms, and they produce samples from the desired distributions after a so-called burning period. So in practice, we always drop some samples from the initial steps (say, for example, 1000 or 5000 steps) and start saving samples after that.

In Bayesian analysis, suppose we have multiple parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. Then we can draw the posterior samples of $\boldsymbol{\theta}$ using a multi-stage Gibbs sampler and at each stage, we draw θ_i from the conditional distribution $\pi(\theta_i|\boldsymbol{\theta}_{[-i]}, \text{Data})$ where $\boldsymbol{\theta}_{[-i]}$ denotes the $(K-1)$ parameters except θ_i. The reason for Gibbs samplers is that in many cases the conditional distribution of $\pi(\boldsymbol{\theta}|\text{Data})$ is not of closed form, while all those conditionals are.

How to describe the data generating process for a sequence of binary variables $X_1, X_2, \ldots$?

- (Frequentist) X_i iid $\sim$ Bern(p)
- (Bayesian) X_i's are exchangeable. Then by de Finetti's Theorem, X_i's are generated in the following way:
  1. Generate $p \sim \pi(\cdot)$
  2. Conditioning on p, X_i's iid $\sim$ Bern(p).

Pólya's Urn Process

- Suppose that we have an urn that initially contains r red and b blue balls
- At each trial, we select a ball from the urn and then return the ball to the urn along with c new balls of the same color.
  - $c = -1$: sampling without replacement
  - $c = 0$: sampling with replacement
- The random process is known as Pólya's urn process, which is a generalization of the standard models of sampling with and without replacement.

Let X_i denote the color of the ball selected at time i, where 1 denotes red and 0 denotes blue. What's the distribution of $(X_1, \ldots, X_n)$? [a]

$$
\begin{aligned}
& P(X_1, \ldots, X_n) \\
= \; & P(X_1) P(X_1 \mid X_2) \cdots P(X_n \mid X_1, \ldots, X_{n-1}) \\
= \; & \frac{r(r+c)\cdots(r+(s-1)c)b(b+c)\cdots(b+(n-s-1)c)}{(r+b)(r+b+c)(r+b+2c)\cdots(r+b+(n-1)c)},
\end{aligned}
$$

where the distribution only depends on $s = \sum_{i=1}^{n} X_i$. So $(X_1, \ldots, X_n)$ is an exchangeable sequence.

[a] Assume $c \geq 0$, so n can be any integer.

- $(X_1, \ldots, X_n)$ is an iid sequence of random variables if and only if $c = 0$ (sampling with replacement);
- When $c > 0$, $(X_1, \ldots, X_n)$ are not independent, but dependent.

Pólya's urn is one of the most famous examples of a random process in which the outcome variables are exchangeable, but dependent (in general).

By de Finetti's Theorem, [a] we know that the joint distribution of $(X_1, \ldots, X_n)$ can be written as

$$p(X_1, \ldots, X_n) = \int \Big[\prod_{i=1}^{n} p(X_i \mid \theta) \Big] \pi(\theta) d\theta.$$

What's $\pi(\theta)$?

$$\pi(\theta) = \mathsf{Be}\Big(\frac{\alpha}{c}, \frac{\beta}{c}\Big).$$

[a] An infinite exchangeable sequence is distributed as a mixture of iid sequences.

A Mixture Model Example

Suppose $X_1, X_2, \ldots, X_n$ iid from

$$w\ \mathsf{N}(\mu_1, \sigma_1^2) + (1-w)\ \mathsf{N}(\mu_2, \sigma_2^2).$$

For each X_i, we introduce a latent variable Z_i indicating which component X_i is generated from and

$$P(Z_i = 1) = w, \quad P(Z_i = 2) = 1 - w.$$

The parameters of interest are $\theta = (w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and their prior distributions are specified as follows

$$w \sim \mathsf{Be}(1,1), \quad \mu_1, \mu_2 \sim \mathsf{N}(0, \tau^2), \quad \sigma_1^2, \sigma_2^2 \sim \mathsf{InvGa}(\alpha, \beta).$$

$$\begin{aligned}\pi(\theta, \mathbf{Z} \mid \mathbf{X}) &= \prod_{i=1}^{n} \left[p(Z_i \mid \theta) p(X_i \mid Z_i, \theta)\right] \\ &\quad \times \pi(w)\pi(\mu_1)\pi(\mu_2)\pi(\sigma_1^2)\pi(\sigma_2^2)\end{aligned}$$

In a Gibbs sampler, we iteratively sample each element from

$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, Z_1, \ldots, Z_n)$$

conditioning on the other elements being fixed. For example,

- how to sample Z_i?

$$P(Z_i = 1 \mid \mathbf{X}, \text{ others}) \propto w p(X_i \mid Z_i = 1)\pi(w).$$

- how to sample μ_1?

$$\mu_1 \mid \mathbf{X}, \text{ others} \propto \Big[\prod_{i:Z_i=1} p(X_i \mid \mu_1, \sigma_1^2) \Big] \times \pi(\mu_1).$$

The Gibbs sampler iterates the following steps:

1. Draw Z_i from Bernoulli for $i = 1, \ldots, n$;
2. Draw w from Beta;
3. Draw μ_1, μ_2 from Normal;
4. Draw σ_1^2, σ_2^2 from InvGa.

If we just want to obtain a MAP estimate of the parameters, we can use the EM algorithm. Recall that

$$\hat{\theta} = \arg\max_{\theta} P(\mathbf{X} \mid \theta)\pi(\theta).$$

$$\begin{aligned}
& \log p(\mathbf{x}, \mathbf{Z} \mid \theta)\pi(\theta) \\
= \quad & \sum_i \mathbf{1}(Z_i = 1) \times \Big[\log \phi_{\mu_1,\sigma_1^2}(x_i) + \log w\Big] \\
& + \sum_i \mathbf{1}(Z_i = 2) \times \Big[\log \phi_{\mu_2,\sigma_2^2}(x_i) + \log(1-w)\Big] \\
& + \log \pi(w) + \log \pi(\mu_1) + \log \pi(\mu_2) + \log \pi(\sigma_1^2) + \log \pi(\sigma_2^2)
\end{aligned}$$

Similar to the EM algorithm for MLE:

- at the E-step, we replace $\mathbf{1}(Z_i = 1)$ and $\mathbf{1}(Z_i = 2)$ by its expectation, i.e., the probability of $Z_i = 1$ or 2 conditioning on the data $\mathbf{X}$ and some initial guess of the parameter θ_0

$$\gamma_i = P(Z_i = 1 \mid x_i, \theta_0) = \frac{w\phi_{\mu_1,\sigma_1^2}(x_i)}{w\phi_{\mu_1,\sigma_1^2}(x_i) + (1-w)\phi_{\mu_2,\sigma_2^2}(x_i)};$$

- at the M-step, we update θ,
- and iterative between the E and M steps, until convergence.

The M-step is slightly different from the one for MLE.

- Without the Beta prior, we would update w by γ_+/n. But with the Beta prior on w, we need to add the pseudo-counts.
- Similarly, without the prior, we would update μ_1 by

$$\frac{1}{\gamma_+}\sum_i \gamma_i x_i,$$

  but with the prior, we would update μ_1 by a weighted average of the value above and the prior mean for μ_1.

A Mixture with Infinite Components

In the model-based clustering approach, we model the data $(x_1, \ldots, x_n)$ by a mixture with K components:

$$\sum_{k=1}^{K} w_k f_k(x \mid \theta_k).$$

We specify prior distribution for $(\mathbf{w}, \theta_1, \ldots, \theta_K)$ as

$$\mathbf{w} \sim \mathsf{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right), \quad \theta_k \text{ iid } \sim G_0.$$

For example, if f_k is normal with $\theta_k = (\mu_k, \sigma_k^2)$, then

$$G_0(\theta) = \mathsf{N}(\mu; 0, \tau^2) \times \mathsf{InvGa}(\sigma^2; \alpha_0, \beta_0).$$

Introduce the latent variable $Z_i = 1, \ldots, K$, where

$$P(Z_i = k \mid \mathbf{w}) = w_k, \quad k = 1, \ldots, K.$$

In Bayesian analysis, if we are not interested in the inference on $\mathbf{w}$, we integrate it over. That is, we work directly with

$$P(Z_1, \ldots, Z_n) = \int P(Z_1, \ldots, Z_n \mid \mathbf{w}) \pi(\mathbf{w}) d\mathbf{w}.$$

Next let's see what the distribution $P(Z_1, \ldots, Z_n)$ looks like.

$$P(Z_1, \ldots, Z_n) = P(Z_1) P(Z_2 \mid Z_1) \times \cdot \times P(Z_n \mid Z_1, \ldots, Z_{n-1}).$$

We have $P(Z_1 = k) = \alpha/K$, [a]

$$P(Z_{i+1} = k \mid z_1, \ldots, z_i) = \frac{P(Z_{i+1} = k, z_1, \ldots, z_i)}{P(z_1, \ldots, z_i)}$$

$$= \frac{\int w_1^{n_1+\alpha/K-1} \cdots w_k^{n_k+1+\alpha/K-1} \cdots w_K^{n_K+\alpha/K-1} d\mathbf{w}}{\int w_1^{n_1+\alpha/K-1} \cdots w_k^{n_k+\alpha/K-1} \cdots w_K^{n_K+\alpha/K-1} d\mathbf{w}}$$

$$= \frac{\Gamma(i+\alpha)}{\Gamma(i+\alpha+1)} \frac{\Gamma(n_k+\alpha/K+1)}{\Gamma(n_k+\alpha/K)}$$

$$= \frac{n_k+\alpha/K}{i+\alpha}$$

where $n_k = \#\{j : z_j = k, j = 1 : i\}$. Next Let $K \to \infty$.

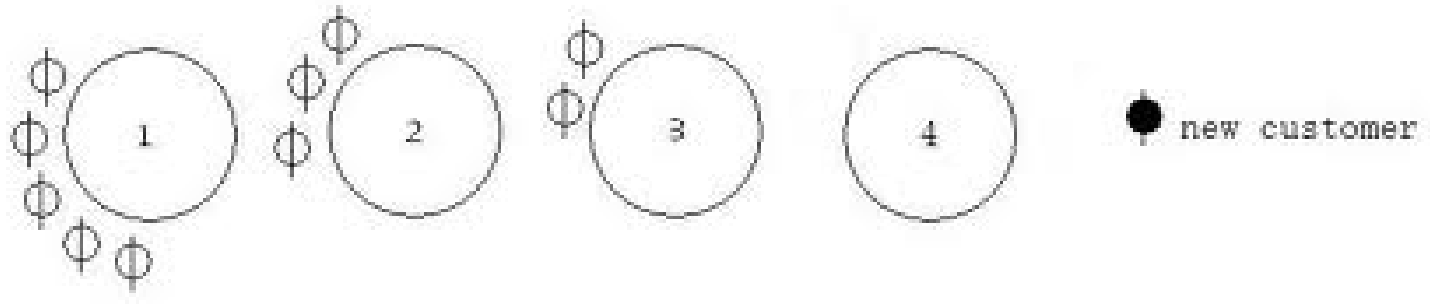
[a] Or equivalently, $P(Z_1 = 1)$ if we always label Z_1 as 1.

Then we have

$$P(Z_{i+1} = k \mid z_1, \ldots, z_i) = \frac{n_k}{i + \alpha},$$

$$P(Z_{i+1} \neq z_j \text{ for all } j = 1 : i \mid z_1, \ldots, z_i) = \frac{\alpha}{i + \alpha},$$

Known as the Chinese restaurant process. [a]



[a] It can be viewed as the continuous version of the Polya Urn model.

Clustering with Chinese Restaurant Process

A mixture model

$$X_i \mid Z_i = k \sim f_{\theta_k^*}.$$

Prior on θ_k^*:

$$\theta_k^* \text{ iid } \sim G_0.$$

Prior on Z_i's is CRP(α): $Z_1 = 1$;

for $i \geq 1$, suppose $Z_1, \ldots, Z_i$ form m tables with size $n_1, \ldots, n_m$, then

$$\begin{aligned} P(Z_{i+1} = k \mid Z_1, \ldots, Z_i) &= \frac{n_k}{i+\alpha}, \quad k = 1, \ldots, m; \\ P(Z_{i+1} = m+1 \mid Z_1, \ldots, Z_i) &= \frac{\alpha}{i+\alpha}. \end{aligned}$$

Alternatively, you can describe the prior first and then the likelihood (which gives you a clear idea of how data are generated):

- Set $Z_1 = 1$, generate $\theta_1^* \sim G_0$ and $X_1 \sim f_{\theta_1^*}$
- Loop over $i = 1, \ldots, n-1$: suppose the previous i samples form m tables

$$\begin{aligned} P(Z_{i+1} = k \mid Z_1, \ldots, Z_i) &= \frac{n_k}{i+\alpha}, \quad k = 1, \ldots, m; \\ P(Z_{i+1} = m+1) \mid Z_1, \ldots, Z_i) &= \frac{\alpha}{i+\alpha}. \end{aligned}$$

If $Z_{i+1} = m+1$, generate $\theta_{m+1}^* \sim G_0$. Then generate

$$X_{i+1} \sim f_{\theta_{Z_{i+1}}^*}.$$

Advantages

- We do not need to specify K.
- K is treated as a random variable, and its (posterior) distribution is learned from the data.
- Can model unseen data: for any new sample X^*, there is always a positive chance that it can start a new cluster.

Posterior Inference

Let's consider a simple case, where

$$f_{\theta_k^*} = \mathsf{N}(\theta_k^*, 1), \quad G_0 = \mathsf{N}(0, 1), \quad \alpha = \alpha_0.$$

Posterior distribution

$$\pi(\mathbf{Z}, \boldsymbol{\theta}^* \mid \mathbf{x}) \propto \left[\prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}^*, Z_i)\right] \times p(Z_1, \ldots Z_n | \alpha_0) \times \pi(\boldsymbol{\theta}^*)$$

Posterior sampling: MCMC or Variational Bayes. [a]

[a] Check Neal (2000) for MCMC and Blei and Jordan (2004) for VB.

How to sample $(Z_i|\mathbf{x}, \mathbf{Z}_{[-i]}, \boldsymbol{\theta}^*)$? Suppose $\mathbf{Z}_{[-i]}$ form m tables.

$$\begin{aligned}
\pi(Z_i \mid \cdots) &\propto p(x_i \mid \boldsymbol{\theta}^*, Z_i) \times p(\mathbf{Z}_{[-i]}, Z_i) \\
&\propto p(x_i \mid \boldsymbol{\theta}^*, Z_i) \times p(Z_i \mid \mathbf{Z}_{[-i]}) \\
&\propto \sum_{k=1}^{m} \mathbf{1}(Z_i = k) f_{\theta^*}(x_i) + \mathbf{1}(Z_i = m+1) g(x_i)
\end{aligned}$$

where

$$g(x) = \int f_\theta(x) G_0(\theta) d\theta.$$

Non-parametric Bayesian (NB) Models for Clustering

- The finite mixture model:

$$X_i \mid Z_i = k \sim f_{\theta_k^*}, \qquad P(Z_i = k) = p_k.$$

$$\theta_k^* \text{ iid } \sim G_0, \qquad \mathbf{p} \sim \text{Dir}(\alpha/K, \ldots, \alpha/K).$$

- Alternatively,

$$X_i \mid \theta_i \sim f_{\theta_i}, \qquad \theta_i \mid G \sim G$$

$$G(\cdot) = \sum_{k=1}^{K} p_k \delta_{\theta_k^*}(\cdot), \qquad \mathbf{p} \sim \text{Dir}\Big(\frac{\alpha}{K}, \cdots\Big), \quad \theta_k^* \sim G_0.$$

The prior on G is a K-element discrete dist. In the NB approach, we'll drop this restriction.

- A DPM model for clustering

$$X_i \mid \theta_i \sim f_{\theta_i}, \qquad \theta_i \mid G \sim G$$

$$G \sim \text{DP}(\alpha, G_0)$$

where $\text{DP}(\alpha, G_0)$ denotes a Dirichlet Process with a scale (precision) parameter α and a base measure G_0.

Dirichlet Process (DP)

$$\theta_i \mid G \text{ iid } G, \quad G \sim \text{DP}(\alpha, G_0)$$

- Define DP as a distribution over distributions (Ferguson, 1973)
- Describe DP as a stick-breaking process (Sethuraman, 1994)
- If we integrate over G (wrt DP), the resulting prior on $(\theta_1, \ldots, \theta_n)$,

$$\pi(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} G(\theta_i) d\Pi(G)$$

  is the Chinese restaurant process (CRP).

Dirichlet Process

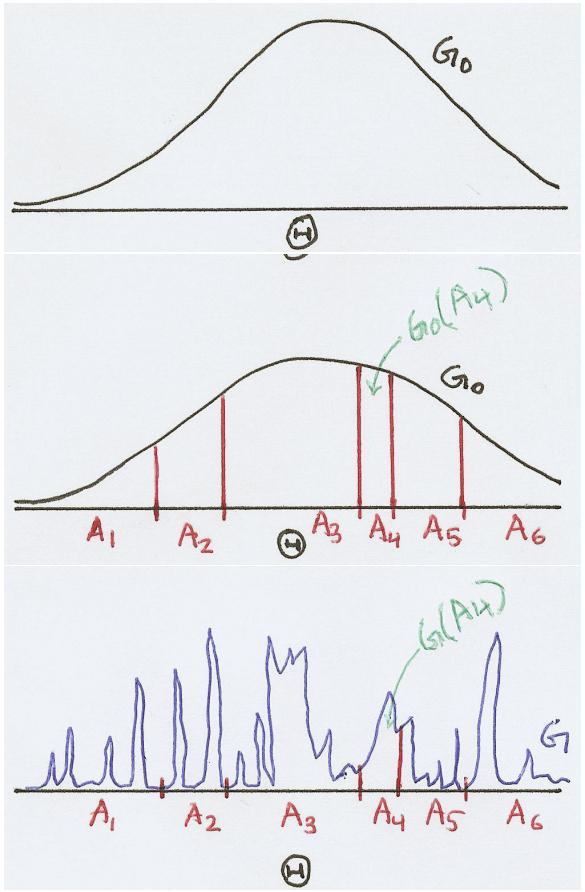
Let Θ be a measurable space, G_0 be a probability measure on Θ, and α a positive real number.

For all $(A_1, \ldots A_K)$ finite partitions of Θ,

$$G \sim \mathsf{DP}(\cdot | G_0, \alpha)$$

means that

$$(G(A_1), \ldots, G(A_K)) \sim \mathsf{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$



(Ferguson, 1973)

Dirichlet Process

$G \sim \text{DP}(\cdot|G_0, \alpha)$ OK, but what does it look like?

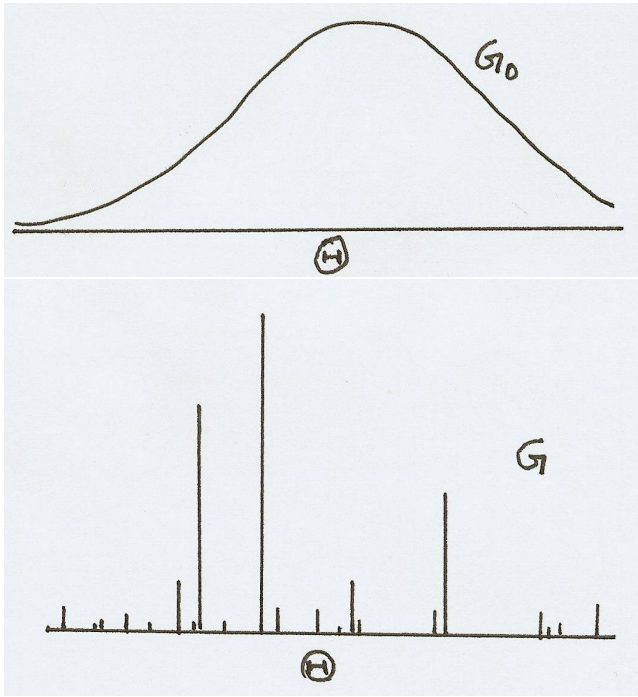
Samples from a DP are discrete with probability one:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

where $\delta_{\theta_k}(\cdot)$ is a Dirac delta at θ_k, and $\theta_k \sim G_0(\cdot)$.

Note: $E(G) = G_0$

As $\alpha \to \infty$, G looks more like G_0.



Dirichlet Processes: Stick Breaking Representation

$$G \sim \text{DP}(\cdot|G_0, \alpha)$$

Samples G from a DP can be represented as follows:

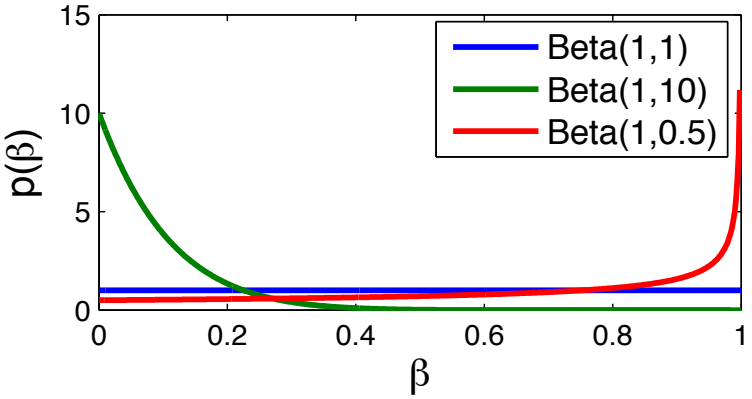
$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot)$$

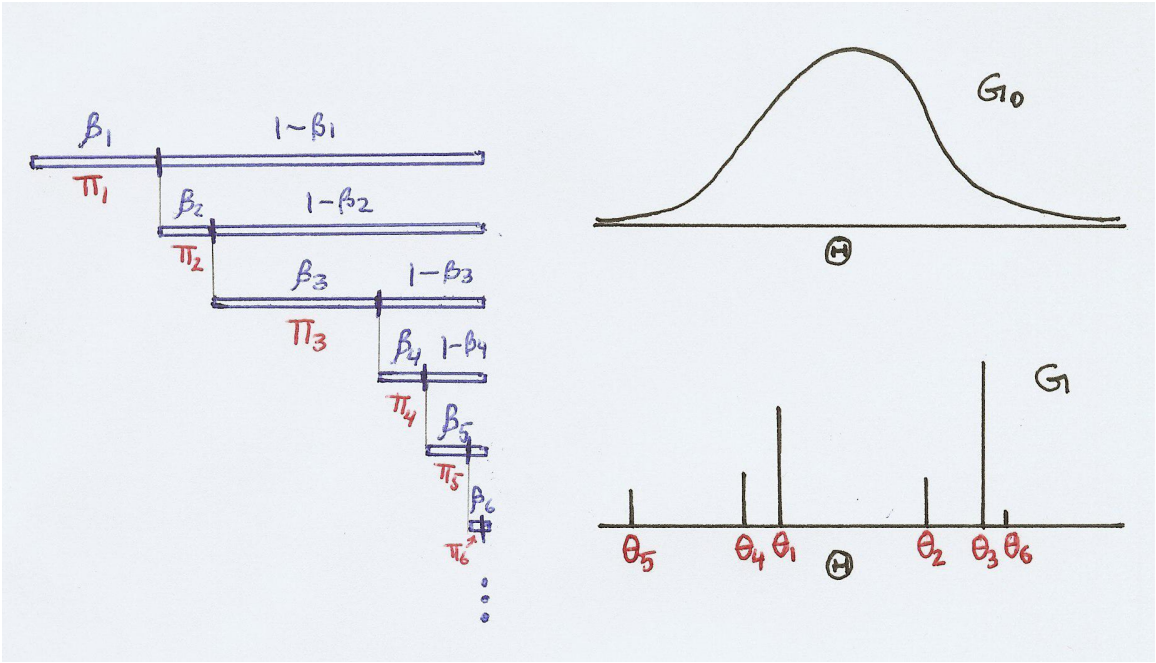
where $\theta_k \sim G_0(\cdot)$, $\sum_{k=1}^{\infty} \pi_k = 1$,

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$$

and

$$\beta_k \sim \text{Beta}(\cdot|1, \alpha)$$





(Sethuraman, 1994)

Dirichlet Process: Conjugacy

$$G \sim \mathsf{DP}(\cdot|G_0, \alpha)$$

If the prior on G is a DP:

$$P(G) = \mathsf{DP}(G|G_0, \alpha)$$

...and you observe θ...

$$P(\theta|G) = G(\theta)$$

...then the posterior is also a DP:

$$P(G|\theta) = \mathsf{DP}\left(\frac{\alpha}{\alpha+1}G_0 + \frac{1}{\alpha+1}\delta_\theta,\ \alpha+1\right)$$

Generalization for n observations:

$$P(G|\theta_1, \ldots, \theta_n) = \mathsf{DP}\left(\frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i},\ \alpha+n\right)$$

Analogous to Dirichlet being conjugate to multinomial observations.

Dirichlet Process

Blackwell and MacQueen's (1973) urn representation

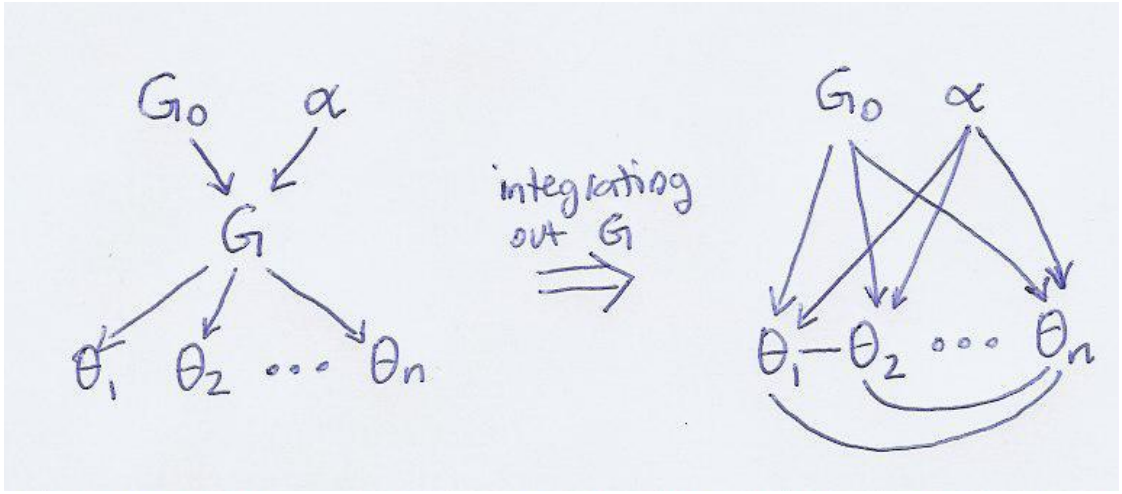
$$G \sim \mathrm{DP}(\cdot|G_0, \alpha) \quad \text{and} \quad \theta|G \sim G(\cdot)$$

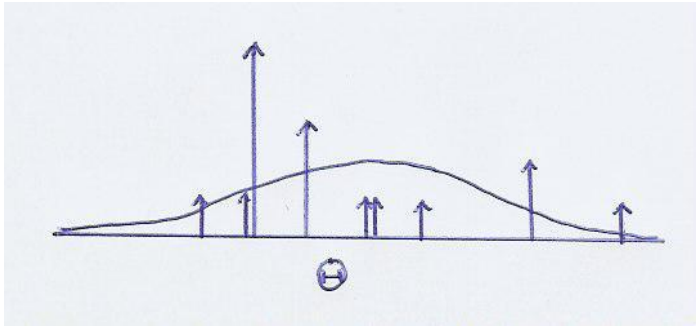
Then

$$\theta_n|\theta_1, \ldots \theta_{n-1}, G_0, \alpha \quad \sim \quad \frac{\alpha}{n-1+\alpha} G_0(\cdot) + \frac{1}{n-1+\alpha} \sum_{j=1}^{n-1} \delta_{\theta_j}(\cdot)$$

$$P(\theta_n|\theta_1, \ldots \theta_{n-1}, G_0, \alpha) \quad \propto \quad \int dG \prod_{j=1}^{n} P(\theta_j|G) P(G|G_0, \alpha)$$

The model exhibits a "clustering effect".





Chinese Restaurant Process (CRP)

This shows the clustering effect explicitly.
Restaurant has infinitely many tables $k = 1, \ldots$.
Customers are indexed by $i = 1, \ldots$, with values ϕ_i
Tables have values θ_k drawn from G_0

K = total number of occupied tables so far.
n = total number of customers so far.
n_k = number of customers seated at table k

Generating from a CRP:
customer 1 enters the restaurant and sits at table 1.
$\phi_1 = \theta_1$ where $\theta_1 \sim G_0$, $K = 1$, $n = 1$, $n_1 = 1$
for $n = 2, \ldots,$

customer n sits at table $\begin{cases} k & \text{with prob } \frac{n_k}{n-1+\alpha} \quad \text{for } k = 1 \ldots K \\ K+1 & \text{with prob } \frac{\alpha}{n-1+\alpha} \quad \text{(new table)} \end{cases}$

if new table was chosen **then** $K \leftarrow K + 1$, $\theta_{K+1} \sim G_0$ **endif**
set ϕ_n to θ_k of the table k that customer n sat at; set $n_k \leftarrow n_k + 1$
endfor

Clustering effect: New students entering a school join clubs in proportion to how popular those clubs already are ($\propto n_k$). With some probability (proportional to α), a new student starts a new club.

(Aldous, 1985)