# A Short Review on Bayesian Analysis

- Binomial, Multinomial, Normal, Beta, Dirichlet
- Posterior mean, MAP, credible interval, posterior distribution
- Gibbs sampling

# 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 $\theta$ by producing a probability distribution for $\theta$.

# 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 $\theta$.

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 $\theta$ can be obtained from the posterior distribution $\pi(\theta|x)$. For example, if one wants a

- point estimate of $\theta$, 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 $\theta$, we can report the 95% credible interval, which is a region with 0.95 posterior probability. So 95% credible interval (1.2, 3.5) means that

$$\mathbb{P}(\theta \in (1.2, 3.5)) = 0.95$$

where $\mathbb{P}$ corresponds to the posterior distribution over $\theta$.

For a 95% confidence interval (1.2, 3.5), we CANNOT say that "(1.2, 3.5) covers the true $\theta$ with prob 95%."

- The mode estimate is often referred to as the MAP (maximum a posteriori) estimate, which is the solution of

$$\max_{\theta} \log \pi(\theta|x) = \max_{\theta} \Big[ \log p(x|\theta) + \log \pi(\theta) \Big].$$

Note that MAP is also the solution of

$$\min_{\theta} \Big[ -\log p(x \mid \theta) - \log \pi(\theta) \Big].$$

So MAP is related to the regularization approach, with the penalty term being $(-\log)$ of the prior.

- Your first resistance to Bayesian inference may be the prior choice. 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 $\theta$, where $\theta$ is the probability of getting a head.
- Without any information about the coin, we can put a uniform prior on $\theta$, that is,

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

Next we calculate the posterior distribution of $\theta$ 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 $\theta$,

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

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

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 $\theta$ 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. Note that when $n$ gets large, the prior gets "washed out".

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

The extra counts–one for head and one for tail–are often called the pseudo-counts. Having pseudo-counts is appealing in cases where $\theta$ is likely to take extreme values close to $1$ or $0$. For example, to estimate $\theta$ 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 $\theta$; in fact, uniform is a special case of Beta. Suppose the prior on $\theta$ is Beta$(\alpha, \beta)$,

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

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 $\alpha$ heads and $\beta$ 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 $\theta$ 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 $\mathbb{R}^k$,

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

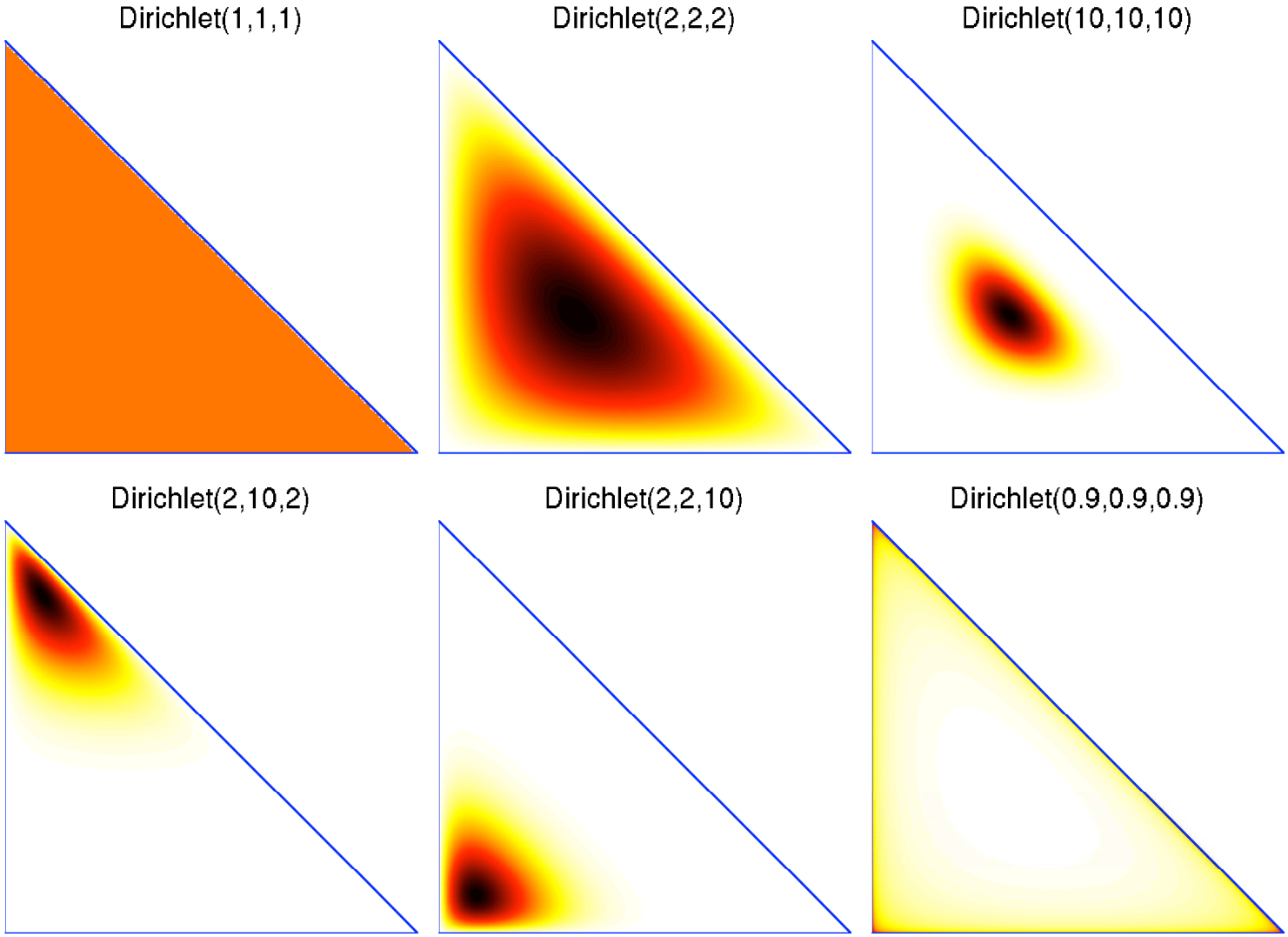
- A Dirichlet distribution on $\mathcal{S}$, $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, is an extension of the Beta distribution, with a density function given by

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

  The Dirichlet distributions are conjugate priors for Multinomial models.

# Dirichlet Distributions

Examples of Dirichlet distributions over $\mathbf{p} = (p_1, p_2, p_3)$ which can be plotted in 2D since $p_3 = 1 - p_1 - p_2$:



## A Normal Example

Assume $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The parameters here are $(\theta, \sigma^2)$ 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 $\theta$, the conjugate prior is normal.

$$\begin{aligned} \bar{X} \mid \theta, \sigma^2 &\sim \mathsf{N}(\theta, \sigma^2/n) \\ \theta &\sim \mathsf{N}(\mu_0, \tau_0^2), \end{aligned}$$

then $\theta|\sigma^2, \mathbf{X} \sim \mathsf{N}(\mu, \tau^2)$ where

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

For the scale parameter $\sigma^2$, 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 Inv$\chi^2$ 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

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

$v_0$ pseudo samples and each contributes $s_0^2$ into RSS.

## Gibbs Sampling 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 a 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.

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.

At each stage, we draw $\theta_i$ from the conditional distribution $\pi(\theta_i|\boldsymbol{\theta}_{[-i]}, \text{Data})$ where $\boldsymbol{\theta}_{[-i]}$ denotes the $(K-1)$ parameters except $\theta_i$.

Why Gibbs samplers?In many cases the conditional distribution of $\pi(\boldsymbol{\theta}|\text{Data})$ is not of closed form, while all those conditionals are.

## Revisit the Gaussian Mixture Model

- EM for MAP
- Collapsed Gibbs sampling
- Chinese restaurant process, nonparametric clustering

# A Gaussian Mixture Model

Suppose the data $x_1, x_2, \ldots, x_n$ 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

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

# EM for MAP

The MAP estimate is defined to be

$$\hat{\theta} = \arg\max_{\theta} P(\mathbf{x} \mid \theta)\pi(\theta) = \arg\max_{\theta} \Big[\sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z} \mid \theta)\Big]\pi(\theta).$$

We can use the EM algorithm:

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

Recall 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 the current estimate of the parameter $\theta_0$

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

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

For MAP, the E-step is the same; the M-step is slightly different:

- Without the Beta prior, we would update $w$ by $\gamma_+/n$. But with the Beta prior on $w$, we need to add the pseudo-counts.
- Similarly, without the prior, we would update $\mu_1$ by

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

but with the prior, we would update $\mu_1$ by a weighted average of the value above and the prior mean for $\mu_1$.

## Gibbs Sampling from the Posterior Distribution

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

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 $Z_i$? Bernoulli

$$\begin{aligned} P(Z_i = 1 \mid \mathbf{x}, \mathbf{Z}_{[-i]}, \text{ others}) &\propto P(Z_i = 1|\theta) \times P(x_i \mid Z_i = 1, \theta) \\ &= wp(x_i \mid \mu_1, \sigma_1^2) \end{aligned}$$

- How to sample $\mu_1$? Normal

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

For a general Gaussian mixture model with $K$ components, we have the prior on the mixing weights as

$$\mathbf{w} = (w_1, \ldots, w_K) \sim \text{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}),$$

and again, normal prior on $\mu_k$'s and InvGa prior on $\sigma_k^2$'s.

The Gibbs sampler iterates the following steps:

1. Draw $Z_i$ from a Multinomial distribution for $i = 1, \ldots, n$;
2. Draw $\mathbf{w}$ from Dirichlet;
3. Draw $\mu_1, \ldots, \mu_K$ from Normal;
4. Draw $\sigma_1^2, \ldots, \sigma_K^2$ from InvGa.

## Collapsed Gibbs sampling

The mixing weights $\mathbf{w}$ is used in sampling $Z_i$'s:

$$P(Z_i = k \mid \mathbf{x}, \mathbf{z}_{[-i]}, \theta) \propto P(x_i \mid Z_i = k, \theta) P(Z_i = k \mid \mathbf{z}_{[-i]}, \theta),$$

where the 2nd term is equal to $P(Z_i = k \mid \mathbf{z}_{[-i]}, \mathbf{w}) = P(Z_i = k|\mathbf{w})$, i.e., when conditioning on $\mathbf{w}$, $Z_i$'s are independent.

Let's eliminate $\mathbf{w}$ from the parameter list, i.e., integrate over $\mathbf{w}$. So we need to compute

$$P(Z_i = k \mid \mathbf{z}_{[-i]}) = \frac{P(z_1, \ldots, z_{i-1}, Z_i = k, z_{i+1}, \ldots, z_n)}{P(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)}$$

Note that $Z_i$'s are exchangeable (its meaning will be made clearly in class). So it suffices to compute the sampling distribution for the last observation.

$$P(Z_n = k \mid z_1, \ldots, z_{n-1}) = \frac{P(Z_n = k, z_1, \ldots, z_i)}{P(z_1, \ldots, z_i)}$$

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

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

$$= \frac{n_k+\alpha/K}{n-1+\alpha}$$

where $n_k = \#\{j : z_j = k, j = 1 : (n-1)\}$.

The Collapsed Gibbs sampler iterates the following steps:

1. Draw $Z_i$ from a Multinomial$(\gamma_{i1}, \ldots, \gamma_{iK})$, for $i = 1, \ldots, n$ where

$$\gamma_{ik} \propto \frac{n_k^{(i)} + \alpha/K}{n - 1 + \alpha} \times P(x_i \mid \mu_k, \sigma_k^2).$$

2. No need to sample $\mathbf{w}$ from Dirichlet;

3. Draw $\mu_1, \ldots, \mu_K$ from Normal;

4. Draw $\sigma_1^2, \ldots, \sigma_K^2$ from InvGa.

## Infinite Many Clusters

$K =$ A Large Value $(> n)$?

At the $t$-th iteration of the algorithm, there must be some empty clusters.

Let's re-label the clusters: $1$ to $K^*$ being the non-empty clusters (of course, the value of $K^*$ changes from iteration to iteration), and the remaining are empty ones.

- Update $\{\mu_k, \sigma_k^2\}_{k=1}^{K^*}$ from posterior (Normal/InvGa)
- Update $\{\mu_k, \sigma_k^2\}_{k=K^*+1}^{K}$ from prior (Normal/InvGa)

- Update $Z_i$ from a Multinomial$(\gamma_{i1}, \ldots, \gamma_{iK})$, where

$$\gamma_{ik} \propto \frac{n_k^{(i)} + \alpha/K}{n - 1 + \alpha} \times P(x_i \mid \mu_k, \sigma_k^2).$$

$Z_i$ may start a new cluster, i.e., $Z_i > K^*$.

It doesn't matter which value $Z_i$ takes from $(K^* + 1, \ldots, K)$, since we can always label this new cluster as the $(K^* + 1)$th cluster, and then immediately update $(\mu_{K*+1}, \sigma^2_{K^*+1})$ based on the corresponding posterior (conditioning on $x_i$).

What't the chance that $Z_i$ starts a new cluster?

$$\begin{aligned}
\sum_{k=K^*+1}^{K} \gamma_{ik} &= \frac{\alpha/K}{n-1+\alpha} \sum_{k=K^*+1}^{K} P(x_i \mid \mu_k, \sigma_k^2) \\
&= \frac{\alpha\frac{K-K^*}{K}}{n-1+\alpha} \left[ \frac{1}{K-K^*} \sum_{k=K^*+1}^{K} P(x_i \mid \mu_k, \sigma_k^2) \right] \\
&\to \frac{\alpha}{n-1+\alpha} \iint P(x_i \mid \mu, \sigma^2)\pi(\mu, \sigma^2) d\mu d\sigma^2,
\end{aligned}$$

when $K \to \infty$, where

$$\iint P(x_i \mid \mu, \sigma^2)\pi(\mu, \sigma^2) d\mu d\sigma^2 = m(x_i)$$

is the integrated (wrt to our prior $\pi$) likelihood for a sample $x_i$.

# Clustering with Chinese Restaurant Process (CRP)

A mixture model

$$X_i \mid Z_i = k \sim P(\cdot \mid \mu_k, \sigma_k^2), \quad i = 1:n.$$

Prior on $Z_1, \ldots, Z_n$ and $(\mu_k, \sigma_k^2)_{k=1}^K$ (known as the Chinese Restaurant Process)

- $Z_1 = 1$ and $(\mu_1, \sigma_1^2) \sim \pi(\cdot)$
- for $i \geq 1$, suppose $Z_1, \ldots, Z_i$ form $m$ clusters with size $n_1, \ldots, n_m$, then

$$\begin{aligned} P(Z_{i+1} = k \mid Z_1, \ldots, Z_i) &= \frac{n_k}{i+\alpha}, \quad k = 1, \ldots, m; \\ P(Z_{i+1} = m+1 \mid Z_1, \ldots, Z_i) &= \frac{\alpha}{i+\alpha}, \quad (\mu_{m+1}, \sigma_{m+1}^2) \sim \pi(\cdot). \end{aligned}$$

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 $(\mu_1, \sigma_1^2) \sim \pi(\cdot)$ and $X_1 \sim P(\cdot \mid \mu_1, \sigma_1^2)$.
- Loop over $i = 1, \ldots, n-1$: suppose the previous $i$ samples form $m$ clusters with cluster-specific parameters $(\mu_k, \sigma_k^2)_{k=1}^m$; then

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

If $Z_{i+1} = m+1$, generate $(\mu_{m+1}, \sigma_{m+1}^2) \sim \pi(\cdot)$. Then generate

$$X_{i+1} \sim P(\cdot \mid \mu_k, \sigma_k^2).$$

## 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.

# Exchangeability of $Z_i$'s

In CRP, the labels $Z_i$'s are generated sequentially, but in fact they are exchangeable (up to a permutation of the cluster labels – labels should start from 1, 2, ...)

$$\begin{aligned} P(11122) &= P(12121) = P(12221) \\ &= \frac{\alpha^2(2!)(1!)}{(1+\alpha)(2+\alpha)\cdots(4+\alpha)}. \end{aligned}$$

In general, suppose $z_1, \ldots, z_n$ form $m$ clusters with size $n_1, \ldots, n_m$, then

$$P(z_1, \ldots, z_n) = \frac{\alpha^m (n_1 - 1)! \cdots (n_m - 1)!}{\prod_{i=2}^{n} (i - 1 + \alpha)}.$$

So the order of $z_i$'s doesn't matter; what matters is the partition of the $n$ samples implied by $z_i$'s.

# Posterior Sampling for Clustering with CRP

Same as the Gibbs sampler we have derived for $K \to \infty$. At the $t$th iteration, repeat the following:

- Suppose $\mathbf{Z}_{[-i]}$ form $K$ clusters, labeled from 1 to $K$, of size $n_k^{(i)}$.

  Sample $Z_i$ from a Multinomial with

$$P(Z_i = k) \propto \frac{n_k^{(i)}}{n-1+\alpha} \times P(x_i \mid \mu_k, \sigma_k^2), \quad k = 1, \ldots, K$$

$$P(Z_i = K+1) \propto \frac{\alpha}{n-1+\alpha} m(x_i),$$

  where $m(x_i) = \iint P(x_i|\mu, \sigma^2)\pi(\mu, \sigma^2) d\mu d\sigma^2$.

- Update $\{\mu_k, \sigma_k^2\}$ from posterior (Normal/InvGa)

- The exchangeability of $Z_i$'s plays an important role in the algorithm. Where we use this property?
- The marginal likelihood $m(\cdot)$ is easy to compute if the prior $\pi(\mu, \sigma^2)$ is conjugate, otherwise, we need to figure a way to compute $m(\cdot)$.
- For other MCMC algorithms, check Neal (2000); for Variational Bayes, check Blei and Jordan (2004).
- The "ugly" side: labeling issue.

# Non-parametric Bayesian (NB) Models

- The finite mixture model

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

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

- Alternatively,

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

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

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

- A NB approach for clustering

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

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

where $\text{DP}(\alpha, G_0)$ denotes a Dirichlet Process with a scale (precision) parameter $\alpha$ 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 $(\theta_1, \ldots, \theta_n)$,

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

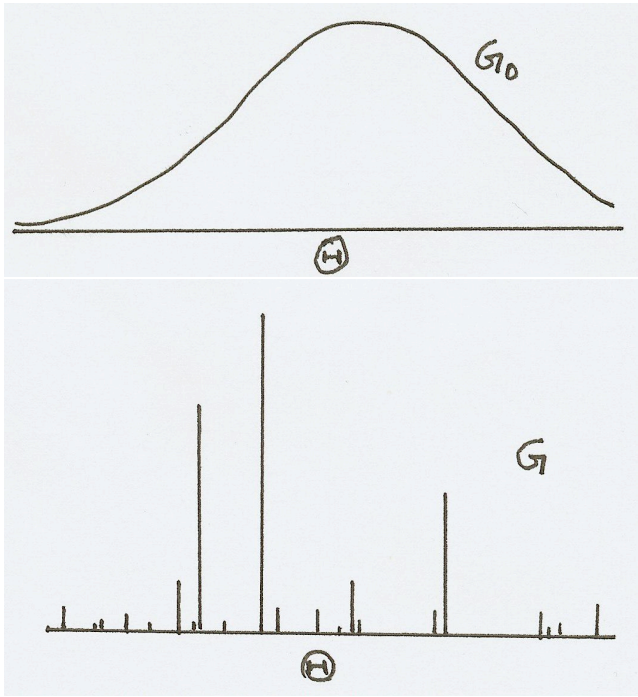
is the Chinese restaurant process (CRP).

# Dirichlet Process

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

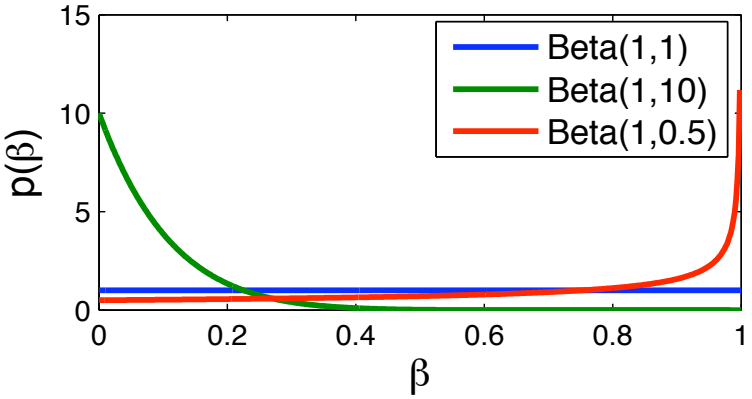
$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot)$$

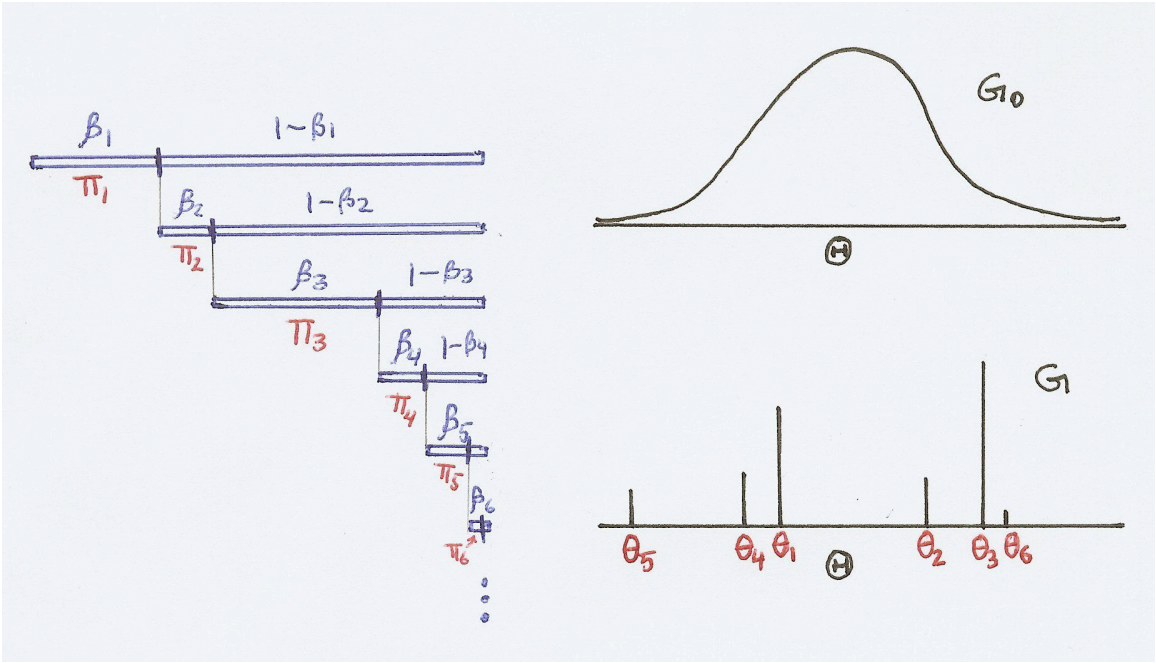
where $\theta_k \sim G_0(\cdot)$, $\sum_{k=1}^{\infty} \pi_k = 1$,

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

and

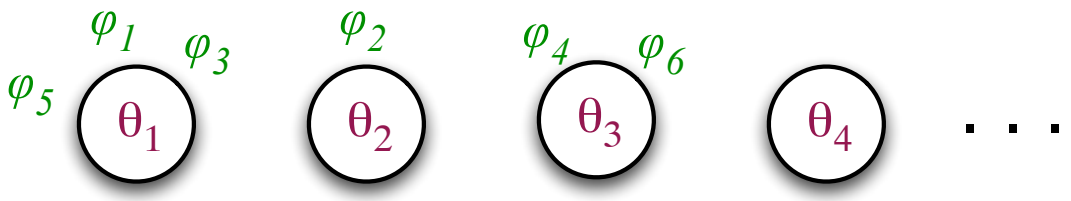
$$\beta_k \sim \text{Beta}(\cdot|1, \alpha)$$




(Sethuraman, 1994)

# Chinese Restaurant Process




**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 $\phi_n$ to $\theta_k$ of the table $k$ that customer $n$ sat at; set $n_k \leftarrow n_k + 1$
**endfor**

The resulting conditional distribution over $\phi_n$:

$$\phi_n | \phi_1, \ldots, \phi_{n-1}, G_0, \alpha \sim \frac{\alpha}{n-1+\alpha} G_0(\cdot) + \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \delta_{\theta_k}(\cdot)$$