Advanced Methods in Statistics (Fall 2016)

J.P.Kim

Dept. of Statistics

Finally modified at September 20, 2016

Preface & Disclaimer

This note is a summary of the lecture Advanced Methods in Statistics (326.521) held at Seoul National University, Fall 2016. Lecturer was Jaeyong Lee, and the note was summarized by J.P.Kim, who is a Ph.D student. Note that in here, each section means each lecture or topic. There are few textbooks and references in this course, which are following.

- *An Introduction to Statistical Learning, James, Witten, Hastie & Tibshirani, 2013.*
- *Machine Learning: a probabilistic perspective, K.P.Murphy, 2012.*
- *Applied Bayesian Modelling, P.Congdon, 2014.*
- *The elements of Statistical Learning, Friedman, Hastie & Tibshirani, 2001.*

Also I referred to following books when I write this note. The list would be updated continuously.

- *Mathematical Statistics: Basic ideas and selected topics, Vol. I., 2nd edition, P.Bickel & K.Doksum, 2007.*
- *Linear Models in Statistics, Rencher & Schaalje, 2008.*
- *A first course in Bayesian statistical methods, P.D.Hoff, 2009.*
- *Statistical decision theory and Bayesian Analysis, James O. Berger, 2013.*
- *Introduction to Nonparametric Regression, K.Takezawa, 2005.*
- *Applied Multivariate Statistical Analysis, R.Johnson & D.Wichern, 2007.*

If you want to correct typo or mistakes, please contact to: `joonpyokim@snu.ac.kr`

Contents

1 Introduction to Bayesian Statistics

1.1 Basic concepts of Bayesian Inference

In statistical inference, we use the parametric model to obtain the information about nature. In here, parameter denotes *the nature state*, and we *observe* the observation. Often we use θ for parameter, and x for the observation. Our model is about 'the distribution of observation when the parameter is given,' i.e., $x|\theta \sim p(x|\theta)$.

Frequentists assume that the nature state is nonrandom, so parameter is unknown constant. However, Bayesians want to represent *all of uncertain information* as a probability distribution. Before we perform data analysis, we have the *belief* or *priori information* about the nature. For example, if we toss a coin, we believe that probability for head is near to 0.5. In Bayesian statistics we define such information as a form of probability distribution, and it is called a **prior distribution**, or in short, a **prior**.

There are three basic elements of Bayesian inference. First is a **prior distribution**, $\theta \sim \pi(\theta)$. Next is an **observation**. We often use the model $x|\theta \sim p(x|\theta)$. Important thing is that *every information used for estimating θ should be from the observation.* Observation is also called as a likelihood. Finally, after observing the data, we can obtain an updated information about θ, which is called a **posterior distribution**. Note that posterior distribution is $\theta|x \sim \pi(\theta|x)$, which means the distribution of the parameter after observing x.

Main goal of Bayesian inference is to obtain posterior distribution. Then how to obtain it? Next theorem makes it possible.

Proposition 1.1 (Bayes' rule)**.** *Posterior distribution is obtained as*

$$\pi(\theta|x) = \frac{p(x,\theta)}{m(x)} = \frac{\pi(\theta)p(x|\theta)}{m(x)} \propto \pi(\theta)p(x|\theta),$$

where $m(x) = \int \pi(\theta)p(x|\theta)d\theta$.

Example 1.2. Suppose that we toss a pin n times. There are two possibilities in the result: one is $\nearrow$, and the other is $\perp$. Let θ be a probability for $\nearrow$. Then our model is $x|\theta \sim Bin(n,\theta)$. Since we don't have any priori information about the simulation, we can use the noninformative prior, $\theta \sim U(0,1)$. Then posterior distribution is obtained as

$$\pi(\theta|x) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and hence $\theta|x \sim Beta(x+1, n-x+1)$.

Remark 1.3. Often we use a *point estimator* of θ, rather than posterior distribution. It would be one point summarization of the distribution. Posterior mean, median or MAP (*Maximum a Posteriori*) is frequently used.

Definition 1.4. *Interval* (L, U) *such that*

$$P(L < \theta < U|x) = 1 - \alpha$$

is called a $100(1-\alpha)\%$ *credible set or confidence interval.*

Remark 1.5. Note that frequentist and Bayesian view for confidence region is different. In Bayesian view, we can say that a **probability** the parameter contained in region is $1-\alpha$. Also note that, Bayesians represent all of uncertainty as a probability distribution.

1.2 Bayesian hypothesis testing

Consider the model $x|\theta \sim f(x|\theta)$ and hypotheses

$$H_0 : \theta = \theta_0 \; vs \; H_1 : \theta = \theta_1.$$

It can be also written as

$$H_0 : x \sim f(x|\theta_0) \; vs \; H_1 : x \sim f(x|\theta_1).$$

First, let $\pi_0 = \pi(H_0)$ and $\pi_1 = \pi(H_1)$ be prior probabilities of H_0 and H_1, respectively. Clearly $0 < \pi_0, \pi_1 < 1$ and $\pi_0 + \pi_1 = 1$ should be held. Then posterior probability of each hypothesis is obtained as

$$\pi(H_0|x) = \frac{\pi_0 f(x|H_0)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)}$$

and

$$\pi(H_1|x) = 1 - \pi(H_0|x) = \frac{\pi_1 f(x|H_1)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)}.$$

Now hypothesis procedure is that, we support H_1 if $\pi(H_1|x)/\pi(H_0|x)$ is sufficiently large.

Definition 1.6 (Bayesian Testing)**.** *Let*

$$\frac{\pi(H_1|x)}{\pi(H_0|x)} = \frac{\pi_1}{\pi_0} \cdot \frac{f(x|H_0)}{f(x|H_1)}$$

be a posterior odds. Then

$$B_{10} := \frac{f(x|H_0)}{f(x|H_1)}$$

is called a ***Bayes factor****. Then clearly we get*

$$\textit{posterior odds} = \textit{prior odds} \times \textit{Bayes factor}.$$

Note that Bayes factor has similar role to the "likelihood ratio" in frequentist's inference.

Remark 1.7. We additionally need a criterion to judge whether B_{10} is large or not. One can use *Jeffreys' criterion* following.

$\log_{10} B_{10}$	B_{10}	Strength of evidence for H_1
$0 \sim 1/2$	$1 \sim 3.2$	Not worth a bare mention
$1/2 \sim 1$	$3.2 \sim 10$	substantial
$1 \sim 2$	$10 \sim 100$	strong
> 2	> 100	decisive

Example 1.8. Consider a pin example again. Suppose that we want to test

$$H_0 : \theta = \frac{1}{2} \; vs \; H_1 : \theta = \frac{2}{3}.$$

Assume that we observed $x = 7$. Then Bayes factor is

$$B_{10} = \frac{\binom{10}{7}(2/3)^7(1/3)^3}{\binom{10}{7}(1/2)^7(1/2)^3} = 2.2197.$$

Also note that

$$\pi(H_1|x) = \frac{\pi_1 f(x|H_1)}{\pi_0 f(x|H_0) + \pi_1 f(x|H_1)} = \frac{\pi_1 B_{10}}{\pi_0 + \pi_1 B_{10}} = 0.6894.$$

Thus we may support H_0.

2 Bayesian Computation

In Bayeisan analysis, every information about θ is summarized in posterior. However, in many cases, characteristic of posterior distribution (such as moments or quantiles) is difficult to find. Then we should estimate or approximate moments or quantiles. In this section, we introduce some estimation methods and random generation algorithm, which is known as Markov Chain Monte Carlo (MCMC), and give some references.

2.1 Monte Carlo

If one wants to estimate

$$I = Ef(X) = \int f(x)g(x)dx,$$

where $g(x)$ is a density and $X \sim g(x)$, then we can use a random sample $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} g$ and estimate I as

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Note that by SLLN, $\hat{I} \xrightarrow[n\to\infty]{a.s} I$. Also, it can be easily verified that

$$\hat{se}(\hat{I}) = \sqrt{\frac{\nu}{n}}$$

where

$$\nu = \widehat{Var} f(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(f(X_i) - \hat{I}\right)^2.$$

Note that error rate is $O(n^{-1/2}$, which does not depend on the dimension of integration. It is a strong point when we compare Monte Carlo method with other numerical approaches, such as trapezoidal rule or quadrature rules.

Example 2.1. Let $X|\theta \sim N(\theta, 1)$ and $\theta \sim Cauchy(0, 1)$. Then by Bayes' rule posterior is obtained as

$$\pi(\theta|x) \propto \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2},$$

so posterior mean (=Bayes estimator) is

$$E(\theta|x) = \frac{\int \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2}}{\int \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2}}.$$

To estimate this, we may use a random sample $\theta_i \overset{i.i.d.}{\sim} N(x,1)$, and use

$$\hat{E}(\theta|x) = \frac{\sum_{i=1}^n \frac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^n \frac{1}{1+\theta_i^2}}.$$

Importance sampling

Suppose that we want to estimate

$$I = \int f(x)g(x)dx = E[f(X)]$$

where $X \sim g$. However, random generation from g is difficult problem, so we can't handle it. Instead, there is $\pi \approx g$ such that we can easily obtain random numbers from π. Then using π we can estimate I as a *weighted average of observed values*. Precisely, let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} \pi$. Then since

$$\int f(x)g(x)dx = \int \frac{f(x)g(x)}{\pi(x)}\pi(x)dx = E\left[\frac{f(X)g(X)}{\pi(X)}\right]$$

we can estimate I as

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^n w_i f(X_i)$$

$$\hat{I}_2 = \frac{\sum_{i=1}^n w_i f(X_i)}{\sum_{i=1}^n w_i}$$

where

$$w_i = \frac{g(X_i)}{\pi(X_i)}.$$

Note that $\hat{I}_2$ is biased estimator for I but has some advantages compared to $\hat{I}_1$. If we use $\hat{I}_2$, we don't have to know normalizing constant of g. To use $\hat{I}_1$, we may estimate $\int g(x)dx$ again to make g a density function.

Or we can do a further step in importance sampling, which is known as *sampling importance sampling (SIR)*. Note that, Monte Carlo can be treated as approximation of pdf to *empirical cdf*, $g(x) \approx n^{-1}\sum \delta_{x_i}$, where δ_{x_0} is Dirac measure. Importance sampling is an approximation $g(x) \approx \sum x_i \delta_{x_i}$. Sampling importance resampling is, resample $X_1^*, \cdots, X_n^*$ from the distribution estimated by (X_i, w_i), i.e., $g(x) \approx \sum w_i \delta_{x_i}$, and estimate I as

$$\hat{I}_3 = \frac{1}{n}\sum_{i=1}^n f(x_i^*).$$

2.2 Markov Chain Monte Carlo

Monte Carlo is useful when θ is high-dimensional, because its error may not depend on the dimension. However, as dimension goes high, it becomes difficult to generate a random number. Importance sampling is one alternative for this, but it may not be a good answer, because if dimension is high, $\pi \approx g$ becomes a hard goal ("curse of dimension"). In this sense, Markov Chain Monte Carlo can be other alternative.

Gibbs Sampling

Gibbs sampling is an algorithm that generates random numbers using previous numbers. Let $f(x_1, x_2, \cdots, x_p)$ be a density function, and denote $x_{-i} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)$.

Algorithm 1 (systematic sweep) Gibbs Sampling

1: Initialize $(X_1^{(0)}, X_2^{(0)}, \cdots, X_p^{(0)})$.
2: **for** $t = 1, 2, 3, \cdots$ **do**
3: Sample $X_1^{(t)} \sim f_{X_1|X_{-1}}(\cdot|X_2^{(t-1)}, \cdots, X_p^{(t-1)})$.
4: Sample $X_2^{(t)} \sim f_{X_2|X_{-2}}(\cdot|X_1^{(t)}, X_3^{(t-1)}, \cdots, X_p^{(t-1)})$.
$\vdots$
5: Sample $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)}, \cdots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \cdots, X_p^{(t-1)})$
$\vdots$
6: Sample $X_p^{(t)} \sim f_{X_p|X_{-p}}(\cdot|X_1^{(t)}, \cdots, X_{p-1}^{(t)})$.
7: **end for**

Note that sequence of random variables $(X^{(0)}, X^{(1)}, X^{(2)}, \cdots)$ obtained via Gibbs sampling becomes a *Markov Chain whose stationary distribution is* $f(x_1, \cdots, x_p)$. Furthermore, under some conditions, sample moments or quantiles of $(X^{(0)}, X^{(1)}, X^{(2)}, \cdots)$ converge to those of population.

We can consider *random sweep* Gibbs sampling algorithm.

Algorithm 2 random sweep Gibbs Sampling

1: Initialize $(X_1^{(0)}, X_2^{(0)}, \cdots, X_p^{(0)})$.
2: **for** $t = 1, 2, 3, \cdots$ **do**
3: Choose $j \in \{1, 2, \cdots, p\}$ randomly.
4: Sample $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t)} \cdots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \cdots, X_p^{(t-1)})$ and let $X_i^{(t)} = X_i^{(t-1)}$ $\forall i \neq j$.
5: **end for**

Example 2.2. In this example, our goal is to apply Gibbs sampling to find random numbers

with stationary distribution

$$\pi = N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

First initialize $(x_1^{(0)}, x_2^{(0)})$. Then update x_1 and x_2 as following.

(1) $x_1^{(t)} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2^{(t-1)} - \mu_2), \sigma_1^2(1-\rho^2)\right)$

(2) $x_2^{(t)} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1^{(t)} - \mu_1), \sigma_2^2(1-\rho^2)\right)$

Finally we can estimate population characteristics. For example,

$$\int h(x_1, x_2)\pi(dx_1, dx_2) \approx \frac{1}{n}\sum_{t=1}^{n} h(x_1^{(t)}, x_2^{(t)}),$$

$$\text{and } P(X_1 \geq 0, X_2 \geq 0) \approx \frac{1}{n}\sum_{t=1}^{n} I(x_1^{(t)} \geq 0, x_2^{(t)} \geq 0).$$

Metropolis-Hastings Algorithm

Goal of Metropolis-Hastings Algorithm is to find Markov chain with stationary distribution $\pi(\theta)$, where π is given. Note that if kernel $K(x, dy)$ of Markov chain is determined, Markov chain is also determined. For this, kernel K should satisfy

$$\int \pi(dx)K(x, dy) = \pi(dy).$$

It is known that if K *satisfies detained balance condition*

$$\pi(dx)K(x, dy) = \pi(dy)K(y, dx)\ \forall x, y \in S,$$

then π is a stationary distribution of $K(x, dy)$. Now our question is: *for arbitrary proposal kernel $q(x, y)$, can we find a kernel K which satisfies detailed balance condition?* If we choose α satisfying

$$\alpha(x, y)\pi(x)q(x, y) = \pi(y)q(y, x)$$

and define *Metropolis-Hastings kernel*

$$K(x, dy) = \alpha(x, y)q(x, y)dy + (1 - \alpha(x))\delta_x(dy)$$

where

$$\alpha(x) = \int \alpha(x,y)q(x,y)dy,$$

then K satisfies detailed-balance condition. We can make this as choosing

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right).$$

Algorithm 3 Metropolis-Hastings Algorithm

1: Initialize $x^{(0)}$.
2: **for** $t = 1, 2, 3, \cdots, m$ **do**
3: Draw $x \sim q(x^{(t-1)}, \cdot)$ and $u \sim U(0,1)$ independently.
4: Obtain acceptance rate

$$\alpha(x^{(t-1)}, x) = \min\left(1, \frac{\pi(x)q(x, x^{(t-1)})}{\pi(x^{(t-1)}q(x^{(t-1)}, x)}\right).$$

5: Define

$$x^{(t)} = \begin{cases} x & \text{if} \quad u \le \alpha(x^{(t-1)}, x) \\ x^{(t-1)} & \text{if} \quad u > \alpha(x^{(t-1)}, x) \end{cases}.$$

6: **end for**

Convergence diagnostics

To judge whether generated random numbers converged or not, we may use *time series plot, cumulative sum plot, ACF plot, log-likelihood or log-posterior plot.* If first some samples did not converge and may not represent the stationary distribution, then we would *not* use such samples, and only use the numbers generated later. Such approach is called *burn-in.* Or, if ACF is too high to say that samples are independent, then we may use not all of samples, but some of them, until ACF becomes small. Such approach is called *thinning.*

Now we will see *effective sample size.* Note that for a stationary time series y_t and

$$H_m = \frac{1}{m}\sum_{t=1}^{m} h(y_t),$$

$$\sqrt{m}(H_m - Eh(y)) \xrightarrow[n\to\infty]{d} N(0, \nu^2)$$

holds for

$$\nu^2 = \sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

under some conditions. If there are m MCMC samples with ACF ρ_k, *effective sample size* is

defined as

$$M = \frac{m}{1 + 2\sum_{k=1}^{\infty} \rho_k}.$$

Note that sample mean based on m MCMC samples has the same variance with sample mean based on M i.i.d. samples.

3 Hierarchical Models

Consider a data obtained as following: There are 71 different groups rats. We are interested in the rate of *endometrial stromal polyps* in the different groups. The number of rats varies from group to group. Let y_i be the number of tumors in group i. Our model is based on binomial model,

$$y_i|\theta_i \overset{indep}{\sim} Bin(n_i, \theta_i).$$

If we assume that $\theta_1 = \theta_2 = \cdots = \theta_{71}$, then the problem becomes simple, but the assumption has many problems. First, can we believe $\theta_1 = \cdots = \theta_{71}$ holds? Under such assumption, for

$$\bar{p} = \frac{\sum x_i}{\sum n_i},\ \hat{p}_i = \frac{x_i + 0.5}{n_i + 1},$$

$$z_i = \frac{\hat{p}_i - \bar{p}}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n_i}}$$

should be $N(0,1)$ distributed approximately. In here, we used $\hat{p}_i = (x_i + 0.5)/(n_i + 1)$ because in the data $x_i = 0$ is observed for many i. However, if we plot histogram or Q-Q plot, we can easily verify that such assumption is not reasonable.

We can solve this problem by using prior distribution,

$$\theta_i \sim Beta(\alpha, \beta).$$

Even though $\theta_1 = \cdots = \theta_{71}$ did not make sense, it is reasonable to suppose that θ_is have similar values. it can be reflected on the assumption that *each θ_i was drawn from the same distribution.* There are other 70 groups that we observed, and using the observations we can estimate hyperparameters α and β. We can use MLE or MME. In this lecture, MME was used. Such approach is called *empirical Bayes.*

Although it seems very reasonable, there are some problems. First, if we wanted to infer $\theta_1, \theta_2, \cdots$ as well as θ_{71}, then we should repeat such procedures in many times. In addition, we used estimated values α and β *just as we have the exact information about prior*, even if there is uncertainty from estimation of hyperparameters.

One alternative we can use is **hierarchical model**. In here, our model is:

$$y_i|\theta_i \sim Bin(n_i, \theta_i),\ i = 1, 2, \cdots, 71,$$

$$\theta_i \sim Beta(\alpha, \beta),\ i = 1, 2, \cdots, 71.$$

We employee Gamma *hyperprior* in this lecture,

$$\alpha \sim Gamma(a_\alpha, b_\alpha),\ \beta \sim Gamma(a_\beta, b_\beta).$$

Since we don't have information about α and β, we may set $Var(\alpha)$ and $Var(\beta)$ large. For example, we can suppose that

$$E(\alpha) = \frac{a_\alpha}{b_\alpha} = 1,\ Var(\alpha) = \frac{a_\alpha}{b_\alpha^2} = 10^3,$$

and similar for β.

Before finishing, there are some remarks. First, in hierarchical model, prior can be interpreted as a part of model, and hyper-prior has a role of prior. Note that Bayesian employees prior distribution to represent *uncertainty of the nature*. Next, we can also think further-hyperprior of hyperparemeter, but it is not justified because (i) There are no data more that would be used, and (ii) uncertainty of hyperprior distribution may affects less than prior, so it may not be needed. Finally, if we choose a prior, there can be two approaches: (i) *mimic* "non-information," or (ii) reflect information that we have. One example of using latter view is an empirical Bayes approach on hyperprior.

4 Dirichlet Process

4.1 Definition and properties of Dirichlet process

Consider a Bayesian nonparametric model

$$X_1, X_2, \cdots, X_n | F \overset{i.i.d.}{\sim} F.$$

Note that in parametric model, the model is

$$X_1, X_2, \cdots, X_n | \theta \overset{i.i.d.}{\sim} f(x|\theta)$$

and we assume prior for θ on Θ. However, in nonparametric case, there is no assumption about F, so we should consider a prior distribution on

$$\mathcal{M}(\mathbb{R}) := \{F : F \text{ is a probability distribution on } \mathbb{R}\}.$$

Then it becomes *a probability distribution on the set of probability distribution.*

Definition 4.1 (Dirichlet process)**.** *Let α be a finite measure on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is a complete separable metric space, and $\mathcal{B} = \mathcal{B}(\mathcal{X})$ is a Borel σ-field on $\mathcal{X}$. Let P be a random probability measure satisfying*

$$(P(B_1), P(B_2), \cdots, P(B_k)) \sim Dirichlet(\alpha(B_1), \alpha(B_2), \cdots, \alpha(B_k)),$$

for any measurable partition $\{B_i\}_{i=1}^k$ on $\mathcal{X}$, then P is said to be distributed according to the Dirichlet Process, and denoted as

$$P \sim DP(\alpha).$$

Definition 4.2 (Alternative view: from Wikipedia)**.** *Let H be a base probability distribution on $\mathcal{X}$ and $\alpha > 0$. Then the Dirichlet process $DP(H, \alpha)$ is a stochastic process whose sample path is a probability distribution over $\mathcal{X}$ and the following holds: For any measurable finite partition of $\mathcal{X}$, say $\{B_i\}_{i=1}^n$,*

$$X \sim DP(H, \alpha) \Rightarrow (X(B_1), X(B_2), \cdots, X(B_n)) \sim Dirichlet(\alpha H(B_1), \cdots, \alpha H(B_n)).$$

Remark 4.3. Note that, if P is a random probability measure, then all of $P(B)$ becomes random

variable. $P \sim DP$ means that P is a probability measure on $\mathcal{X}$, which is random. "Dirichlet" process is termed because *every finite dimensional distribution* of P,

$$(P(B_1), \cdots, P(B_n)),$$

is Dirichlet distributed. It can be also interpreted as following briefly: We defined a "marginal" (finite-dimensional) distribution of P as a Dirichlet distribution, to define a "joint" feature of P. (Remark that: it is not a formal term!) Because P and partition B_i's should satisfy that

$$0 \leq P(B_i) \leq 1 \text{ and } \sum_{i=1}^{n} P(B_i) = 1,$$

it is natural to think Dirichlet distribution. Note that P is itself a probability measure, or distribution.

There are some important properties of DP. Proof is given in *Advanced Bayesian Analysis* course.

Proposition 4.4 (Conjugacy, or Posterior distribution)**.** *Let $P \sim DP(\alpha)$ and the model be*

$$X_1, \cdots, X_n | P \overset{i.i.d.}{\sim} P.$$

Then posterior distribution of P also becomes DP. In fact,

$$P|X_1, \cdots, X_n \sim DP\left(\alpha + \sum_{j=1}^{n} \delta_{X_i}\right),$$

where δ_c denotes Dirac measure.

Proposition 4.5 (Marginal property, Blackwell & MacQueen (1973))**.** *Let $P \sim DP(\alpha)$ and $X_1, X_2, \cdots | P \overset{i.i.d.}{\sim} P$. Then marginal $(X_1, X_2, \cdots)$ forms **Pólya urn sequence**, i.e.,*

$$\begin{aligned} X_1 &\sim \frac{\alpha}{\alpha(\mathcal{X})} \\ X_{n+1}|X_1, \cdots, X_n &\sim \frac{\alpha + \sum_{i=1}^{n} \delta_{x_i}}{\alpha(\mathcal{X}) + n}. \end{aligned}$$

(See Pólya urn scheme or Chinese restaurant process in section 4.2.)

Proposition 4.6 (Sethuramen representation)**.** *Let*

$$\theta_1, \theta_2, \cdots \overset{i.i.d.}{\sim} Beta(1, \alpha(\mathcal{X}))$$
$$Y_1, Y_2, \cdots \overset{i.i.d.}{\sim} \frac{\alpha}{\alpha(\mathcal{X})}.$$

Consider a ***stick-breaking process***

$$\begin{aligned} p_1 &= \theta_1 \\ p_2 &= \theta_2(1-\theta_1) \\ &\vdots \\ p_n &= \theta_n \prod_{i=1}^{n-1}(1-\theta_i) \\ &\vdots \end{aligned}$$

which makes $\sum_{n=1}^{\infty} p_i = 1$*. Then*

$$P = \sum_{i=1}^{\infty} p_i \delta_{Y_i} \sim DP(\alpha).$$

Remark 4.7. Note that both p_i and Y_i are random. Also, from Sethuramen representation, we can find that if $P \sim DP(\alpha)$, P is a discrete probability measure *w.p. 1.*

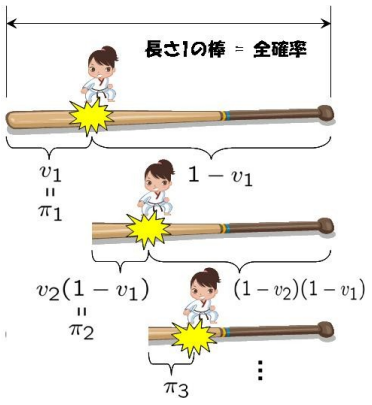

Figure 1: Stick-breaking process. Image from `http://d.hatena.ne.jp/b3s/20081021/1224569652`

Proposition 4.8 (Gamma process representation)**.** *Let* α *be a finite measure on* $[0, \infty)$ *and define a cumulative "measure" function* $A(t) = \alpha[0, t]$*. Also, let* $S(t) \sim GP(A(t), 1)$ *be a Gamma*

process. Then,

$$F(t) = \frac{S(t)}{S(\infty)} \sim DP(\alpha)$$

holds.

Also, following is known about the support of Dirichlet process.

$$supp(DP_\alpha) = \{P : supp(P) \subseteq supp(\alpha)\}.$$

4.2 Description

Contents of this section are mostly quoted from Wikipedia.

Pólya urn scheme

There is a description of Dirichlet process, which is called a Pø'lya urn scheme. Imagine that we start with an urn filled with $\alpha(\mathcal{X})$ black balls. Then we proceed as follows:

1. Each time we need an observation, we draw a ball from the urn.
2. If the ball is *black*, we generate a new (non-black) color uniformly, label a new ball this color, drop the **new ball** into the urn along with the ball we drew, and return the color we generated.
3. Otherwise, (i.e., if the ball is non-black, label a new ball with the color of the ball we drew, drop the **new ball** into the urn along with the ball we drew, and return the color we observed.

This scheme is related to Pólya urn sequence, described in proposition 4.5. Consider following situation. First, draw X_1 from the distribution $\alpha/\alpha(\mathcal{X})$. Next, for $n > 1$, with probability $\frac{\alpha}{\alpha + n - 1}$ draw X_n from $\alpha/\alpha(\mathcal{X})$, which is corresponding to *return the color generated newly*. Or, with probability $\frac{n_x}{\alpha + n - 1}$, set $X_n = x$, where n_x is the number of previous observations $X_j, j < n$ such that $X_j = x$. This procedure is corresponding to *return the color we observed*.

Chinese restaurant process

Similar description for Dirichlet process is the one so-called Chinese restaurant process. Imagine an infinitely large restaurant containing an infinite number of tables, and able to serve an infinite number of dishes. The restaurant in question operates a somewhat unusual seating policy

whereby new diners are seated either at a currently occupied table with probability proportional to the number of guests already seated there, or at an empty table with probability proportional to a constant. Guests who sit at an occupied table must order the same dish as those currently seated, whereas guests allocated a new table are served a new dish at random. The distribution of dishes after n guests are served is a sample drawn as described above.

4.3 Applications

First application we see is an estimation of cdf. Consider the model

$$P \sim DP(\alpha),\ X_1, \cdots, X_n | P \overset{i.i.d.}{\sim} P.$$

Now let the loss function be

$$L(F, G) = \int (F(t) - G(t))^2 dt.$$

We can easily find that Bayes estimator

$$\hat{F}^B(t) = \arg\min_G \mathbb{E}^\pi \left[(F - G)^2(t) | X_1, \cdots, X_n \right]$$

is obtained as

$$\hat{F}^B(t) = \mathbb{E}^\pi \left[F(t) | X_1, \cdots, X_n \right] = \int F(t) \pi(dF | X_1, \cdots, X_n).$$

Note that by definition of DP, for given t,

$$F(t) | X_1, \cdots, X_n \sim Beta(\alpha(t) + nF_n(t), \alpha(\mathbb{R}) - \alpha(t) + n(1 - F_n(t))),$$

because

$$(P((-\infty, t]), P((t, \infty))) \sim Dirichlet(\alpha((-\infty, t]) + \sum \delta_{x_i}(-\infty, t], \alpha((t, \infty)) + \sum \delta_{x_i}(t, \infty)).$$

In here, F_n denotes an empirical cdf. Then denoting $\bar{\alpha}(t) = \alpha(t)/\alpha(\mathbb{R})$, we get

$$\hat{F}^B(t) = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} \bar{\alpha}(t) + \frac{n}{\alpha(\mathbb{R}) + n} F_n(t).$$

Thus, $\alpha(\mathbb{R})$ may be interpreted as "prior size", which gives a "reliability of prior we have."

Next application is Dirichlet Process mixture model. Consider the model

$$\begin{aligned} X_i|\theta_i &\overset{indep}{\sim} f(x_i|\theta_i),\ i = 1, 2, \cdots, n \\ \theta_i|P &\overset{i.i.d.}{\sim} P \\ P &\sim DP. \end{aligned}$$

Then it can be obtained that

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} \sum_{j=1}^{\infty} P_j f(x|\theta_j)$$

where P_j's are from Sethuramen representation

$$P = \sum_{j=1}^{\infty} P_j \delta_{\theta_j} \sim DP.$$

There are some applications of this model. First, we can estimate a density $f(x|\theta)$ of continuous random variable, even though P from DP is discrete. Also, we can fit complex data using simple model f, if we employee mixture model. Furthermore, if we gather observations with the same θ_i, we can find a **clustering** algorithm without pre-determination of the number of cluster.

4.4 Sampling algorithm from the posterior

Escobar (1994) and Escobar & West (1995) suggested MCMC algorithm for Dirichlet Process mixture model. It employees a *collapsed Gibbs sampler*. First, consider the model

$$\begin{aligned} y_i|\theta_i &\overset{indep}{\sim} F(\cdot|\theta_i),\ i = 1, 2, \cdots, n \\ \theta_i|G &\overset{i.i.d.}{\sim} G,\ i = 1, 2, \cdots, n \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where $\alpha > 0$ is a constant and G_0 is a given distribution. Note that joint distribution of y, θ, and G is obtained as

$$p(dy, d\theta, dG) = \prod_{i=1}^{n} F(dy_i|\theta_i) \prod_{i=1}^{n} G(d\theta_i) DP(dG|G_0, \alpha).$$

However, since G is from infinite dimensional space, it is difficult to sample G. Thus, we will integrate out ("collapse") G. Then we get

$$\begin{aligned}
p(dy, d\theta) &= \int_G p(dy, d\theta, dG) \\
&= \int_G prod_{i=1}^n F(dy_i|\theta_i) \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha) \\
&= \prod_{i=1}^n F(dy_i|\theta_i) \underbrace{\int_G \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha)}_{=\text{marginal of } \theta} \\
&= \prod_{i=1}^n F(dy_i|\theta_i) p(d\theta|G_0, \alpha) \\
&= \prod_{i=1}^n f(y_i|\theta_i) dy_i \cdot p(d\theta|G_0, \alpha)
\end{aligned}$$

where density $f(\cdot|\theta_i)$ of $F(\cdot|\theta_i)$ exists. In here, $p(\theta|G_0, \alpha)$ denotes a joint distribution of Pólya sequence. Note that, $G(d\theta_i)$ is in fact $G(d\theta_i|G)$, and hence

$$\int_G \prod_{i=1}^n G(d\theta_i) DP(dG|G_0, \alpha)$$

is a marginal distribution of θ_i. Using this we can find a Gibbs sampler algorithm from $[\theta_i|\theta_{-i}, y]$. First note that Pólya sequence has a conditional distribution

$$p(\theta_i|\theta_{-i}, G_0, \alpha) = \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(d\theta_i) + \frac{\alpha}{n-1+\alpha} G_0(d\theta_i).$$

Now by definition of δ_{θ_j} we get

$$\begin{aligned}
p(d\theta_i|\theta_{-i}, y) &\propto f(y_i|\theta_i) p(d\theta_i|\theta_{-i}, G_0, \alpha) \\
&= f(y_i|\theta_i) \left[\frac{1}{n-1+\alpha} \sum_{j \neq i} \delta_{\theta_j}(d\theta_i) + \frac{\alpha}{n-1+\alpha} G_0(d\theta_i) \right] \\
&\propto \sum_{j \neq i} f(y_i|\theta_j) \delta_{\theta_j}(d\theta_i) + \alpha f(y_i|\theta_i) G_0(d\theta_i) \\
&\propto \sum_{j \neq i} q_{ij} \delta_{\theta_j}(d\theta_i) + r_i H_i(d\theta_i),
\end{aligned}$$

where

$$q_{ij} = \frac{f(y_i|\theta_j)}{\sum_{l \neq i} f(y_i|\theta_l) + \alpha \int f(y_i|\theta) dG_0(\theta)}$$

$$r_i = \frac{\alpha \int f(y_i|\theta) dG_0(\theta)}{\sum_{l \neq i} f(y_i|\theta_l) + \alpha \int f(y_i|\theta) dG_0(\theta)}$$

$$H_i(d\theta_i) = \frac{f(y_i|\theta) dG_0(\theta)}{\int f(y_i|\theta) dG_0(\theta)}.$$

It implies that, to implement this algorithm, we should know the integrated value of $\int f(y_i|\theta) dG_0(\theta)$.

Normalizing constants are determined to make them satisfy

$$\sum_{j \neq i} q_{ij} + r_i = 1.$$

So we obtain the algorithm: iterating sampler from

$$p(d\theta_i|\theta_{-i}, y) = \sum_{j \neq i} q_{ij} \delta_{\theta_j}(d\theta_i) + r_i H_i(d\theta_i).$$