BIO 249 Lab 1: Introduction to Bayesian Statistics

Sep 5, 2016

Recommended reading: Notes Lectures 1-2, Gelman Chapter 1, Bayes Refresher PDF.

Every course in bayesian statistics inevitably begins by contrasting frequentist and bayesian paradigms. When I took my first statistics class, I was not aware of these statistical "religions" and naively thought what I was learning was the only way of thinking about statistics. The term "frequentist" did not enter my vocabulary until I discovered bayesian statistics. The underlying difference between these two paradigms is that frequentists assume that the data are random and parameters are fixed, while bayesians assume the data are fixed and parameters can be summarized probabilistically (note that this is not the same as assuming that parameters are random - more on that below). In very simple settings such as a linear regression, these two paradigms typically lead to similar results. One does not have to strictly adhere to either paradigm, as each method has its time and place depending on the problem. Bayesian statistics have increased in popularity due to the modern advancements in computing power [citation needed]. One of my favourite quotes I found on this subject is as follows:

> *"A frequentist is a person whose long-run ambition is to be wrong 5% of the time.*
> *A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule."*

Lorenzo is motivating the use of bayesian statistics through the likelihood principle (all relevant information from the data about a parameter θ is contained in the likelihood function). The likelihood principle is maintained in the bayesian paradigm because we only consider the data that was observed, whereas frequentists consider data that could have been observed (through the sampling distribution). We see examples from Berry (1987), and Slide 51 from the Lecture 1 notes.

The use of priors on parameters, and subsequently the entire bayesian framework, is motivated through De Finetti's Theorem. We first define a sequence of random variables $(y_1, y_2, \ldots)$ as **infinitely exchangeable** if, for all n, the joint probability function $p\left(y_1, y_2, \ldots, y_n\right)$ is invariant to the permutation of the indices. That is,

$$p\left(y_1, y_2, \ldots, y_n\right) = p\left(y_{\pi_1}, y_{\pi_2}, \ldots, y_{\pi_n}\right)$$

for any permutation $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$ of $(1, 2, \ldots, n)$. Note that iid random variables are infinitely exchangeable, but infinite exchangeability is broader than being iid (e.g. if $(x_1, x_2, \ldots)$ are iid and x_0 is a non-trivial random variable independent of the others, then $(x_0 + x_1, x_0 + x_2, \ldots)$ is infinitely exchangeable but not iid). Any sequence where x_n depends on x_{n-1} will be an example of a sequence that is not infinitely exchangeable.

De Finetti's Theorem states (simplified) that a sequence of random variables $(y_1, y_2, \ldots)$ is infinitely exchangeable iff, for all n,

$$p\left(y_1, y_2, \ldots, y_n\right) = \int \prod_{i=1}^{n} p\left(y_i|\theta\right) p\left(\theta\right) d\theta$$

The theorem states that if we have an infinitely exchangeable sequence $\left(y_1, y_2, \ldots, y_n\right)$, then

1. There exists a parameter θ with distribution $p\left(\theta\right)$.
2. There exists a likelihood $p\left(\boldsymbol{y}|\theta\right)$.
3. The data are conditionally iid given θ.

I looked at various sources trying to highlight the importance of this theorem, and decided on the following summary (using Bernoulli trials) from Genesis of Bayesian Analysis in 1999. We know that a mixture of conditionally iid Bernoulli trials is exchangeable. The difficult and important result is that any exchangeable sequence must be a mixture of conditionally iid Bernoulli trials. This result characterizes exchangeable binary sequences. In other words, if you assess your uncertainty about a sequence of trials in such a way that labeling of the trials is irrelevant, then you are equivalently asserting the existence of a random variable Z such that given $Z = z$, the trials are iid $\text{Bernoulli}\left(z\right)$. But the orthodox (frequentist) view is precisely that the trials are iid $\text{Bernoulli}\left(\theta\right)$, where the unknown θ represents the long-run relative frequency of successes. The connection between the subjective view and the orthodox view is complete if we simply add a probability distribution over θ. That is, if we treat the long-run relative frequency itself as a random variable (called Z, say). A subjective assessment of a sequence of binary trials which takes the primitive exchangeability assumption is equivalent to placing a prior distribution on the long-run success probability within the orthodox view. *Bayesian analysis rests on this formulation. That is, we accept the orthodox view of iid trials, but we place probability distributions over the parameters of the sampling distributions.*

A typical bayesian formulation involves

1. **Prior distribution** $\pi\left(\boldsymbol{\theta}\right)$: a summary of prior belief on the parameters (vaguely expecting a horse).
2. **Likelihood** $\mathcal{L}\left(\boldsymbol{y}|\boldsymbol{\theta}\right)$: from the distribution of the data that we collect, based on unknown parameters (catching a glimpse of a donkey).
3. **Posterior distribution** $p\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$: a summary of updated belief on the topic, given the prior and likelihood (believes in seeing a mule).

These three functions are linked together through Bayes' Theorem:

$$f\left(\boldsymbol{\theta}|\boldsymbol{y}\right) = \frac{f\left(\boldsymbol{y}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)}{f\left(\boldsymbol{y}\right)} = \frac{f\left(\boldsymbol{y}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)}{\int_{\boldsymbol{\theta}} f\left(\boldsymbol{y}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}} \propto f\left(\boldsymbol{y}|\boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)$$

I will typically write this relationship as

$$p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \propto \mathcal{L}\left(\boldsymbol{y}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right)$$

Note that we only need to solve for $p(\boldsymbol{\theta}|\boldsymbol{y})$ up to a proportionality constant. If $p(\boldsymbol{\theta}|\boldsymbol{y})$ has a recognizable kernel, we can immediately write down its distribution in closed form. One reason bayesian inference is heavily reliant on computing is that we often do not have closed form solutions for $p(\boldsymbol{\theta}|\boldsymbol{y})$, and must rely on iterative numerical methods.

If the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ is in the same family as the prior $\pi(\boldsymbol{\theta})$, then the posterior and prior are called conjugate distributions, with the prior called a **conjugate prior**. Some examples include

1. Beta-Bernoulli (Bernoulli data, Beta prior)

$$\begin{aligned}
p(\theta|\boldsymbol{y}) &\propto \mathcal{L}(\boldsymbol{y}|\theta)\,\pi(\theta) \\
&= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{\alpha-1+\sum_{i=1}^{n} y_i}(1-\theta)^{\beta-1+n-\sum_{i=1}^{n} y_i} \\
&\sim \text{Beta}\left(\alpha+\sum_{i=1}^{n} y_i, \beta+n-\sum_{i=1}^{n} y_i\right)
\end{aligned}$$

2. Poisson-Gamma (Poisson data, Gamma prior)

$$\begin{aligned}
p(\lambda|\boldsymbol{y}) &\propto \mathcal{L}(\boldsymbol{y}|\lambda)\,\pi(\lambda) \\
&= \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda) \\
&\propto \exp(-n\lambda)\lambda^{\sum_{i=1}^{n} y_i}\lambda^{\alpha-1}\exp(-\beta\lambda) \\
&= \lambda^{\alpha-1+\sum_{i=1}^{n} y_i}\exp[-(n+\beta)\lambda] \\
&\sim \text{Gamma}\left(\alpha+\sum_{i=1}^{n} y_i, \beta+n\right)
\end{aligned}$$

After obtaining the posterior distribution, we may be interested in prediction. In that case, we would like the **posterior predictive distribution**

$$\begin{aligned}
p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) &= \int_{\boldsymbol{\theta}} \mathcal{L}(\tilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \mathcal{L}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y})\,p(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \mathcal{L}(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\boldsymbol{y})\,d\boldsymbol{\theta} \text{ by independence of } \tilde{\boldsymbol{y}} \text{ and } \boldsymbol{y}
\end{aligned}$$

Problem 1 Let y denote the number of successes observed in a study of n trials such that $y|\theta \sim \text{Binom}\left(n, \theta\right).$ Assume that the prior on θ is $\text{U}\left(0,1\right).$ Now consider a new trial $\tilde{y}$, which is independent of the first n and has the same distribution. What is the posterior predictive probability that $\tilde{y}$ will equal 1?

Solution 1 We are looking for

$$\begin{aligned}
P\left(\tilde{y}=1|y\right) &= \int_\theta P\left(\tilde{y}=1|\theta\right)p\left(\theta|y\right)d\theta \\
&= \int_\theta \theta p\left(\theta|y\right)d\theta \\
&= E\left(\theta|y\right) \text{ (the posterior mean)}
\end{aligned}$$

The posterior distribution is

$$\begin{aligned}
p\left(\theta|y\right) &\propto \mathcal{L}\left(y|\theta\right)\pi\left(\theta\right) \\
&= \binom{n}{y}\theta^y\left(1-\theta\right)^{n-y}(1) \\
&\propto \theta^y\left(1-\theta\right)^{n-y} \\
&\sim \text{Beta}\left(y+1, n-y+1\right)
\end{aligned}$$

Then the posterior predictive probability that $\tilde{y}=1$ is $\frac{y+1}{y+1+n-y+1} = \frac{y+1}{n+2}.$

Problem 2 Consider $y_1, \ldots y_n|\theta \sim \text{Gamma}\left(\alpha, \theta\right),$ where α is known. Assume that the prior on θ is $\text{Gamma}\left(\alpha_0, \beta_0\right).$ Find the posterior distribution $p\left(\theta|\boldsymbol{y}\right).$ Let $\tilde{\boldsymbol{y}}$ be an m-vector of new observations. Find the posterior predictive distribution $p\left(\tilde{\boldsymbol{y}}|\boldsymbol{y}\right).$

Solution 2

$$\begin{aligned}
p\left(\theta|\boldsymbol{y}\right) &\propto \mathcal{L}\left(\boldsymbol{y}|\theta\right)\pi\left(\theta\right) \\
&= \prod_{i=1}^n \frac{\theta^\alpha}{\Gamma\left(\alpha\right)}y_i^{\alpha-1}\exp\left(-\theta y_i\right)\frac{\beta_0^{\alpha_0}}{\Gamma\left(\alpha_0\right)}\theta^{\alpha_0-1}\exp\left(-\beta_0\theta\right) \\
&\propto \theta^{n\alpha}\exp\left(-\theta\sum_{i=1}^n y_i\right)\theta^{\alpha_0-1}\exp\left(-\beta_0\theta\right) \\
&= \theta^{n\alpha+\alpha_0-1}\exp\left[-\left(\sum_{i=1}^n y_i+\beta_0\right)\theta\right] \\
&\sim \text{Gamma}\left(\alpha_0+n\alpha, \beta_0+\sum_{i=1}^n y_i\right)
\end{aligned}$$

$$\begin{aligned}
p\left(\tilde{\boldsymbol{y}}|\boldsymbol{y}\right) &= \int_\theta p\left(\tilde{\boldsymbol{y}}|\theta\right)p\left(\theta|\boldsymbol{y}\right)d\theta \\
&= \int_\theta \prod_{i=1}^m \frac{\theta^\alpha}{\Gamma\left(\alpha\right)}\tilde{y}_i^{\alpha-1}\exp\left(-\theta\tilde{y}_i\right)\frac{\left(\beta_0+\sum_{i=1}^n y_i\right)^{\alpha_0+n\alpha}}{\Gamma\left(\alpha_0+n\alpha\right)}\theta^{\alpha_0+n\alpha-1}\exp\left[-\left(\beta_0+\sum_{i=1}^n y_i\right)\theta\right]d\theta \\
&= \frac{\prod_{i=1}^m \tilde{y}_i^{\alpha-1}}{\Gamma\left(\alpha\right)^m}\frac{\left(\beta_0+\sum_{i=1}^n y_i\right)^{\alpha_0+n\alpha}}{\Gamma\left(\alpha_0+n\alpha\right)}\int_\theta \theta^{m\alpha}\exp\left(-\theta\sum_{i=1}^m \tilde{y}_i\right)\theta^{\alpha_0+n\alpha-1}\exp\left[-\left(\beta_0+\sum_{i=1}^n y_i\right)\theta\right]d\theta
\end{aligned}$$

$$= \frac{\prod_{i=1}^{m} \tilde{y}_i^{\alpha-1}}{\Gamma(\alpha)^m} \frac{\left(\beta_0 + \sum_{i=1}^{n} y_i\right)^{\alpha_0+n\alpha}}{\Gamma(\alpha_0+n\alpha)} \int_\theta \theta^{(m+n)\alpha+\alpha_0-1} \exp\left[-\left(\sum_{i=1}^{m} \tilde{y}_i + \beta_0 + \sum_{i=1}^{n} y_i\right)\theta\right] d\theta$$

Note that the integrand

$$\theta^{(m+n)\alpha+\alpha_0-1} \exp\left[-\left(\sum_{i=1}^{m} \tilde{y}_i + \beta_0 + \sum_{i=1}^{n} y_i\right)\theta\right]$$

is a kernel of a Gamma $\left(\alpha_0 + (m+n)\alpha, \beta_0 + \sum_{i=1}^{m} \tilde{y}_i + \sum_{i=1}^{n} y_i\right)$, whose pdf integrates to 1 over θ. The integrand is missing the normalizing constant (with respect to θ) $\frac{\left(\beta_0+\sum_{i=1}^{m} \tilde{y}_i+\sum_{i=1}^{n} y_i\right)^{\alpha_0+(m+n)\alpha}}{\Gamma[\alpha_0+(m+n)\alpha]}$. Then we can conclude

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \frac{\prod_{i=1}^{m} \tilde{y}_i^{\alpha-1}}{\Gamma(\alpha)^m} \frac{\left(\beta_0 + \sum_{i=1}^{n} y_i\right)^{\alpha_0+n\alpha}}{\Gamma(\alpha_0+n\alpha)} \frac{\Gamma[\alpha_0+(m+n)\alpha]}{\left(\beta_0 + \sum_{i=1}^{m} \tilde{y}_i + \sum_{i=1}^{n} y_i\right)^{\alpha_0+(m+n)\alpha}}$$

Problem 3 Suppose that in each individual of a large population there is a pair of genes, each of which can be either x or X, that controls eye colour: those with xx have blue eyes, while heterozygotes (those with Xx or xX) and those with XX, have brown eyes. The proportion of blue-eyed individuals is p^2 and of heterozygotes is $2p(1-p)$, where $0 < p < 1$. Each parent transmits one of its own genes to the child; if a parent is a heterozygote, the probability that it transmits the gene of type X is $\frac{1}{2}$. Assuming random mating, show that among brown-eyed children of brown-eyed parents, the expected proportion of heterozygotes is $\frac{2p}{1+2p}$. Suppose Judy, a brown-eyed child of brown-eyed parents, marries a heterozygote, and they have n children, all brown-eyed. Find the posterior probability that Judy is a heterozygote.

Solution 3 We are first looking for

$$\begin{aligned} & P(\text{HetZ}|\text{Brown Eye} \cap \text{Brown Eye Parents}) \\ = & \frac{P(\text{Brown Eye} \cap \text{Brown Eye Parents}|\text{HetZ}) P(\text{HetZ})}{P(\text{Brown Eye} \cap \text{Brown Eye Parents})} \\ = & \frac{P(\text{Brown Eye Parents}|\text{HetZ}) P(\text{HetZ})}{P(\text{Brown Eye} \cap \text{Brown Eye Parents})} \text{ heterozygotes have brown eyes} \\ = & \frac{\frac{P(\text{HetZ}|\text{Brown Eye Parents}) P(\text{Brown Eye Parents})}{P(\text{HetZ})} P(\text{HetZ})}{P(\text{Brown Eye} \cap \text{Brown Eye Parents})} \\ = & \frac{P(\text{HetZ}|\text{Brown Eye Parents}) P(\text{Brown Eye Parents})}{P(\text{Brown Eye}|\text{ Brown Eye Parents}) P(\text{Brown Eye Parents})} \\ = & \frac{P(\text{HetZ}|\text{Brown Eye Parents})}{P(\text{Brown Eye}|\text{ Brown Eye Parents})} \end{aligned}$$

We know that $P(\text{xx}) = p^2, P(\text{XX}) = (1-p)^2, P(\text{Xx}) = 2p(1-p)$. In the numerator, we have

$$\begin{aligned} P(\text{HetZ}|\text{Brown Eye Parents}) &= P(\text{HetZ}|\text{Brown Eye Parents} \cap \text{Parents XXXX}) P(\text{Parents XXXX}) \\ &\quad +2P(\text{HetZ}|\text{Brown Eye Parents} \cap \text{Parents XXXx}) P(\text{Parents XXXx}) \\ &\quad +P(\text{HetZ}|\text{Brown Eye Parents} \cap \text{Parents XxXx}) P(\text{Parents XxXx}) \\ &= 0(1-p)^2(1-p)^2 + 2\left(\frac{1}{2}\right)(1-p)^2 2p(1-p) + \frac{1}{2}(2p)(1-p)2p(1-p) \\ &= 2p(1-p)^3 + 2p^2(1-p)^2 \end{aligned}$$

$$
\begin{aligned}
&= 2p\left(1-p\right)^2\left(1-p+p\right) \\
&= 2p\left(1-p\right)^2
\end{aligned}
$$

In the denominator, we have

$$
\begin{aligned}
P\left(\text{Brown Eye}|\text{Brown Eye Parents}\right) &= P\left(\text{Brown Eye}|\text{Brown Eye Parents}\cap\text{Parents XXXX}\right)P\left(\text{Parents XXXX}\right) \\
&\quad +2P\left(\text{Brown Eye}|\text{Brown Eye Parents}\cap\text{Parents XXXx}\right)P\left(\text{Parents XXXx}\right) \\
&\quad +P\left(\text{Brown Eye}|\text{Brown Eye Parents}\cap\text{Parents XxXx}\right)P\left(\text{Parents XxXx}\right) \\
&= 1\left(1-p\right)^2\left(1-p\right)^2+2\left(1\right)\left(1-p\right)^2 2p\left(1-p\right)+\frac{3}{4}\left(2p\right)\left(1-p\right)2p\left(1-p\right) \\
&= \left(1-p\right)^4+4p\left(1-p\right)^3+3p^2\left(1-p\right)^2 \\
&= \left(1-p\right)^2\left[\left(1-p\right)^2+4p\left(1-p\right)+3p^2\right] \\
&= \left(1-p\right)^2\left(1-2p+p^2+4p-4p^2+3p^2\right) \\
&= \left(1-p\right)^2\left(1+2p\right)
\end{aligned}
$$

Then $P\left(\text{HetZ}|\text{Brown Eye}\cap\text{Brown Eye Parents}\right)=\frac{2p}{1+2p}$, as desired.

The marginal (prior) probability that Judy is a heterozygote is $\frac{2p}{1+2p}$, since we are told she has brown eyes and brown-eyed parents. Note that Judy is either a heterozygote or has genes XX. We want to update this probability with the information that she had n brown-eyed kids with another heterozygote (call data_n as the event Brown Eye$\cap$Brown Eye Parents$\cap n$ brown-eyed kids with HetZ). Before proceeding, we first need to know

$$P\left(\text{data}_1|\text{Judy HetZ}\right)=1-P\left(\text{child has blue eyes}\right)=\frac{3}{4}$$

Then by independence, $P\left(\text{data}_n|\text{Judy HetZ}\right)=\left(\frac{3}{4}\right)^n$. We are looking for

$$
\begin{aligned}
P\left(\text{Judy HetZ}|\text{data}\right) &= \frac{P\left(\text{data}|\text{Judy HetZ}\right)P\left(\text{Judy HetZ}\right)}{P\left(\text{data}\right)} \\
&= \frac{P\left(\text{data}|\text{Judy HetZ}\right)P\left(\text{Judy HetZ}\right)}{P\left(\text{data}|\text{Judy Hetz}\right)P\left(\text{Judy HetZ}\right)+P\left(\text{data}|\text{Judy XX}\right)P\left(\text{Judy XX}\right)} \\
&= \frac{\left(\frac{3}{4}\right)^n\frac{2p}{1+2p}}{\left(\frac{3}{4}\right)^n\frac{2p}{1+2p}+\left(1\right)\left(1-\frac{2p}{1+2p}\right)} \\
&= \frac{\left(\frac{3}{4}\right)^n 2p}{\left(\frac{3}{4}\right)^n 2p+1}\to 0\text{ as }n\to\infty
\end{aligned}
$$

This makes sense because if Judy kept having brown-eyed kids, it would become more and more unlikely that she is a heterozygote (otherwise she likely would have had a blue-eyed kid).

BIO 249 Lab 2: Conjugate Priors for Exponential Families

Sep 12, 2016

Recommended reading: Conjugate Priors for Exponential Families (1979), Billingsley Sections 2, 10, 15-17

Review of Probability Theory

Probability theory begins with a foundation in measure theory. One of the most fundamental concepts in Euclidean geometry is that of the measure $m(E)$ of a solid body E in one or more dimensions. In one, two, and three dimensions, we refer to this measure as the length, area, or volume of E respectively. With the advent of analytic geometry, however, Euclidean geometry became re-interpreted as the study of Cartesian products $\mathbb{R}^d$ of the real line $\mathbb{R}$. Using this analytic foundation rather than the classical geometrical one, it was no longer intuitively obvious how to define the measure $m(E)$ of a general subset E of $\mathbb{R}^d$. The standard solution to the problem of measure has been to abandon the goal of measuring *every* subset E of $\mathbb{R}^d$, and instead to settle for only measuring a certain subclass of subsets of $\mathbb{R}^d$, which are then referred to as the measurable sets. For more motivation on measure theory and examples of non-measurable sets, look up information on the Banach-Tarski paradox and Vitali sets.

The class of sets which are measurable should be sufficiently rich. In particular, we want to keep measurability of sets if we perform simple operations like taking the complement or taking (countable) unions and intersections. This leads to the definition of σ-fields.

Defn 1: A class $\mathscr{F}$ of subsets of Ω is called a **σ-field** (or σ-algebra) if it contains Ω and is closed under the formation of complements and countable unions:

(i) $\Omega \in \mathscr{F}$

(ii) $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$

(iii) For every sequence $A_1, A_2, \dots$ of elements of $\mathscr{F}$, the countable union $\bigcup_{k=1}^{\infty} A_n \in \mathscr{F}$

The largest σ-field over Ω is the power set 2^Ω of all subsets of Ω. The smallest σ-field is the trivial one which contains only $\{\emptyset, \Omega\}$. The intersection of an arbitrary family of σ-fields over Ω is again a σ-field over Ω. The σ-field generated by a class $\mathscr{A}$ of subsets of Ω is the intersection of all the σ-fields containing $\mathscr{A}$, and is denoted $\sigma(\mathscr{A})$.

Defn 2: A set function μ on a field $\mathscr{F}$ in Ω is a **measure** if it satisfies these conditions:

1. $\mu(A) \in [0, \infty)$ for $A \in \mathscr{F}$
2. $\mu(\emptyset) = 0$
3. If $A_1, A_2, \dots$ is a disjoint sequence of $\mathscr{F}$-sets and if $\bigcup_{k=1}^{\infty} A_k \in \mathscr{F}$, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k)$$

The measure μ is finite or infinite as $\mu(\Omega) < \infty$ or $\mu(\Omega) = \infty$; it is a probability measure if $\mu(\Omega) = 1$. If $\Omega = A_1 \cup A_2 \cup \ldots$ for some finite or countable sequence of $\mathscr{F}$-sets satisfying $\mu(A_k) < \infty$, then μ is σ-finite.

If Ω is a set and $\mathscr{F}$ is a σ-field on Ω, then the pair $(\Omega, \mathscr{F})$ is called a measurable space. If μ is a measure on $\mathscr{F}$, the triplet $(\Omega, \mathscr{F}, \mu)$ is a measure space. The elements of $\mathscr{F}$ are called measurable sets. If $\mu(A^c) = 0$ for an $\mathscr{F}$-set A, then A is a support of μ. For a finite measure, A is a support if an only if $\mu(A) = \mu(\Omega)$.

Defn 3: Consider the measurable space $(\Omega, \mathscr{F})$ for countable Ω. Define the σ-finite **counting measure** $\nu : \mathscr{F} \to [0, \infty)$ by

$$\nu(A) = |A| \text{ for all } A \in \mathscr{F}$$

where $|A|$ denotes the cardinality of A.

Defn 4: The d-dimensional Lebesgue measure is a measure on a suitable σ-field over $\mathbb{R}^d$ which coincides with the usual geometric length/area/volume $(d = 1, 2, 3)$.

Defn 5: We would like to define the **integral** of a (measurable) function f **with respect to measure** μ. The integral may exist as a real number (in which case we say f is integrable), or may exist as $\pm\infty$, or may not exist at all. The integral is denoted by $\int f d\mu, \int f(x) d\mu(x), \int f(x) \mu(dx)$.

Terminology and Notation from Paper

The paper uses several of the above terminology.

- μ is a σ-finite measure on the Borel sets of $\mathbb{R}^d$
  - the measure space $\left(\mathbb{R}^d, \mathscr{F}, \mu\right)$ where $\mathscr{F}$ is the σ-field generated by the Borel sets of $\mathbb{R}^d$ (this is the smallest σ-field containing all open sets on $\mathbb{R}^d$)
- $\mathscr{X}$ is the interior of the convex hull of the support of μ
  - the support of a finite measure μ satisfies $\mu[\text{supp}(\mu)] = \mu\left(\mathbb{R}^d\right)$
- $\theta \in \mathbb{R}^d$
  - I would normally denote vectors as bolded, but I did not to be consistent with the paper

For our exercises, discrete distributions can be written as probability measures with respect to the counting measure ν. Continuous distributions can be written as probability measures with respect to the Lebesgue measure.

Paper Discussion

This paper presents a general framework for conjugate priors for exponential families. The definition of a conjugate prior in general demands that, for a given observation model (likelihood), and a given class of priors, all corresponding posteriors are again in the prior class. This property is known as closure under sampling. The exponential family is presented in this paper in its measure-theoretic formulation characterized by the cumulative distribution function $P_\theta(x)$. This presentation allows us to encompass discrete, continuous, and mixture distributions with one general form. For the data, define the exponential family $\{P_\theta\}$ of probability measures through the measure μ as

$$dP_\theta\left(x\right) = e^{x\theta - M(\theta)} d\mu\left(x\right), \theta \in \Theta \subseteq \mathbb{R}^d \tag{1}$$

where $M\left(\theta\right) = \log \int e^{\theta x} d\mu\left(x\right)$ (the cumulant function) is obtained by requiring that the integral of (1) is equal to 1. For the prior distribution, consider the family $\{\tilde{\pi}_{n_0,x_0}\}$ of measures through the Lebesgue measure $d\theta$ as (we use Lebesgue measure here since conjugate prior distributions are continuous)

$$d\tilde{\pi}_{n_0,x_0}\left(\theta\right) = e^{n_0 x_0 \theta - n_0 M(\theta)} d\theta, n_0 \in \mathbb{R}, x_0 \in \mathbb{R}^d \tag{2}$$

For (2) to be a proper prior distribution, it must be integrable (we need to be able to normalize the integral in (2) to be equal to 1), meaning $\int_\Theta d\tilde{\pi}_{n_0,x_0}\left(\theta\right) d\theta < \infty$. The normalized prior distribution no longer has the tilde and is denoted by its cumulative distribution function $\pi_{n_0,x_0}\left(\theta\right)$. Theorem 1 gives sufficient or necessary and sufficient conditions on the hyperparameters (n_0, x_0) of a conjugate prior distribution for it to be proper. We can formulate the posterior distribution as

$$\begin{aligned} p\left(\theta|\boldsymbol{x}\right) &\propto \mathcal{L}\left(\boldsymbol{x}|\theta\right)\pi\left(\theta\right) \\ &= e^{\theta \sum_{i=1}^n x_i - nM(\theta)} e^{n_0 x_0 \theta - n_0 M(\theta)} \\ &= e^{(n_0 x_0 + n\bar{x})\theta - (n_0+n)M(\theta)} \\ &\sim \pi_{n_0+n, \frac{n_0 x_0 + n\bar{x}}{n_0+n}} \text{ using the notation from (2)} \end{aligned}$$

This result shows that the posterior distribution is in the same family of distributions as the prior distribution, with new parameters $n^* = n_0 + n$ and $x_0^* = \frac{n_0 x_0 + n\bar{x}}{n_0+n}$. Then, from Theorem 2, we know the posterior mean is

$$E\left[\nabla M\left(\theta\right)|X_1, \ldots, X_n\right] = x_0^* = \frac{n_0 x_0 + n\bar{x}}{n_0 + n} = \frac{n}{n_0+n}\bar{x} + \frac{n_0}{n_0+n}x_0 \tag{3}$$

We see that the posterior mean is a linear combination of the the hyperparameter x_0 and the sample mean $\bar{x}$. The weights on these two elements are proportional to n_0 and the sample size n, so n_0 can be interpreted as a prior sample size. Theorem 3 is concerned with the converse to Theorem 2: can one conclude from the linearity at (3) that θ had a conjugate prior? The answer is yes, if it is assumed that the support of μ is sufficiently rich (the support of μ contains an open interval I_0 in $\mathbb{R}^d$). For a sample size of one, Theorem 3 says that if

$$E\left[\nabla M\left(\theta\right)|X\right] = aX + b$$

then

$$d\tau\left(\theta\right) = ce^{a^{-1}b\theta - a^{-1}(1-a)M(\theta)} d\theta \tag{4}$$

is absolutely continuous with respect to the Lebesgue measure, ensuring integrability. From (3), if we set $n = 1$ and $a = \frac{1}{n_0+1}$ and $b = \frac{n_0}{n_0+1}x_0$, we find that

$$d\tau\left(\theta\right) = ce^{n_0 x_0 - n_0 M(\theta)} d\theta = d\tilde{\pi}_{n_0,x_0}\left(\theta\right)$$

Example 1: Bernoulli Distribution, Beta Prior

The underlying measure here is the counting measure. The probability mass function of the Bernoulli distribution can be written as

$$\begin{aligned} dP_\xi(x) &= \xi^x(1-\xi)^{1-x}\,d\nu(x) \\ &= \exp\left(\log\left[\xi^x(1-\xi)^{1-x}\right]\right)d\nu(x) \\ &= \exp\left[x\log\xi+(1-x)\log(1-\xi)\right]d\nu(x) \\ &= \exp\left[\left(\log\frac{\xi}{1-\xi}\right)x+\log(1-\xi)\right]d\nu(x) \end{aligned}$$

Then

$$\begin{aligned} dP_\theta(x) &= e^{x\theta-\log\left(1+e^\theta\right)}d\nu(x)\,, x=0,1,\theta\in\mathbb{R} \\ \theta &= \log\frac{\xi}{1-\xi} \\ M(\theta) &= -\log(1-\xi)=\log\left(1+e^\theta\right) \end{aligned}$$

We can verify

$$\begin{aligned} M(\theta) &= \log\int_{\mathscr{X}}e^{\theta x}d\nu(x)=\log\sum_{x=0}^{1}e^{\theta x}=\log\left(e^0+e^\theta\right)=\log\left(1+e^\theta\right) \\ \int_{\mathscr{X}}e^{x\theta-\log\left(1+e^\theta\right)}d\nu(x) &= \sum_{x=0}^{1}e^{x\theta-\log\left(1+e^\theta\right)}=e^{-\log\left(1+e^\theta\right)}+e^{\theta-\log\left(1+e^\theta\right)}=\frac{1}{1+e^\theta}+\frac{e^\theta}{1+e^\theta}=1 \end{aligned}$$

The kernel of the Beta prior for ξ can be written as

$$\begin{aligned} d\tilde{\pi}_{\alpha,\beta}(\xi) &= \xi^{\alpha-1}(1-\xi)^{\beta-1}\,d\xi \\ &= \exp\left(\log\left[\xi^{\alpha-1}(1-\xi)^{\beta-1}\right]\right)d\xi \\ &= \exp\left[(\alpha-1)\log\xi+(\beta-1)\log(1-\xi)\right]d\xi \\ d\tilde{\pi}_{\alpha,\beta}(\theta) &= \exp\left[(\alpha-1)\left[\theta-\log\left(1+e^\theta\right)\right]-(\beta-1)\log\left(1+e^\theta\right)\right]\frac{e^\theta}{\left(1+e^\theta\right)^2}d\theta \\ &= \exp\left[(\alpha-1)\theta-(\alpha+\beta-2)\log\left(1+e^\theta\right)+\theta-2\log\left(1+e^\theta\right)\right]d\theta \\ &= \exp\left[\alpha\theta-(\alpha+\beta)M(\theta)\right]d\theta \end{aligned}$$

Here $n_0=\alpha+\beta, x_0=\frac{\alpha}{\alpha+\beta}$. To verify Theorem 2, check that

$$E\left[M'(\theta)\right]=E\left(\frac{e^\theta}{1+e^\theta}\right)=E(\xi)=\frac{\alpha}{\alpha+\beta}=x_0$$

Example 2: Poisson Distribution, Gamma Prior

The underlying measure here is the counting measure. The probability mass function of the Poisson distribution can be written as

$$\begin{aligned} dP_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!} d\nu(x) \\ &= \exp\left[(\log\lambda)x - \lambda\right]\frac{1}{x!}d\nu(x) \\ &= \exp\left[(\log\lambda)x - \lambda\right] d\nu^*(x) \end{aligned}$$

Then

$$\begin{aligned} dP_\theta(x) &= e^{x\theta - e^\theta} d\nu(x), x = 0, 1, \ldots, \theta \in \mathbb{R} \\ \theta &= \log\lambda \\ M(\theta) &= \lambda = e^\theta \end{aligned}$$

We can verify

$$\begin{aligned} M(\theta) &= \log\int_{\mathscr{X}} e^{\theta x} d\nu^*(x) = \log\sum_{x=0}^{\infty} e^{\theta x}\frac{1}{x!} = \log\left[1 + \frac{e^\theta}{1!} + \frac{\left(e^\theta\right)^2}{2!} + \ldots\right] = \log\left(e^{e^\theta}\right) = e^\theta \\ \int_{\mathscr{X}} e^{x\theta - e^\theta} d\nu^*(x) &= \sum_{x=0}^{\infty} e^{x\theta - e^\theta}\frac{1}{x!} = \sum_{x=0}^{\infty} e^{x\log\lambda - \lambda}\frac{1}{x!} = \sum_{x=0}^{\infty}\frac{\lambda^x e^{-\lambda}}{x!} = 1 \end{aligned}$$

The kernel of the Gamma prior for λ can be written as

$$\begin{aligned} d\tilde{\pi}_{\alpha,\beta}(\lambda) &= \lambda^{\alpha-1}e^{-\beta\lambda}d\lambda \\ &= \exp\left[\log\left(\lambda^{\alpha-1}e^{-\beta\lambda}\right)\right]d\lambda \\ &= \exp\left[(\alpha-1)\log\lambda - \beta\lambda\right]d\lambda \\ d\tilde{\pi}_{\alpha,\beta}(\theta) &= \exp\left[(\alpha-1)\theta - \beta e^\theta\right]e^\theta d\theta \\ &= \exp\left[\alpha\theta - \beta M(\theta)\right]d\theta \end{aligned}$$

Here $n_0 = \beta, x_0 = \frac{\alpha}{\beta}$. To verify Theorem 2, check that

$$E\left[M'(\theta)\right] = E\left(e^\theta\right) = E(\lambda) = \frac{\alpha}{\beta} = x_0$$

Example 3: Exponential Distribution, Gamma Prior

The underlying measure here is the Lebesgue measure. The probability density function of the Exponential distribution can be written as

$$\begin{aligned} dP_\lambda(x) &= \lambda \exp(-\lambda x)\, dx \\ &= \exp(\log\lambda - \lambda x)\, dx \\ &= \exp(-\lambda x + \log\lambda)\, dx \\ dP_\theta(x) &= e^{\theta x + \log(-\theta)} dx \end{aligned}$$

Then

$$\begin{aligned} dP_\theta &= e^{\theta x + \log(-\theta)} dx, x \geq 0, \theta \in \mathbb{R}, \theta < 0 \\ \theta &= -\lambda \\ M(\theta) &= -\log\lambda = -\log(-\theta) \end{aligned}$$

We can verify

$$\begin{aligned} M(\theta) &= \log \int_{\mathscr{X}} e^{\theta x} dx = \log \int_0^\infty e^{\theta x} dx = \log \left[\frac{1}{\theta} e^{\theta x}\right]_0^\infty = \log\left(-\frac{1}{\theta}\right) = -\log(-\theta) \\ \int_{\mathscr{X}} e^{\theta x + \log(-\theta)} dx &= \int_0^\infty (-\theta)\, e^{\theta x} dx = -\theta \left[\frac{1}{\theta} e^{\theta x}\right]_0^\infty = -\theta\left(-\frac{1}{\theta}\right) = 1 \end{aligned}$$

The kernel of the Gamma prior for δ can be written as

$$\begin{aligned} d\tilde{\pi}_{\alpha,\beta}(\lambda) &= \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\ &= \exp\left[\log\left(\lambda^{\alpha-1} e^{-\beta\lambda}\right)\right] d\lambda \\ &= \exp\left[(\alpha-1)\log\lambda - \beta\lambda\right] d\lambda \\ d\tilde{\pi}_{\alpha,\beta}(\theta) &= \exp\left[-(\alpha-1) M(\theta) + \beta\theta\right] d\theta \\ &\propto \exp\left[\beta\theta - (\alpha-1) M(\theta)\right] d\theta \end{aligned}$$

Here $n_0 = \alpha - 1, x_0 = \frac{\beta}{\alpha-1}$. To verify Theorem 2, check that

$$E\left[M'(\theta)\right] = E\left(-\frac{1}{\theta}\right) = E\left(\frac{1}{\lambda}\right)$$

where

$$\begin{aligned} E\left(\frac{1}{\lambda}\right) &= \int \frac{1}{\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int \lambda^{\alpha-2} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha-1)}{\beta^{\alpha-1}} \int \frac{\beta^{\alpha-1}}{\Gamma(\alpha-1)} \lambda^{\alpha-2} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta}{\alpha-1} \end{aligned}$$

First Inequality

1. μ_A is a probability measure, since it's non-negative, $\mu_A(\Omega) = 1$, and it satisfies countable additivity.

2. For any set B in the Borel sets of $\mathbb{R}^d$, we have

$$\mu_A(B) = \int_B \frac{1}{\mu(A)} \mathbb{I}_A(z)\, d\mu(z) \text{ see Billingsley Ex. 16.9}$$

3. Applying Billingsley 16.12 (pg. 214), we have

$$\begin{aligned} \mu(A) \int e^{z\theta} d\mu_A(z) &= \mu(A) \int e^{z\theta} \frac{1}{\mu(A)} \mathbb{I}_A(z)\, d\mu(z) \\ &= \int e^{z\theta} \mathbb{I}_A(z)\, d\mu(z) \\ &\leq \int e^{z\theta} d\mu(z) \end{aligned}$$

   where the last inequality is from the monotonicity of the Lebesgue integral, since $0 \leq e^{z\theta} \mathbb{I}_A(z) \leq e^{z\theta}$.

4. Taking the inverse of both sides gives the desired result

$$\frac{1}{\int e^{z\theta} d\mu(z)} \leq \frac{1}{\mu(A) \int e^{z\theta} d\mu_A(z)}$$

Second Inequality

1. $e^{(\cdot)}$ is a convex function, so from Jensen's Inequality, we have

$$e^{\int z\theta d\mu_A(z)} \leq \int e^{z\theta} d\mu_A(z)$$

2. Taking the inverse of both sides gives the desired result

$$\begin{aligned} e^{-\int z\theta d\mu_A(z)} &\geq \frac{1}{\int e^{z\theta} d\mu_A\left(z\right)} \\ \left(\int e^{z\theta} d\mu_A\left(z\right)\right)^{-1} &\leq e^{-\theta x_A} \end{aligned}$$

BIO 249 Lab 3: Introduction to Bayesian Computation

Sep 19, 2016

Recommended reading: Notes Lecture 5, Gelman Chapter 10

Monte Carlo

Monte Carlo represents the second "MC" of MCMC. The idea of Monte Carlo simulation is to draw iid samples $\left\{x^{(i)}\right\}_{i=1}^{N}$ from a target distribution $p\left(x\right).$ Of particular use in Bayesian statistics, these N samples can be used to approximate integrals $I\left(f\right)$ with tractable sums $I_N\left(f\right)$ that converge as follows:

$$I_N\left(f\right)=\frac{1}{N}\sum_{i=1}^{N}f\left(x^{(i)}\right)\overset{a.s.}{\to}I\left(f\right)=\int f\left(x\right)p\left(x\right)dx$$

That is, the estimate $I_N\left(f\right)$ is unbiased for $I\left(f\right)$ and converges almost surely to $I\left(f\right)$ by the strong law of large numbers. Some examples include

$$\begin{aligned}E\left(X\right) &= \int xp\left(x\right)dx\doteq\frac{1}{N}\sum_{i=1}^{N}x^{(i)}\\ P\left(X>0\right) &= \int_0^\infty p\left(x\right)dx\doteq\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[x^{(i)}>0\right]\end{aligned}$$

Sampling randomly allows us to apply statistics to interpret results when the analytical work is too complex. Over the years, Monte Carlo became a technical term for simulation of random processes. Any time you run simulations to compute something empirically, you are using Monte Carlo methods. The remainder of this lab discusses methods for drawing the iid samples $\left\{x^{(i)}\right\}_{i=1}^{N}$ if their distribution is not standard.

Rejection Sampling

We can sample from a distribution $p\left(x\right),$ which is known up to a proportionality constant, by sampling from another easy-to-sample proposal distribution $q\left(x\right)$ satisfying $p\left(x\right)\leq Mq\left(x\right), M<\infty$. We might want to do this in situations where we don't know the exact distribution of $p\left(x\right),$ or are unable to easily generate samples from $p\left(x\right).$ The general framework of rejection sampling involves the following steps:

For R iterations indexed by i,

1. Sample $x^{(i)}\sim q\left(x\right)$
2. Generate $u\sim\text{Unif}\left(0,1\right)$
3. If $u<\frac{p\left(x^{(i)}\right)}{Mq\left(x^{(i)}\right)}$ then accept $x^{(i)}$

How Rejection Sampling Works

Conceptually, we are accepting the generated $x^{(i)}$ with probability $\frac{1}{M}$. To see this, first note that, for a random variable A with support in $(0,1)$,

$$\begin{aligned}
P\left(U\leq A\right) &= \int_0^1\int_0^a f_U\left(u\right)f_A\left(a\right)duda \\
&= \int_0^1\left[\int_0^a f_U\left(u\right)du\right]f_A\left(a\right)da \\
&= \int_0^1 P\left(U\leq a\right)f_A\left(a\right)da \\
&= \int_0^1 af_A da \\
&= E\left(A\right)
\end{aligned}$$

Then, since $0\leq p\left(x\right)\leq Mq\left(x\right)\Rightarrow 0\leq\frac{p(x)}{Mq(x)}\leq 1$,

$$\begin{aligned}
P\left(U<\frac{p\left(X\right)}{Mq\left(X\right)}\right) &= E\left(\frac{p\left(X\right)}{Mq\left(X\right)}\right) \\
&= \int\frac{p\left(t\right)}{Mq\left(t\right)}q\left(t\right)dt \\
&= \frac{1}{M}\int p\left(t\right)dt \\
&= \frac{1}{M}
\end{aligned}$$

Let Y be the draws that are accepted. We would like to show that the rejection sampling procedure produces Y that represent draws from $X\sim p\left(x\right)$. That is, we would like to show that the conditional distribution of Y given $U\leq\frac{p(Y)}{Mq(Y)}$ is indeed $p\left(x\right)$, or that

$$P\left(Y\leq y\middle|U\leq\frac{p\left(Y\right)}{Mq\left(Y\right)}\right)=P\left(X\leq y\right)$$

First note that

$$\begin{aligned}
P\left(U\leq\frac{p\left(Y\right)}{Mq\left(Y\right)}\middle|Y\leq y\right) &= \frac{P\left[U\leq\frac{p(Y)}{Mq(Y)},Y\leq y\right]}{P\left(Y\leq y\right)} \\
&= \int_{-\infty}^{y}\frac{P\left[U\leq\frac{p(y)}{Mq(y)},Y=t\leq y\right]}{P\left(Y\leq y\right)}q\left(t\right)dt \\
&= \frac{1}{P\left(Y\leq y\right)}\int_{-\infty}^{y}\frac{p\left(t\right)}{Mq\left(t\right)}q\left(t\right)dt \\
&= \frac{P\left(X\leq y\right)}{MP\left(Y\leq y\right)}
\end{aligned}$$

Then

$$
\begin{aligned}
P\left(Y \le y \middle| U \le \frac{p(Y)}{Mq(Y)}\right) &= \frac{P\left(U \le \frac{p(Y)}{Mq(Y)} \middle| Y \le y\right) P(Y \le y)}{P\left[U \le \frac{p(Y)}{Mq(Y)}\right]} \\
&= \frac{\frac{P(X \le y)}{MP(Y \le y)} P(Y \le y)}{\frac{1}{M}} \\
&= P(X \le y)
\end{aligned}
$$

Example 1: Generating the Beta $(2,2)$ Distribution with Rejection Sampling

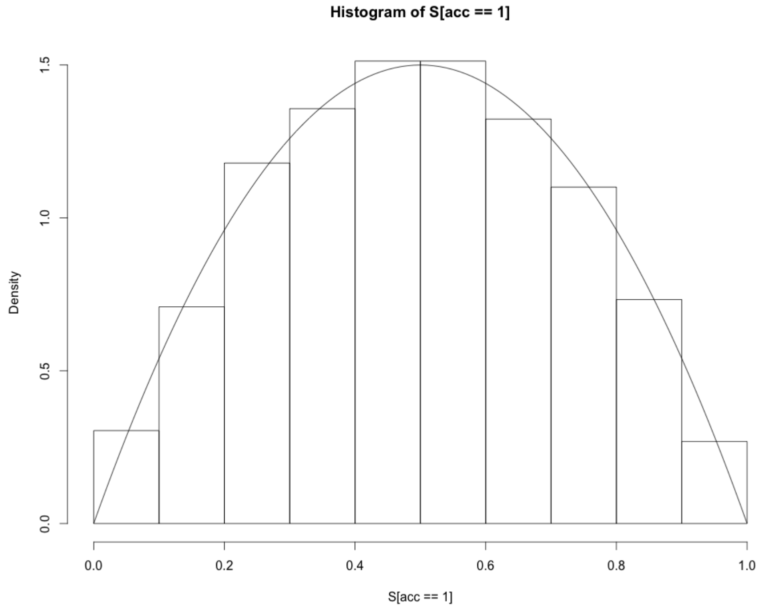
Suppose we live in a world where the Beta distribution is unknown, or standard statistics software is unable to generate iid draws of the Beta distribution. All we know is that our distribution of interest is

$$p(x) \propto x(1-x), 0 < x < 1$$

Note that the exact distribution of $X \sim \text{Beta}(2,2)$ has normalizing constant 6.

We know that $p(x) \le 0.25$ for $0 < x < 1$, so we can write $p(x) \le 0.25q(x)$ where $q(x) = 1$ is the density of the uniform distribution (which we do know). Note that in spirit of the previous notation, $p(x)$ should actually include its normalizing constant of 6, which makes $M = 1.5$, giving us an acceptance probability of about 67%, which is verified in the simulations. In practice, we often fine tune various things to get desirable acceptance probabilities, rather than calculate them ahead of time.

```
# create a grid of values for x
x<-seq(0,1,0.01)

# number of iterations
R<-10000

# number of acceptances
acc<-rep(0,R)

# main loop
S<-rep(NA,R)
for (i in 1:R) {
  t<-runif(1,0,1) # proposal
  u<-runif(1,0,1)
  S[i]<-t
  if (u<t*(1-t)*4) {
    acc[i]<-1
  }
}

mean(acc)
# [1] 0.6704

hist(S[acc==1],probability=T)
lines(x,6*x*(1-x))
```



Markov Chains

Markov chains represent the first "MC" of MCMC. A Markov chain is a stochastic process that operates sequentially, transitioning from one state to the next within an allowed set of states. Markov chain is defined by three elements:

1. A state space $\mathcal{X}$, which is a set of values that the chain is allowed to take (can be discrete or continuous).
2. A transition operator $P\left(X_{n+1}|X_n\right)$ that defines the probability of moving from X_n to X_{n+1}.
   - The Markov chain is memoryless if $P\left(X_{n+1}|X_n\right) = P\left(X_{n+1}|X_n, \ldots, X_1\right)$
   - This is known as the Markov property
   - For discrete Markov chains, these probabilities can be represented with a transition matrix $\mathbf{P}$ where $\mathbf{P}_{ij} = P\left(X_{n+1} = j|X_n = i\right)$
3. An initial condition distribution $\pi^{(0)}$, which defines the probability of being in any one of the possible states at the initial iteration $n = 0$.

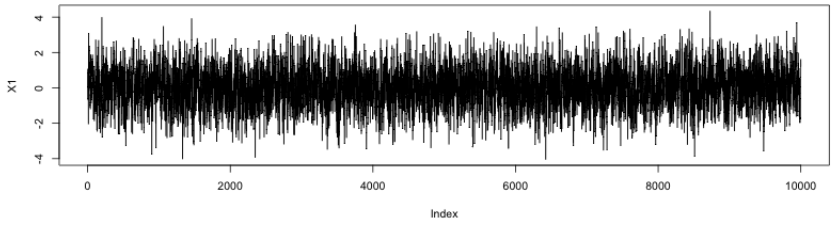
The Markov chain starts at some initial state, sampled from $\pi^{(0)}$, then transitions from one state to another according to $P\left(X_{n+1}|X_n\right)$. If $P\left(X_{n+1}|X_n\right)$ does not change across transitions (does not depend on n), the Markov chain is called time homogeneous. In that case, as $n \to \infty$, the chain will reach an equilibrium that is called the chain's stationary distribution, denoted π. For MCMC, we would like a Markov chain that moves towards our distribution of interest (posterior distribution). Classical examples of discrete Markov chains include weather prediction, where π represents the probability of observing a type of weather on a particular day.

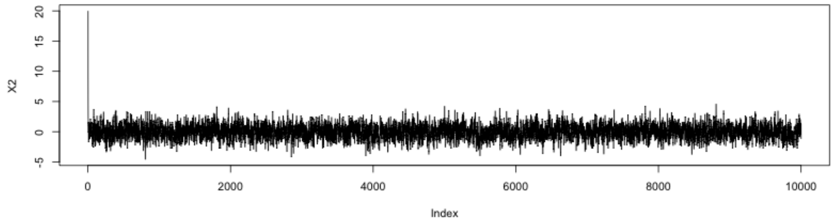
Example 2: Continuous Markov Chain

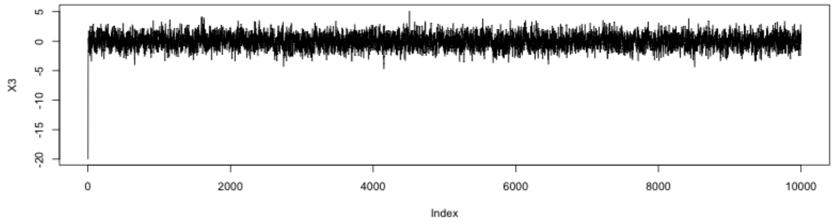
For continuous Markov chains, the transition operator cannot be represented as a matrix, but is instead a continuous function. Consider the transition operator defined by

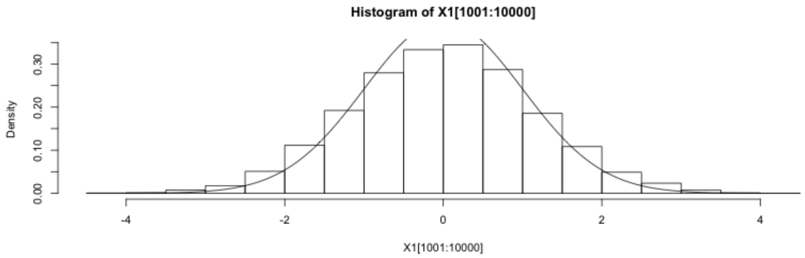
$$\begin{aligned} X_{n+1} &\sim \mathcal{N}\left(\frac{X_n}{2}, 1\right), n \geq 1 \\ X_1 &\sim \mathcal{N}\left(0, 1\right) \end{aligned}$$

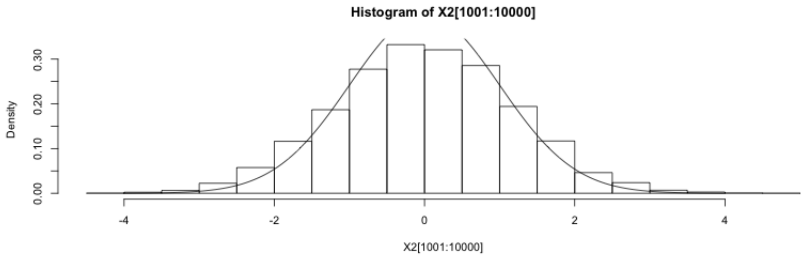
Then $X_n \overset{d}{\to} \mathcal{N}\left(0,1\right)$ as $n \to \infty$, so our stationary distribution is a standard normal. That is, we could simulate draws from the standard normal distribution using the continuous Markov chain described above (trivial example). One problem with this is the dependency of the draws, which we will address in the future. In the plots below, I started the chain X1 with initial value 0. Chain X2 had initial value 20, and chain X3 had initial value -20.

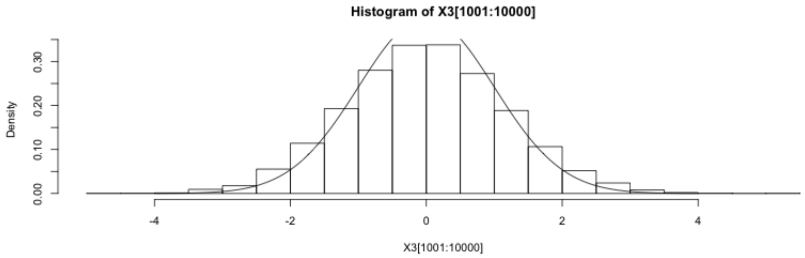












Detailed Balance (Reversibility)

Let π represent the stationary distribution. A Markov chain is reversible if it satisfies the detailed balance equations

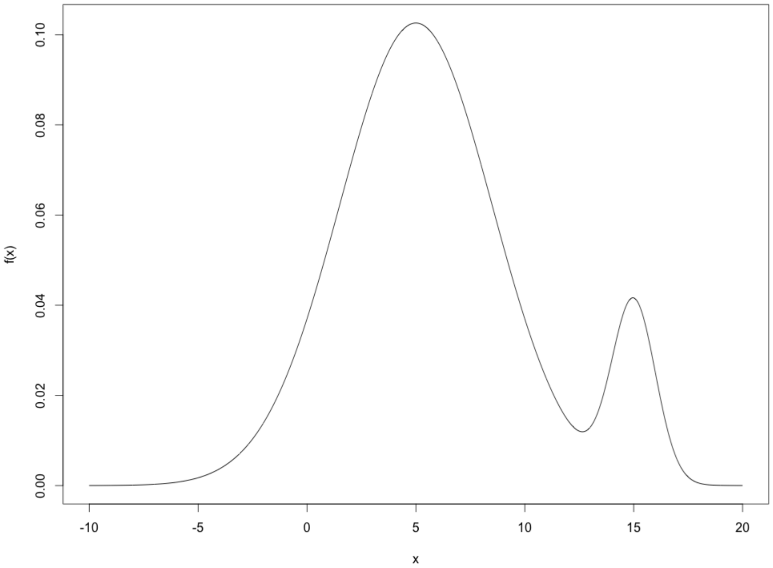
$$\pi_i P\left(X_{n+1} = j | X_n = i\right) = \pi_j P\left(X_{n+1} = i | X_n = j\right)$$

Detailed balance is a strong condition for Markov chains than possessing a stationary distribution.

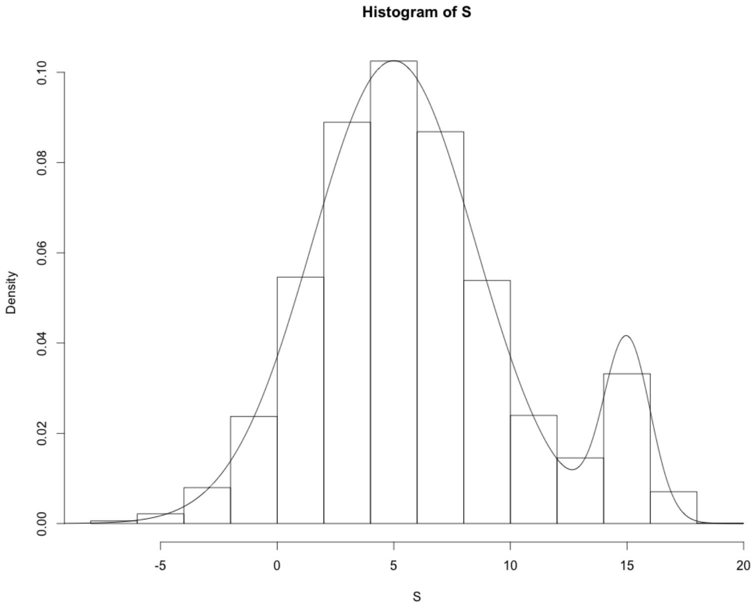
Example 3: Rejection Sampling with Random Walk

Consider the mixture distribution $f(x) \sim 0.9\mathcal{N}(5, 3.5) + 0.1\mathcal{N}(15, 1)$. We can use a random walk algorithm to estimate the density of $f(x)$.

```
# create a grid of values for x
x<-seq(-10,20,0.01)

# calculate f
f<-function(x) {
  0.9*dnorm(x,5,3.5)+0.1*dnorm(x,15,1)
}

# plot f
plot (x,f(x),type='l')
```



```
# set R (# iterations) and S (samples to be collected)
# initial value for S is 0
R<-50000
S<-rep(NA,R)
S[1]<-0

# number of accepted samples
acc<-0

# main loop
for (i in 2:R) {
  S[i]<-rnorm(1,mean=S[i-1],sd=3)
  r<-f(S[i])/f(S[i-1])
  if (runif(1)>r) {
    S[i]<-S[i-1] # reject S[i]
  } else {
    acc<-acc+1
  }
}

# acceptance rate
acc/R

hist(S,probability=T) lines(x,f(x))
```


Histogram of S
Density
0.00
0.02
0.04
0.06
0.08
0.10
-5
0
5
10
15
20
S

BIO 249 Lab 4: Metropolis-Hastings

Sep 26, 2016

Recommended reading: Notes Lecture 6-8, Understanding the Metropolis-Hastings Algorithm (1995)

Overview

Let $\boldsymbol{y}$ denote the observations of data, and let $\boldsymbol{\theta}$ denote the parameter or set of parameters for the data. Bayesian methods combine prior information on $\boldsymbol{\theta}$ summarized in $\pi(\boldsymbol{\theta})$ with the likelihood $\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta})$ to produce the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. From the posterior distribution we can extract any information - not simply "the most likely value" of a parameter, as with maximum likelihood estimators. However, until the advent of Monte Carlo Markov Chain methods, it was not straightforward to sample from the posterior distribution, except in cases where it was analytically defined.

Assume we have a preset initial parameter value $\boldsymbol{\theta}^{(1)}$. Then the MCMC methods involve generating a correlated sequence of sampled values $\left\{\boldsymbol{\theta}^{(t)}\right\}_{t=2,\ldots,R}$ based on the transition distribution

$$q\left(\boldsymbol{\theta}^{(t)}\middle|\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(t-1)}\right)=q\left(\boldsymbol{\theta}^{(t)}\middle|\boldsymbol{\theta}^{(t-1)}\right)$$

that is Markovian in the sense of depending only on $\boldsymbol{\theta}^{(t-1)}$. The general process for Metropolis-Hastings is (at iteration t)

1. Propose $\boldsymbol{\theta}^{*}\sim q$
2. Calculate the acceptance probability a
3. If $a>u\sim\mathrm{U}(0,1)^{\dagger}$ then set $\boldsymbol{\theta}^{*}$ to $\boldsymbol{\theta}^{(t)}$
    - Otherwise set $\boldsymbol{\theta}^{(t)}=\boldsymbol{\theta}^{(t-1)}$
4. Repeat

†But why randomly keep "bad" proposal samples? It turns out that doing this allows the Markov chain to occasionally visit states of low probability under the target distribution. This is a desirable property if we want the chain to adequately sample the entire target distribution, including any tails.

The transition distribution q is chosen to satisfy additional conditions ensuring that the sequence has the target (joint posterior distribution) p as its stationary distribution. These conditions typically reduce to requirements on the proposal distribution (q) and acceptance rule (a) used to generate new parameter samples. Under such conditions, the sampled parameters $\left\{\boldsymbol{\theta}^{(t)}\right\}_{t=b+1,\ldots,R}$ beyond a certain burn-in phase of length b can be viewed as a random sample from the distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. In particular, we would like our Markov chain with transition distribution q to have a unique stationary distribution p. In Metropolis-Hastings, instead of using q to find the stationary distribution p, we know what we want p to be ahead of time and instead use p to define our q. In general, this would require

$$p\left(y\right)=\sum_{x}p\left(x\right)q\left(y|x\right) \tag{1}$$

but instead we use a sufficient (and stronger) condition of detailed balance

$$p\left(y\right)q\left(x|y\right)=p\left(x\right)q\left(y|x\right) \tag{2}$$

It is easy to show that (2) ⇒ (1) because

$$\begin{aligned}\sum_{x}p\left(x\right)q\left(y|x\right) &= \sum_{x}p\left(y\right)q\left(x|y\right)\ \text{ from }\ (2)\\ &= p\left(y\right)\sum_{x}q\left(x|y\right)\\ &= p\left(y\right)\end{aligned}$$

We now would like to find q and a satisfying (2). That is, we would like

$$\begin{aligned}p\left(y\right)q\left(x|y\right)a\left(x|y\right) &= p\left(x\right)q\left(y|x\right)a\left(y|x\right)\\ \frac{a\left(y|x\right)}{a\left(x|y\right)} &= \frac{p\left(y\right)q\left(x|y\right)}{p\left(x\right)q\left(y|x\right)}\end{aligned}$$

from which we get (details provided in 2014 Notes)

$$a\left(y|x\right)=\min\left\{1,\frac{p\left(y\right)q\left(x|y\right)}{p\left(x\right)q\left(y|x\right)}\right\}$$

In this notation, y can be thought of as the candidate at iteration t, and x can be thought of as the last draw at iteration $t-1$.

In practice, MCMC methods may be applied separately to individual parameters or groups ("blocks") of more than one parameter. Parameters in a particular block may be updated jointly. Different MCMC methods may be applied to different parameters or blocks, such as a Gibbs sampler that requires Metropolis-Hastings to generate one of its conditional posteriors.

Using the Metropolis (no Hastings) algorithm, detailed balance is assured if the proposal distribution is symmetric. However, using a symmetric proposal distribution may not be reasonable to adequately or efficiently sample all possible target distributions. For example, if a target distribution is bounded on the positive numbers, we would like to use a proposal distribution that has the same support, and will thus be not symmetric. This is where the Metropolis-Hastings sampling algorithm comes in.

Metropolis-Hastings

The Metropolis-Hastings algorithm involves a proposal distribution $q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right)$, where $\boldsymbol{\theta}^*$ is a candidate for a draw of the posterior distribution of $\boldsymbol{\theta}$. The acceptance probability is

$$a = \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right) q\left(\boldsymbol{\theta}^{(t-1)}\middle|\,\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y}\right) q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right)}\right\}$$

Often we will see the dependence on $\boldsymbol{y}$ omitted in this notation. When q is symmetric, we get

$$q\left(\boldsymbol{\theta}^{(t-1)}\middle|\,\boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right)$$

so that

$$a = \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*|\,\boldsymbol{y}\right)}{p\left(\boldsymbol{\theta}^{(t)}\middle|\,\boldsymbol{y}\right)}\right\}$$

In this case, the algorithm is called the Metropolis algorithm (without the contribution of Hastings). For example, if we have a normal proposal distribution

$$\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)} \sim \mathcal{N}\left(\boldsymbol{\theta}^{(t-1)}, \hat{\sigma}^2\right)$$

then

$$q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{1}{2\hat{\sigma}^2}\left(\boldsymbol{\theta}^*-\boldsymbol{\theta}^{(t-1)}\right)^2} = q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right)$$

Special Case 1: IID Sampling (Monte Carlo Simulation)

If $q\left(\boldsymbol{\theta}^*|\,\boldsymbol{\theta}^{(t-1)}\right) = p\left(\boldsymbol{\theta}^*|\,\boldsymbol{y}\right)$ then

$$a = \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*|\boldsymbol{y}\right) p\left(\boldsymbol{\theta}^{(t-1)}\middle|\,\boldsymbol{y}\right)}{p\left(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y}\right) p\left(\boldsymbol{\theta}^*|\,\boldsymbol{y}\right)}\right\} = 1$$

In this case, our proposal distribution is just the posterior distribution itself, so each iteration we propose directly from the posterior distribution and accept the proposal with probability one.

Special Case 2: Gibbs Algorithm

Suppose we have an n-dimensional $\boldsymbol{\theta}$ with full conditionals

$$p\left(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_n\right)$$

Gibbs sampling can be viewed as a special case of Metropolis-Hastings where the proposal q is based on a two stage procedure.

1. A single dimension j of $\boldsymbol{\theta}$ is proposed

2. The proposed value θ_j^* is identical to $\boldsymbol{\theta}^{(t-1)}$, except that its value at dimension j is sampled from the distribution $p\left(\theta_j^*|\boldsymbol{\theta}_{-j}^{(t-1)}\right)$

This procedure can be summarized as

$$q\left(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}\right) = \begin{cases} p\left(\theta_j^*|\boldsymbol{\theta}_{-j}^{(t-1)}\right) & \text{if } \boldsymbol{\theta}_{-j}^* = \boldsymbol{\theta}_{-j}^{(t-1)} \\ 0 & \text{otherwise} \end{cases}$$

The corresponding acceptance probability is

$$\begin{aligned}
a &= \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*\right) q\left(\boldsymbol{\theta}^{(t-1)}\middle|\,\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}\right) q\left(\boldsymbol{\theta}^*\middle|\,\boldsymbol{\theta}^{(t-1)}\right)}\right\} \\
& \qquad \text{from here on, we have } \boldsymbol{\theta}_{-j}^* = \boldsymbol{\theta}_{-j}^{(t-1)} \\
&= \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*\right) p\left(\theta_j^{(t-1)}\middle|\,\boldsymbol{\theta}_{-j}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}\right) p\left(\theta_j^*\middle|\,\boldsymbol{\theta}_{-j}^{(t-1)}\right)}\right\} \\
&= \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*\right) p\left(\theta_j^{(t-1)}\middle|\,\boldsymbol{\theta}_{-j}^{(t-1)}\right)}{p\left(\boldsymbol{\theta}^{(t-1)}\right) p\left(\theta_j^*\middle|\,\boldsymbol{\theta}_{-j}^*\right)}\right\} \\
&= \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*\right) p\left(\theta_j^{(t-1)}, \boldsymbol{\theta}_{-j}^{(t-1)}\right) p\left(\boldsymbol{\theta}_{-j}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}\right) p\left(\theta_j^*, \boldsymbol{\theta}_{-j}^*\right) p\left(\boldsymbol{\theta}_{-j}^{(t-1)}\right)}\right\} \\
&= \min\left\{1, \frac{p\left(\boldsymbol{\theta}^*\right) p\left(\boldsymbol{\theta}^{(t-1)}\right) p\left(\boldsymbol{\theta}_{-j}^*\right)}{p\left(\boldsymbol{\theta}^{(t-1)}\right) p\left(\boldsymbol{\theta}^*\right) p\left(\boldsymbol{\theta}_{-j}^{(t-1)}\right)}\right\} \\
&= \min\left\{1, \frac{p\left(\boldsymbol{\theta}_{-j}^*\right)}{p\left(\boldsymbol{\theta}_{-j}^{(t-1)}\right)}\right\} = 1
\end{aligned}$$

Example 1

Download the bank data set `http://www.ceremade.dauphine.fr/~xian/BCS/bank`

Assume a Probit model without intercept and four covariates $x_1, \ldots, x_4$ as main effects.

$$\begin{aligned}
y_i &\sim \text{Bernoulli}\left(p_i\right) \\
p_i &= \Phi\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}\right)
\end{aligned}$$

Assume a diffuse normal prior on the four regression coefficients.

Problem A: Obtain samples from the marginal posterior distribution $p\left(\beta_1, \beta_2, \beta_3, \beta_4|\boldsymbol{y}\right)$ by implementing a Metropolis-Hastings sampler.

Problem B: Obtain samples from the marginal posterior distribution $p\left(\beta_1, \beta_2, \beta_3, \beta_4|\boldsymbol{y}\right)$ by implementing a Gibbs sampler with latent continuous data.

Solutions

For reference, a standard GLM analysis yields

```
   Estimate Std. Error z value Pr(>|z|)
x1  -1.1810     0.2747  -4.299 1.72e-05 ***
x2   0.9515     0.5715   1.665   0.0959 .
x3   0.9217     0.5144   1.792   0.0731 .
x4   1.1028     0.1683   6.554 5.60e-11 ***
```

Problem A

For notational convenience, we suppress the dependence on $\mathbf{X}$ throughout the problem. Let the diffuse normal prior be

$$\pi\left(\boldsymbol{\beta}\right) = \left(2\pi d^2\right)^{-\frac{k}{2}} \exp\left(-\frac{1}{2d^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right), \boldsymbol{\beta} \in \mathbb{R}^k$$

for some large d^2. We formulate the posterior as

$$\begin{aligned} p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &\propto L\left(\boldsymbol{y}|\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right) \\ &\propto \prod_{i=1}^{n} \Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^{y_i}\left[1-\Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]^{1-y_i} \exp\left(-\frac{1}{2d^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \end{aligned}$$

To simplify coding, we note that

$$\log\left[p\left(\boldsymbol{\beta}|\boldsymbol{y}\right)\right] = \sum_{i=1}^{n}\left(y_i \log\Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) + \left(1-y_i\right)\log\left[1-\Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right]\right) - \frac{1}{2d^2}\boldsymbol{\beta}^T\boldsymbol{\beta}$$

Let the proposal distribution be $\mathcal{N}\left[\boldsymbol{\beta}^{(t-1)}, \tau\hat{\boldsymbol{\Sigma}}\right]$ where $\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ is an estimate of the asymptotic covariance matrix. Then the steps are (for iteration $t \geq 2$)

1. Propose $\boldsymbol{\beta}^* \sim \mathcal{N}\left[\boldsymbol{\beta}^{(t-1)}, \tau\hat{\boldsymbol{\Sigma}}\right]$
2. Calculate the acceptance probability $a = \min\left\{1, \frac{p\left(\boldsymbol{\beta}^*|\boldsymbol{y}\right)}{p\left(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{y}\right)}\right\}$
3. If $a > u \sim \mathrm{U}\left(0,1\right)$ then set $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}^{(t)}$
    - Otherwise set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$
4. Repeat

Posterior Summary for $\boldsymbol{\beta}$ using Metropolis Hastings

	25%ile	50%ile	75%ile	Mean	SD
β_1	-1.39	-1.21	-1.03	-1.21	0.26
β_2	0.58	0.97	1.38	0.98	0.60
β_3	0.59	0.94	1.30	0.95	0.53
β_4	1.02	1.13	1.25	1.14	0.17

Problem B

For notational convenience, we suppress the dependence on X throughout the problem. Now we re-express the data in terms of the continuous latent variable $Z_i|\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)$ so that $Y_i = I\left(Z_i > 0\right)$. To see this, note that

$$\begin{aligned} P\left(Y_i = 1\right) &= P\left(Z_i > 0\right) \\ &= P\left(Z_i - \boldsymbol{x}_i^T\boldsymbol{\beta} > -\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \\ &= \Phi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \end{aligned}$$

We formulate the joint posterior as

$$\begin{aligned} p\left(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}\right) &\propto L\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{z}\right)\pi\left(\boldsymbol{\beta}, \boldsymbol{z}\right) \\ &= L\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{z}\right)\pi\left(\boldsymbol{z}|\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right) \\ &\propto \prod_{i=1}^{n}\left[I\left(z_i > 0\right)I\left(y_i = 1\right) + I\left(z_i \le 0\right)I\left(y_i = 0\right)\right]\prod_{i=1}^{n}\phi\left(z_i; \boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)\exp\left(-\frac{1}{2d^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \end{aligned}$$

with conditional posteriors

$$\begin{aligned} p\left(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}\right) &\propto \prod_{i=1}^{n}\phi\left(z_i; \boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)\exp\left(-\frac{1}{2d^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\ &\sim \mathcal{N}\left[\left(d^{-2}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\boldsymbol{z}\right), \left(d^{-2}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\right] \\ p\left(z_i|\boldsymbol{\beta}, \boldsymbol{y}\right) &\sim \begin{cases} \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)I\left(z_i > 0\right) & \text{if } y_i = 1 \\ \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)I\left(z_i \le 1\right) & \text{if } y_i = 0 \end{cases} \end{aligned}$$

Then the steps are (for iteration $t \ge 2$)

1. For $i = 1, \ldots, n$, simulate $z_i^{(t)} \sim \mathcal{N}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1\right)$
   - z_i truncated > 0 if $y_i = 1$
   - z_i truncated < 0 if $y_i = 0$
2. Generate $\boldsymbol{\beta}^{(t)} \sim \mathcal{N}\left[\left(d^{-2}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\boldsymbol{z}^{(t)}\right), \left(d^{-2}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\right]$
3. Repeat

Posterior Summary for β using Gibbs

	25%ile	50%ile	75%ile	Mean	SD
β_1	-1.40	-1.21	-1.03	-1.22	0.27
β_2	0.58	0.97	1.39	0.98	0.60
β_3	0.59	0.95	1.30	0.95	0.53
β_4	1.02	1.13	1.25	1.14	0.17

Example 2

The data below consist of the estimated gestational ages (weeks) and weights (g) of 12 female babies:

Age	40	36	40	38	42	39	40	37	36	38	39	40
Weight	3317	2729	2935	2754	3210	2817	3126	2539	2412	2991	2875	3231

Assume a linear regression model predicting gestational age from weight with non-informative prior distributions on all parameters. Give an interval in which you are 90% sure that the gestational age of a particular 3000g baby will lie. Give a similar interval for the average age of all 3000g babies.

Solution 2

For notational convenience, we suppress the dependence on X throughout the problem. Define our outcome Y as gestational age and covariate X as weight. Then we are assuming the linear regression model

$$\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2,\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}\right), \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

with non-informative prior

$$\pi\left(\boldsymbol{\beta},\sigma^2\right) \propto \frac{1}{\sigma^2}$$

Then the joint posterior distribution is

$$\begin{aligned} p\left(\boldsymbol{\beta},\sigma^2,|\boldsymbol{y}\right) &\propto L\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2\right)\pi\left(\boldsymbol{\beta},\sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\right)\right]\sigma^{-2} \\ &\propto \left(\sigma^2\right)^{-\left(\frac{n}{2}+1\right)} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\right)\right] \end{aligned}$$

From this we get

$$\begin{aligned} p\left(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\right) &\sim \mathcal{N}\left[\boldsymbol{\mu}=\hat{\boldsymbol{\beta}} := \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}, \boldsymbol{\Sigma}=\sigma^2\mathbf{V}_{\boldsymbol{\beta}} := \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right] \\ p\left(\sigma^2|\boldsymbol{y}\right) &\sim \text{Inv-}\chi^2\left[\nu=n-k, \tau^2=s^2 := \frac{1}{n-k}\left(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)\right] \end{aligned}$$

We will need the posterior predictive distribution

$$\begin{aligned} p\left(\tilde{\boldsymbol{y}}|\boldsymbol{y}\right) &= \int\int L\left(\tilde{y}|\boldsymbol{\beta},\sigma^2\right)p\left(\boldsymbol{\beta},\sigma^2|\boldsymbol{y}\right)d\boldsymbol{\beta}d\sigma^2 \\ &= \int p\left(\sigma^2|\boldsymbol{y}\right)\int L\left(\tilde{y}|\boldsymbol{\beta},\sigma^2\right)p\left(\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\right)d\boldsymbol{\beta}d\sigma^2 \end{aligned}$$

The inner integral is equal to

$$p\left(\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y}\right) \quad = \quad \int \exp\left[-\frac{1}{2\sigma^2}\left(\tilde{\boldsymbol{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}\right)^T\left(\tilde{\boldsymbol{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}\right)\right]\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T\mathbf{V}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right]d\boldsymbol{\beta}$$

This is a kernel of a multivariate normal with

$$\begin{aligned}
E\left[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y}\right] &= E\left[E\left[\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y}\right]|\sigma^2,\boldsymbol{y}\right] \\
&= E\left[\tilde{\mathbf{X}}\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\right] \\
&= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \\
Var\left[\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y}\right] &= E\left[Var\left[\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y}\right]|\sigma^2,\boldsymbol{y}\right]+Var\left[E\left[\tilde{\boldsymbol{y}}|\boldsymbol{\beta},\sigma^2,\boldsymbol{y}\right]|\sigma^2,\boldsymbol{y}\right] \\
&= E\left[\sigma^2\mathbf{I}|\sigma^2,\boldsymbol{y}\right]+Var\left[\tilde{\mathbf{X}}\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}\right] \\
&= \left(\mathbf{I}+\tilde{\mathbf{X}}\mathbf{V}_{\boldsymbol{\beta}}\tilde{\mathbf{X}}^T\right)\sigma^2
\end{aligned}$$

Finally,

$$\begin{aligned}
p\left(\tilde{\boldsymbol{y}}|\boldsymbol{y}\right) &= \int p\left(\sigma^2|\boldsymbol{y}\right)p\left(\tilde{\boldsymbol{y}}|\sigma^2,\boldsymbol{y}\right)d\sigma^2 \\
&\sim \mathcal{T}_{n-k}\left[\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}, s^2\left(\mathbf{I}+\tilde{\mathbf{X}}\mathbf{V}_{\boldsymbol{\beta}}\tilde{\mathbf{X}}^T\right)\right] \text{ from 2014 Notes}
\end{aligned}$$

For a 3000 gram baby, $\tilde{\mathbf{X}} = \left[\begin{array}{cc} 1 & 3000 \end{array}\right]$. For reference, the point estimate is $\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} = 39.2$. Note that the posterior predictive distribution in this case will be a univariate t distribution. I generated 10,000 samples from this distribution and computed the 0.05 and 0.95 quantiles to get an interval of $(37.3, 41.2)$ in which I am 90% sure that the gestational age of a particular 3000g baby will lie. This interval is centered around 39.2. The analogous frequentist method would be a prediction interval for a new $\hat{y}$ at $x = 3000$.

For the average age of all 3000g babies, we are looking for the 0.05 and 0.95 quantiles of

$$\left[\begin{array}{cc} 1 & 3000 \end{array}\right]\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is taken from $\boldsymbol{\beta}|\boldsymbol{y}$, which follows a $\mathcal{T}_{n-k}\left[\hat{\boldsymbol{\beta}}, s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]$. We can alternatively write a Gibbs sampler to sample $\sigma^2|\boldsymbol{y}$ followed by $\boldsymbol{\beta}|\sigma^2,\boldsymbol{y}$. I generated 10,000 samples from the distribution of $\boldsymbol{\beta}|\boldsymbol{y}$ and computed the 0.05 and 0.95 quantiles to get an interval of $(38.7, 39.8)$ in which I am 90% sure that the average age of all 3000g babies will lie. This interval is centered around 39.2. Note that it is much more narrow because we don't have to account for the extra variability in sampling an individual 3000g baby. The analogous frequentist method would be a confidence interval for the mean response $\hat{y}$ at $x = 3000$. These answers are nearly identical to the numbers obtained via frequentist inference.

BIO 249 Lab 5: Introduction to Bayesian Nonparametrics

Oct 17, 2016

Recommended reading: Notes Lectures 18 & 21, Gelman Chapter 23

Supplementary Material on Polya Trees

1. Neath, A. A. (2003). Polya Tree Distributions for Statistical Modeling of Censored Data. *Journal of Applied Mathematics and Decision Sciences* **7**, 175-186.

   `ftp://ftp.muni.cz/mount/muni.cz/EMIS/journals/HOA/JAMDS/3607.pdf`

2. Paddock, S. M., Ruggeri, F., Lavine, M., West, M. (2003). Randomized Polya Tree Models for Nonparametric Bayesian Inference. *Statistica Sinica* **13**, 443-460.

   `http://ftp.isds.duke.edu/WorkingPapers/00-11.pdf`

   `http://www3.stat.sinica.edu.tw/statistica/oldpdf/A13n211.pdf`

3. Müller, P. and Rodriguez, A. (2013). Polya Trees. *Nonparametric Bayesian Inference*, 43-51.

   `http://projecteuclid.org/download/pdfview_1/euclid.cbms/1362163749`

1 Dirichlet Distribution and Multinomial Distribution

The Dirichlet distribution $(\theta_1, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$ is a multivariate generalization of the Beta distribution, parameterized by $\alpha_1, \ldots, \alpha_K > 0, \alpha_0 = \sum_{i=1}^K \alpha_i, K \geq 2$, with pdf

$$f(\theta_1, \ldots, \theta_K) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}, \theta_1, \ldots, \theta_K \in (0,1), \sum_{i=1}^K \theta_i = 1$$

The multinomial distribution $(y_1, \ldots, y_K) \sim \text{Multinomial}(\theta_1, \ldots, \theta_K)$ is a multivariate generalization of the binomial distribution, parameterized by $n > 0, \theta_1, \ldots, \theta_K \in (0,1), \sum_{i=1}^K \theta_i = 1, K \geq 2$, with pmf

$$f(y_1, \ldots, y_K | \theta_1, \ldots, \theta_k) = \frac{n!}{\prod_{i=1}^K y_i!} \prod_{i=1}^K \theta_i^{y_i}$$

The Dirichlet distribution is a conjugate prior for the multinomial distribution, with posterior

$$\begin{aligned} f(\theta_1, \ldots, \theta_K | y_1, \ldots, y_K) &\propto f(y_1, \ldots, y_K | \theta_1, \ldots, \theta_k) f(\theta_1, \ldots, \theta_K) \\ &\propto \prod_{i=1}^K \theta_i^{y_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} \\ &\sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_K + y_K) \end{aligned}$$

2 Dirichlet Process

The Dirichlet process is an infinite-dimensional generalization of the Dirichlet distribution that can be used to set a prior on unknown distributions.

2.1 Motivation: Bayesian (Random) Histograms

Consider a simple setting in which we have samples $\boldsymbol{y} \overset{\text{iid}}{\sim} f$ and the goal is to obtain a Bayes estimate of the density f. Assume we have pre-specified knots $(\xi_0, \ldots, \xi_k)$ to define our histogram bins, with $\xi_0 < \cdots < \xi_K$, and $y_i \in [\xi_0, \xi_K]$. A probability model for the density that is analogous to the histogram is

$$f(y) = \sum_{h=1}^{K} \mathbb{I}(\xi_{h-1} < y \leq \xi_h) \frac{\pi_h}{\xi_h - \xi_{h-1}}, y \in \mathbb{R},$$

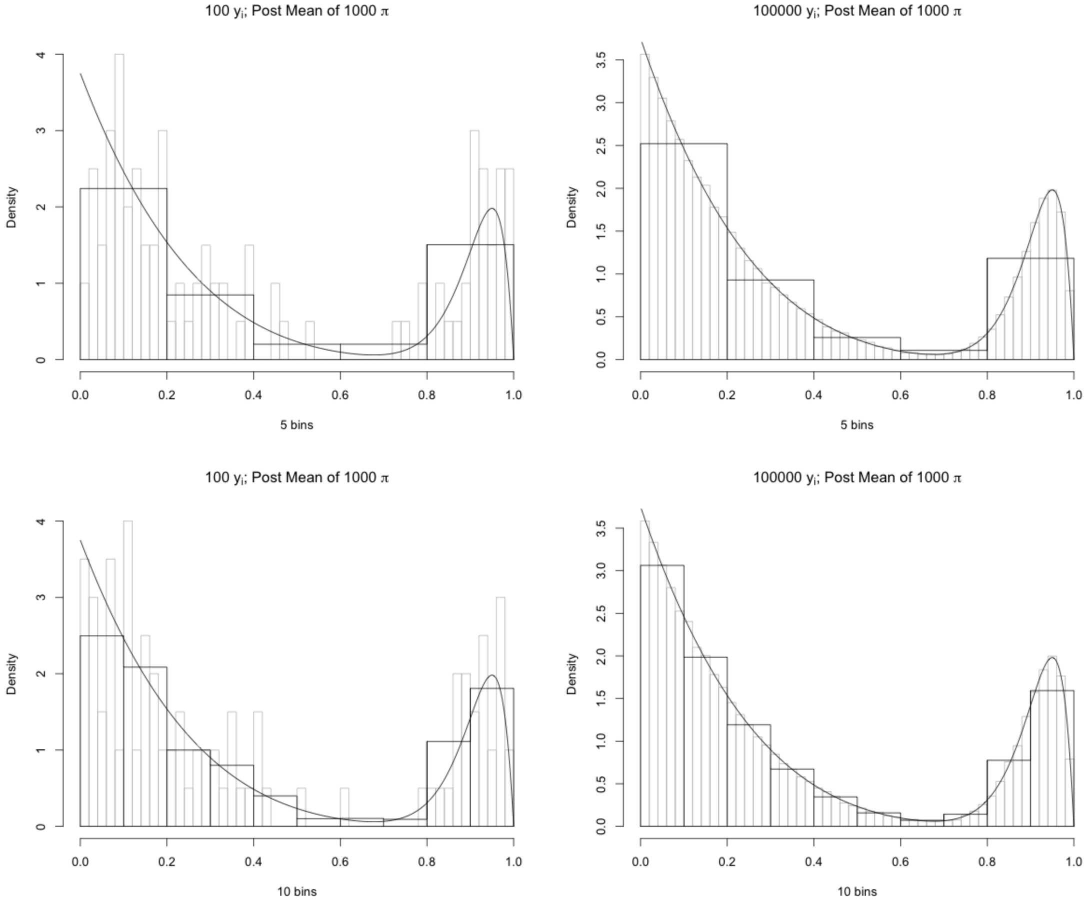
with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ an unknown probability vector summing to 1. The height of each bin is then $\frac{\pi_h}{\xi_h - \xi_{h-1}}$, so that the area under the histogram is 1. That is, the height of each bin is uncertain, and we would like to use our data, along with a prior distribution, to get a posterior distribution of the heights of the bins. If we assume a Dirichlet $(a_1, \ldots, a_K)$ prior for $\boldsymbol{\pi}$, we get the posterior

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\pi}) f(\boldsymbol{\pi}) \\ &\propto \prod_{i:y_i \in (\xi_h - \xi_{h-1}]} \frac{\pi_h}{\xi_h - \xi_{h-1}} \prod_{h=1}^{K} \pi_h^{\alpha_h - 1} \\ &= \prod_{h=1}^{K} \pi_h^{n_h} \prod_{h=1}^{K} \pi_h^{\alpha_h - 1} \\ &\sim \text{Dirichlet}(a_1 + n_1, \ldots, a_K + n_K) \end{aligned}$$

where $n_h = \sum_i \mathbb{I}(\xi_{h-1} < y \leq \xi_h)$ is the number of observations falling in the h^{th} histogram bin. The Dirichlet specification ensures that $\boldsymbol{\pi}$ has elements summing to 1, giving a proper distribution.

In the simulations below, the Bayesian histograms (posterior mean) are plotted in black, and the standard histogram of the samples are plotted in grey. Note that the results are sensitive to the number and location of the knots.

Source code: `Lab 5 2.1 Random Histogram.R`



2.2 Dirichlet Process Prior

Histograms have the unappealing characteristics of bin sensitivity and approximating a smooth density with piecewise constants. Additionally, extending histograms to multiple dimensions is problematic due to the number of bins needed. We look for a model that bypasses the need to explicitly specify bins ahead of time.

Let P be the unknown probability measure over $(\Omega, \mathscr{B})$, where Ω has finite measurable partition $B_1, \ldots, B_K$ (think of these as our histogram bins). We have interest in estimating P. Then

$$P\left(B_1\right), \ldots P\left(B_K\right) = \int_{B_1} f\left(y\right) dy, \ldots, \int_{B_K} f\left(y\right) dy$$

for our density f. The above probabilities are random variables for a random probability measure P (similar to how $\boldsymbol{\pi}$ was unknown in 2.1), meaning we can assign a prior

$$P\left(B_1\right), \ldots, P\left(B_K\right) \sim \text{Dirichlet}\left(\alpha P_0\left(B_1\right), \ldots, \alpha P_0\left(B_K\right)\right),$$

where P_0 is a base probability measure providing an initial guess at P, and α is a prior concentration parameter controlling the degree of shrinkage of P toward P_0. We eliminate sensitivity to the choice of the partition by assuming the prior on P holds for all partitions $B_1, \ldots, B_K$ and K.

The existence of such P is shown by Ferguson (1973). The resulting probability measure P is a Dirichlet process, denoted $P \sim \mathrm{DP}\left(P_0, \alpha\right).$

2.3 Stick Breaking Example

2.3.1 Dirichlet Distribution

We would like to generate a random vector with a Dirichlet $\left(\alpha_1, \ldots, \alpha_K\right)$ distribution. This involves iteratively breaking a stick of length 1 into K pieces in a way that the lengths of the K pieces follow a Dirichlet $\left(\alpha_1, \ldots, \alpha_K\right)$ distribution.

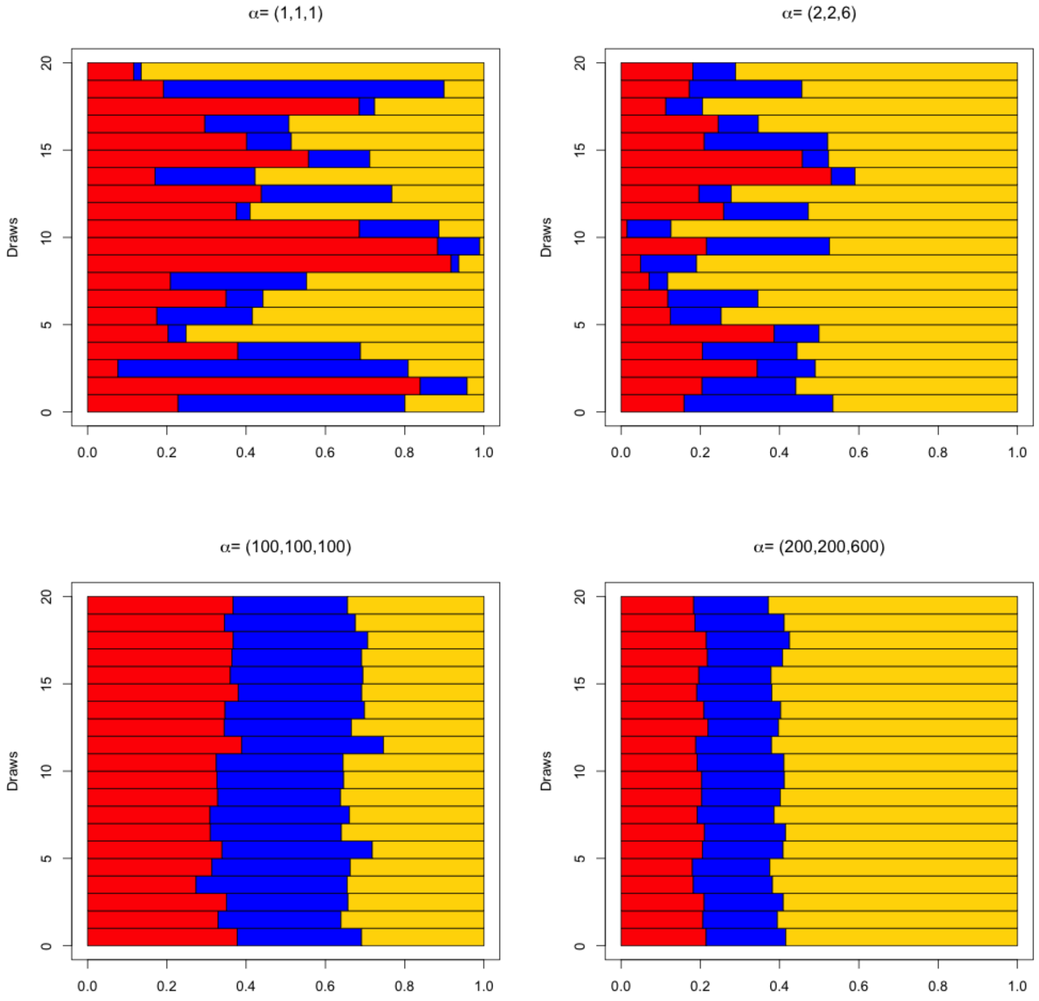
1. Simulate $v_1 \sim \text{Beta}\left(\alpha_1, \sum_{i=2}^K \alpha_i\right)$, and set $\pi_1 = v_1$. The stick has length $1 - \pi_1$ remaining.
2. For $2 \leq h \leq K-1$, simulate $v_h \sim \text{Beta}\left(\alpha_h, \sum_{i=h+1}^K \alpha_i\right)$, and set $\pi_h = v_h \prod_{i=1}^{h-1}\left(1 - v_i\right).$ The stick has length $1 - \sum_{i=1}^h \pi_i$.
3. The last piece has length $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$, the remainder of the stick.

Under this construction, the pieces $\pi_1, \ldots, \pi_K$ follow a Dirichlet $\left(\alpha_1, \ldots, \alpha_K\right)$ distribution.

For $\theta_1, \ldots, \theta_K \sim \text{Dirichlet}\left(\alpha_1, \ldots, \alpha_K\right),$

$$\begin{aligned}
E\left(\theta_i\right) &= \frac{\alpha_i}{\alpha_0} \\
Var\left(\theta_i\right) &= \frac{\alpha_i\left(\alpha_0 - \alpha_i\right)}{\alpha_0^2\left(\alpha_0 + 1\right)} \rightarrow 0 \text{ as } \alpha_0 \rightarrow \infty
\end{aligned}$$

Source code: first half of `Lab 5 2.3 Stick Break.R`



2.3.2 Dirichlet Process Prior

This process is similar to the one described above. The π_h we generate from the stick breaking process in this case represent probabilities of observing corresponding draws from P_0. Suppose there is a stick of length 1. Let $v_h \sim \text{Beta}\,(1, \alpha)$, and regard them as fractions of how much we take away from the remainder of the stick every time. That is,

$$\pi_1 = v_1, \pi_2 = (1 - v_1)\, v_2, \pi_h = v_h \prod_{l<h} (1 - v_l)$$

By this construction, we will have $\sum_{h=1}^{\infty} \pi_h = 1$. We note that $\lim\limits_{h\to\infty} \pi_h \to 0$, with convergence rate depending on α. For a low α, $\text{Beta}\,(1, \alpha)$ generates high values, leading to π_h approaching 0 very quickly. This means the first few π_h will have the most weight. For a high α, the weight is more evenly distributed, as $\text{Beta}\,(1, \alpha)$ generates low values. In simulations, I used generated 10000 (our "infinity") π_h and θ_h.

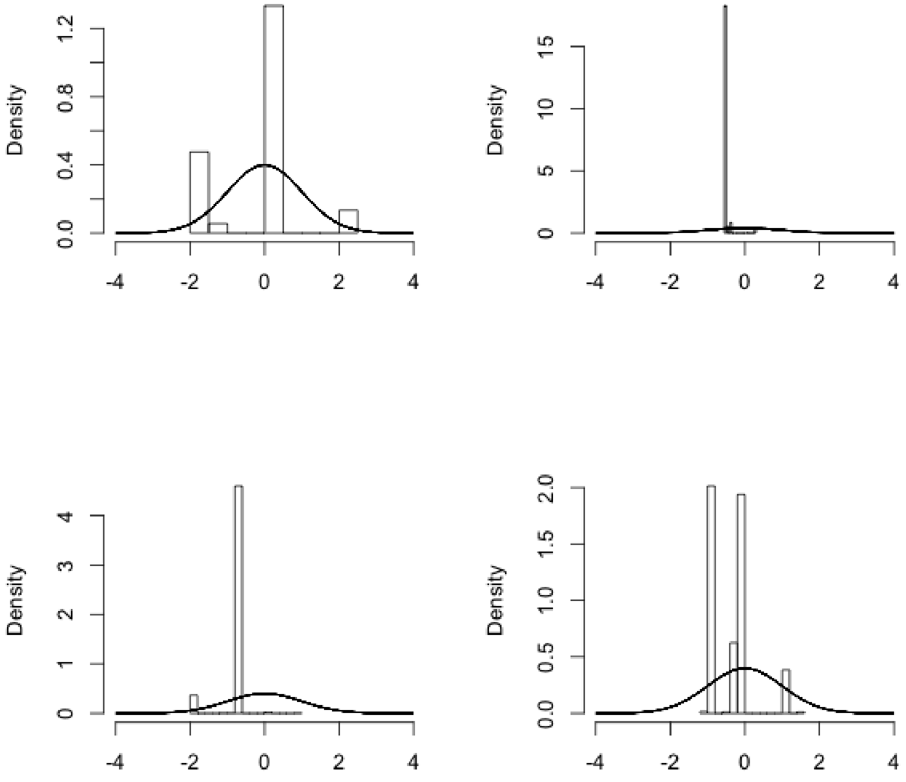
Using this procedure, we have an explicit construction of P :

1. For $h = 1, 2, \ldots, v_h \sim \text{Beta}(1, \alpha), \pi_h = v_h \prod_{l<h} (1 - v_l), \theta_h \sim P_0$.
2. $P(x) = \sum_{h=1}^{\infty} \pi_h \mathbb{I}(x = \theta_h)$.

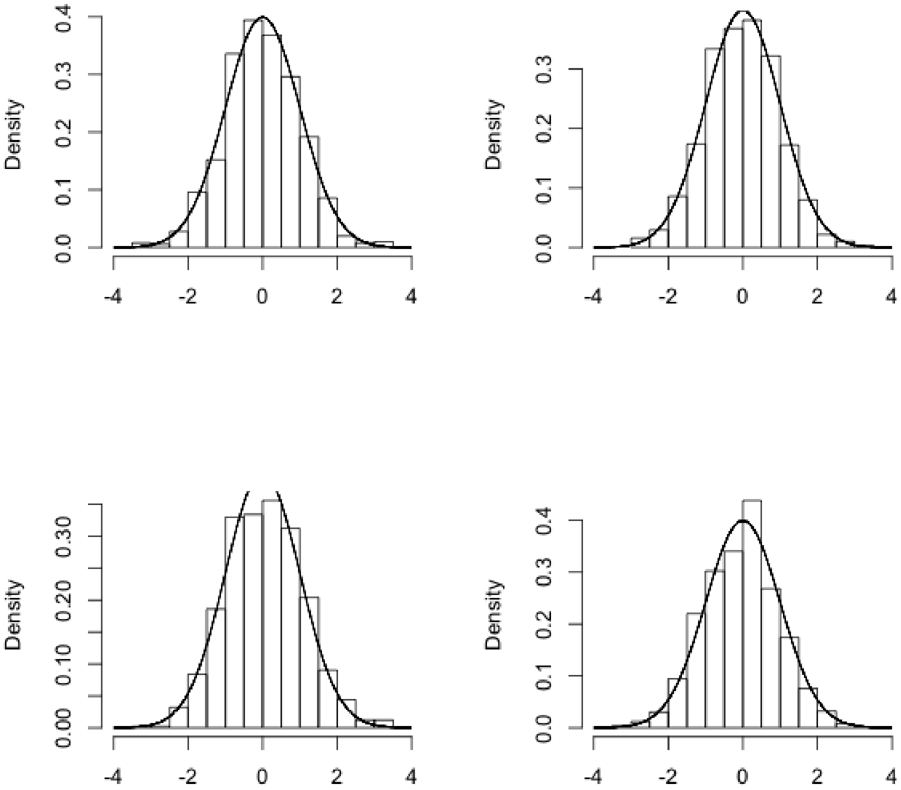
Then $P \sim \text{DP}(P_0, \alpha)$. Note that P is a discrete distribution here, with mass π_h at points θ_h drawn from the base distribution P_0. Generating samples from P involves sampling θ_h with replacement, with corresponding probability π_h. Values of α close to 0 lead to high probabilities on the first few π_h (low concentration at P_0), while high values of α lead to high concentration at P_0.

The histograms below represent four draws from $P \sim \text{DP}(P_0 \sim \mathcal{N}(0,1), \alpha = 1)$. Each draw constitutes a separate distribution $P(x) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(x)$, which is defined by our sampled π_h and θ_h. The histograms are represented by 1000 draws from P (after P itself is drawn). Since $\alpha = 1$, the distribution of P is heavily dominated by the first few values of π_h. The high spike in the northeast plot is due to $\pi_2 = 0.92$ and $\theta_2 = -0.51$.

Source code: second half of `Lab 5 2.3 Stick Break.R`



The next set of plots are four draws of P for $\alpha = 1000$. We see that P very much resembles P_0 in this case.



2.3.3 Posterior Calculation

If we have $\boldsymbol{x} \overset{\text{iid}}{\sim} G$ on Ω and $G \sim \text{DP}\left(G_0, \alpha\right)$, then for a finite measurable partition $A_1, \ldots, A_K$, we can use the conjugacy of the Dirichlet and Multinomial to get the posterior

$$\begin{aligned} G\left(A_1\right), \ldots, G\left(A_K\right) | \boldsymbol{x} &\sim \text{Dirichlet}\left(\alpha G_0\left(A_1\right) + n_1, \ldots, \alpha G_0\left(A_K\right) + n_K\right) \\ G | \boldsymbol{x} &\sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}\right) \end{aligned}$$

where $n_j = \sum_{i=1}^{n} \mathbb{I}\left(x_i \in A_j\right)$ is the number of x_i in partition A_j.

In general, computation of these models is difficult, with a complex MCMC. We see that Polya trees provide simpler posterior updating.

3 Polya Trees

3.1 Definition

I will be following notation from Müller and Rodriguez (2013). Consider the support Ω of the probability measure we are trying to estimate. At the level specified,

0. We have Ω itself.

1. We split Ω into B_0 and B_1.

2. We split B_0 into B_{00} and B_{01} and B_1 into B_{10} and B_{11}.

$m+1$. We split $B_{\varepsilon_1 \ldots \varepsilon_m}$ into $B_{\varepsilon_1 \ldots \varepsilon_m 0}$ and $B_{\varepsilon_1 \ldots \varepsilon_m 1}$ for each possible value of $\boldsymbol{\varepsilon}_m = \varepsilon_1 \ldots \varepsilon_m$.

Here $\boldsymbol{\varepsilon}_m$ is an m-digit number with digits $\varepsilon_1 \ldots \varepsilon_m \in \{0, 1\}$.

I added the subscript m here as a reminder that $\boldsymbol{\varepsilon}$ depends on m

Let $\prod$ be the set of partitions defined this way.

Now consider a marble falling through each branch of the tree, starting from the top. At each $B_{\boldsymbol{\varepsilon}_m}$, a decision is made to move into $B_{\boldsymbol{\varepsilon}_m 0}$ with probability $Y_{\boldsymbol{\varepsilon}_m 0} \sim \text{Beta}\left(\alpha_{\boldsymbol{\varepsilon}_m 0}, \alpha_{\boldsymbol{\varepsilon}_m 1}\right)$, independently across m. The probability of moving into $B_{\boldsymbol{\varepsilon}_m 0}$ is simply $1 - Y_{\boldsymbol{\varepsilon}_m 0}$. The values of $(\alpha_{\boldsymbol{\varepsilon}_m 0}, \alpha_{\boldsymbol{\varepsilon}_m 1})$ control how likely we are to move into either partition. Let $\mathcal{A}$ be the set of all α parameters. Then the Polya tree defines a random probability measure G by assigning any partitioning subset $B_{\boldsymbol{\varepsilon}_m}$ the random probability

$$G\left(B_{\boldsymbol{\varepsilon}_m}\right) = \prod_{j=1;\varepsilon_j=0}^{m} Y_{\varepsilon_1 \ldots \varepsilon_{j-1} 0} \prod_{j=1;\varepsilon_j=1}^{m} \left(1 - Y_{\varepsilon_1 \ldots \varepsilon_{j-1} 0}\right) \tag{*}$$

Underneath all this notation, the Polya tree model is characterized by two key parameters:

1. The set of partitions $\prod$ of Ω.

2. The non-negative parameters $\mathcal{A}$ for the Beta-distributed splitting probabilities.

3. We write $G \sim \text{PT}\left(\prod, \mathcal{A}\right)$.

4. G satisfies (Lavine, 1992)

   (i) All the $Y_{\boldsymbol{\varepsilon}_m}$ are independent.

   (ii) $Y_{\boldsymbol{\varepsilon}_m 0} \sim \text{Beta}\left(\alpha_{\boldsymbol{\varepsilon}_m 0}, \alpha_{\boldsymbol{\varepsilon}_m 1}\right)$.

   (iii) Equation $(^*)$.

One of the important features of the Polya tree prior is that it can generate continuous probability measures. A random probability measure $G \sim \text{PT}\left(\prod, \mathcal{A}\right)$ is absolutely continuous with probability 1 when the α parameters increase sufficiently fast with m. A popular choice is $\alpha_{\boldsymbol{\varepsilon}_m} = cm^2$. In our simulations we use $c = 1$. For α decreasing with respect to m, G can be almost surely discrete. When $\alpha_{\boldsymbol{\varepsilon}_m} = c/2^m$, the Polya tree prior reduces to the Dirichlet process prior.

One of the attractions of the Polya tree model is the ease of centering the model at any desired prior mean G_0. One way to accomplish this is to fix the partitioning subsets B as the dyadic quantiles of G_0. In our simulations I used $G_0 \sim \mathcal{N}(0,1)$. Note that this specification of G_0 is incorporated in the $\prod$ parameter.

3.2 Simulation

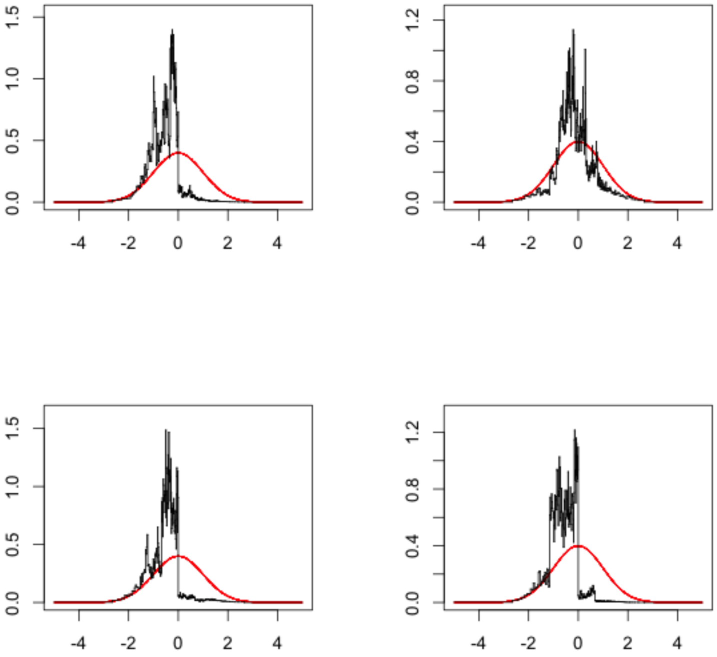
1. Choose some M to be the number of levels of the tree.
2. Choose a base G_0. Here we used $\mathcal{N}(0,1)$.
3. Partition Ω into the dyadic quantiles of G_0.
4. At each $B_{\boldsymbol{\varepsilon}_m}$, generate $Y_{\boldsymbol{\varepsilon}_m 0} \sim \text{Beta}\left(\alpha_{\boldsymbol{\varepsilon}_m 0} = m^2, \alpha_{\boldsymbol{\varepsilon}_m 1} = m^2\right), Y_{\boldsymbol{\varepsilon}_m 1} = 1 - Y_{\boldsymbol{\varepsilon}_m 0}$.
5. Compute the probabilities based on (*) to get $G\left(B_{\boldsymbol{\varepsilon}_M}\right)$ for all $B_{\boldsymbol{\varepsilon}_M}$ in the final row of the tree.

   These should sum to 1, and characterize the distribution we are trying to simulate.

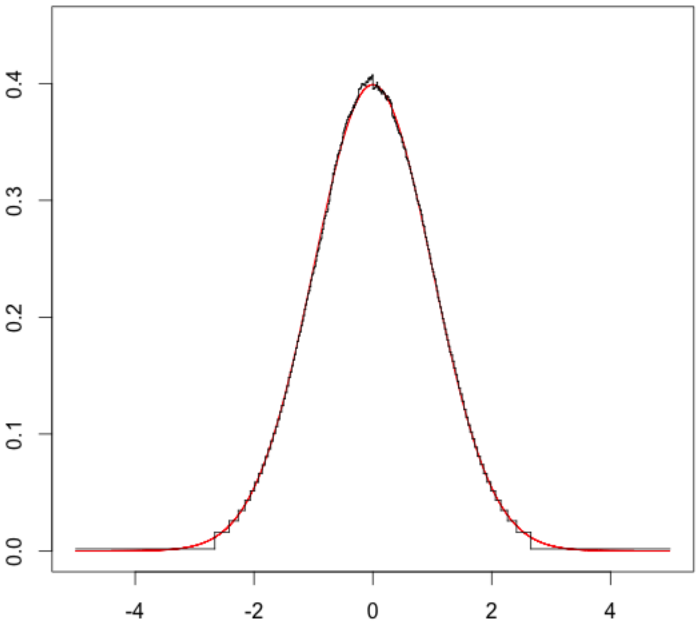
   Each draw of the Polya tree is a set of probabilities for being in each $B_{\boldsymbol{\varepsilon}_M}$, which we can interpret as bin heights (the actual bin height is the probability divided by the width of the bin).

Source code: `Lab 5 3 Polya Trees.R`

These are four draws from $G \sim \text{PT}\left(\prod, \mathcal{A}\right)$, where $\prod$ is defined by the dyadic quantiles of the $\mathcal{N}(0,1)$, and $\mathcal{A}$ is defined as m^2. There are $M = 8$ levels for 256 bins. That is, $\{B_{\boldsymbol{\varepsilon}_M}\}$ has 256 items.



When we take 5000 draws from $G \sim \mathrm{PT}\left(\prod, \mathcal{A}\right)$, we get 5000 sets of 256 probabilities (one for each bin). The plot below is their average. We note that this resembles the base distribution a lot more.



As m increases, $Y_{\boldsymbol{\varepsilon}_m}$ becomes closer and closer to a 50/50 coin flip, by the properties of the Beta distribution, whose expectation approaches 0.5 if its two parameters are equal and large.

3.3 Posterior Inference

The PT is conjugate under iid sampling. If we have $\boldsymbol{x} \overset{\text{iid}}{\sim} G$ with $G \sim \text{PT}\left(\prod, \mathcal{A}\right)$, then the posterior on the unknown probability measure G is again $G|\boldsymbol{x} \sim \text{PT}\left(\prod, \mathcal{A}^*\right)$, with

$$\alpha^*_{\boldsymbol{\varepsilon}_m} = \alpha_{\boldsymbol{\varepsilon}_m} + n_{\boldsymbol{\varepsilon}_m},$$

where $n_{\boldsymbol{\varepsilon}_m}$ is the number of observations that lie in $B_{\boldsymbol{\varepsilon}_m}$.

We would update step 4 in Section 3.2 as

4. At each $B_{\boldsymbol{\varepsilon}_m}$, generate $Y_{\boldsymbol{\varepsilon}_m 0} \sim \text{Beta}\left(\alpha_{\boldsymbol{\varepsilon}_m 0} = m^2 + n_{\boldsymbol{\varepsilon}_m 0}, \alpha_{\boldsymbol{\varepsilon}_m 1} = m^2 + n_{\boldsymbol{\varepsilon}_m 1}\right), Y_{\boldsymbol{\varepsilon}_m 1} = 1 - Y_{\boldsymbol{\varepsilon}_m 0}$.

Intuitively, if a lot of data fall under $B_{\boldsymbol{\varepsilon}_m 0}$, then the probabilities generated for going to that partition is higher, due to the large $n_{\boldsymbol{\varepsilon}_m 0}$. Alternatively, if a lot of data fall under $B_{\boldsymbol{\varepsilon}_m 1}$, then the probabilities generated for going to $B_{\boldsymbol{\varepsilon}_m 0}$ will be smaller, due to the large $n_{\boldsymbol{\varepsilon}_m 1}$. This in turn increases the probability of going to $B_{\boldsymbol{\varepsilon}_m 1}$.

BIO 249 Lab 6: Dirichlet Process

Halloween, 2016

Recommended reading: Bayesian Nonparametrics by Ghosh and Ramamoorthi, Introduction to the Dirichlet Distribution and Related Processes by Frigyik, Kapila, and Gupta.

Some Mathematical Review

Defn (Random Variable): A random variable on a probability space $(\Omega, \mathscr{F}, P)$ is a real-valued function $X = X(\omega)$ that is measurable $\mathscr{F}$. More precisely, $X : (\Omega, \mathscr{F}) \mapsto (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is a function from Ω to the real line, and for every $B \in \mathscr{B}(\mathbb{R})$, we have $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F}$.

Defn (Metric Space): A metric space $\mathbf{S}$, with a distance function or metric ρ, is a set that assigns a real number $\rho(x, y)$ to every pair $x, y \in \mathbf{S}$, where

1. $\rho(x, y) \geq 0$ and $\rho(x, y) = 0 \Leftrightarrow x = y$
2. $\rho(x, y) = \rho(y, x)$
3. $\rho(x, y) + \rho(y, z) \geq \rho(x, z), z \in \mathbf{S}$ (Triangle Inequality)

We denote a metric space with its metric as $(\mathbf{S}, \rho)$. Examples of metric spaces include $(\mathbb{R}, \rho(x, y) = |x - y|)$, and $\left(\mathbf{C}[0,1], d(f, g) = \int_0^1 |f(x) - g(x)|\, dx\right)$, where $\mathbf{C}[0,1]$ is the set of all continuous real functions on the interval $[0,1]$.

Defn (Open Ball): The open ball with center s_0 and radius δ is the set $B(s_0, \delta) = \{s \in \mathbf{S} : \rho(s_0, s) < \delta\}$. For a closed ball, the inequality is strict.

Defn (Separable): The metric space $\mathbf{S}$ is separable if it has a countable dense subset. The real line is a separable metric space because it has a countable dense subset (the rational numbers $\mathbb{Q}$). On the real line, any two real numbers can be separated by a rational number in $\mathbb{Q}$, which is some background on the name. Similarly, the metric space $\mathbb{R}^k$ is separable.

In a separable space $\mathbf{S}$, every open set is a countable union of balls. This fact fails to hold when $\mathbf{S}$ is not separable. The Borel σ-field $\mathscr{B}(\mathbf{S})$ on $\mathbf{S}$ is the one generated by all open sets. If $\mathbf{S}$ is separable then $\mathscr{B}(\mathbf{S})$ is the same as the σ-field generated by open balls. In the absence of separability these two σ-field would be different.

For convergence definitions, refer to pages 12 and 13 of the Ghosh and Ramamoorthi book.

Dirichlet Processes on the Real Line

To avoid confusion, I will distinguish between the Dirichlet distribution and process by abbreviating Dirichlet process to DP.

Proposition 1 (Gamma Representation)

If $Y_i \overset{\text{ind}}{\sim} \text{Gamma}(\alpha_i, 1)$, then $\left(\frac{Y_1}{Y}, \ldots, \frac{Y_K}{Y}\right) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$, and is independent of $Y = \sum_{i=1}^K Y_i$.

Proof: We are transforming $Y_1, \ldots, Y_K$ into $Y, Q_1, \ldots, Q_{K-1}$, and we would first like to find the joint distribution of $Y, Q_1, \ldots, Q_{K-1}$. The transformation itself is

$$\begin{aligned} Y_1, \ldots, Y_{K-1}, Y_K &\mapsto Y, Q_1, \ldots, Q_{K-1}, Q_K \\ Y_1, \ldots, Y_{K-1}, Y_K &= YQ_1, \ldots, YQ_{K-1}, Y\left(1 - \sum_{i=1}^{K-1} Q_i\right) \end{aligned}$$

Then the Jacobian is

$$J = \begin{bmatrix} \frac{\partial Y_1}{\partial Y} & \frac{\partial Y_1}{\partial Q_1} & \cdots & \frac{\partial Y_1}{\partial Q_{K-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_{K-1}}{\partial Y} & \frac{\partial Y_{K-1}}{\partial Q_1} & \cdots & \frac{\partial Y_{K-1}}{\partial Q_{K-1}} \\ \frac{\partial Y_K}{\partial Y} & \frac{\partial Y_K}{\partial Q_1} & \cdots & \frac{\partial Y_K}{\partial Q_{K-1}} \end{bmatrix} = \begin{bmatrix} Q_1 & Y & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ Q_{K-1} & 0 & \cdots & Y \\ 1 - \sum_{i=1}^{K-1} Q_i & -Y & \cdots & -Y \end{bmatrix}$$

with determinant Y^{k-1}. Starting with

$$f(y_1, \ldots, y_K) = \prod_{i=1}^K y_i^{\alpha_i - 1} \frac{e^{-y_i}}{\Gamma(\alpha_i)},$$

the change of variable formula gives

$$\begin{aligned} f\left(y, q_1, \ldots, q_{K-1}\right) &= \left[\prod_{i=1}^{K-1}\left(yq_i\right)^{\alpha_i-1}\frac{e^{-yq_i}}{\Gamma\left(\alpha_i\right)}\right]\left(\left[y\left(1-\sum_{i=1}^{K-1}q_i\right)\right]^{\alpha_K-1}\frac{e^{-y\left(1-\sum_{i=1}^{K-1}q_i\right)}}{\Gamma\left(\alpha_K\right)}\right)z^{k-1} \\ &= \frac{\prod_{i=1}^{K-1}q_i^{\alpha_i-1}\left(1-\sum_{i=1}^{K-1}q_i\right)^{\alpha_K-1}}{\prod_{i=1}^{K}\Gamma\left(\alpha_i\right)}y^{\sum_{i=1}^{K}\alpha_i-1}e^{-y} \end{aligned}$$

This is the product of a Dirichlet and Gamma density. Note that this formulation for the Dirichlet is slightly modified as we are writing q_K as $1-\sum_{i=1}^{K-1}q_i$. We can further integrate over y to get

$$f\left(q_1, \ldots, q_{K-1}, 1-\sum_{i=1}^{K-1}q_i\right) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma\left(\alpha_i\right)}\prod_{i=1}^{K-1}q_i^{\alpha_i-1}\left(1-\sum_{i=1}^{K-1}q_i\right)^{\alpha_K-1}$$

This result provides an easy way to generate the Dirichlet distribution from the Gamma distribution. We first generate $Y_i \overset{\text{iid}}{\sim} \text{Gamma}\left(\alpha_i, 1\right)$, then get

$$\left(\frac{Y_i}{\sum_{i=1}^{K}Y_i}, \ldots, \frac{Y_K}{\sum_{i=1}^{K}Y_i}\right) \sim \text{Dirichlet}\left(\alpha_1, \ldots, \alpha_K\right)$$

Proposition 2 (Aggregation Property)

If $\left(X_1, \ldots, X_K\right) \sim \text{Dir}\left(\alpha_1, \ldots, \alpha_K\right)$ and $Z_j = \sum_{i\in I_j}X_i$ for a given partition $I_1, \ldots, I_m$ of $\{1, \ldots, K\}$, then $\left(Z_1, \ldots, Z_M\right) \sim \text{Dir}\left(\sum_{i\in I_1}\alpha_i, \ldots, \sum_{i\in I_M}\alpha_i\right)$.

Dirichlet Distribution Review

Dirichlet distributions can be used to model the randomness of pmfs (more formally, discrete random probability measures). Each draw from a Dirichlet distribution is a vector of probabilities summing to 1, which can be used to define a pmf. Consider a pmf defined by a six-sided die. Real dice are not exactly uniformly weighted, due to the laws of physics and the reality of manufacturing. A bag of real dice is an example of a sample of random pmfs, and can be modelled with a Dirichlet distribution. Drawing one die from this bag is like obtaining one draw from the Dirichlet distribution. Each draw is a vector of six probabilities summing to 1. Old dice may have pmfs that deviate greatly from a uniform distribution, while newer ones may be close to perfect.

Extension to DP

The Dirichlet distribution is limited in that it assumes a finite set of events. In the previous analogy, the dice have a finite number of faces. The DP enables us to work with an infinite set of events, and hence to model probability distributions over infinite sample spaces. The DP is a random probability measure whose realizations are distributions over an arbitrary (possibly infinite) sample space. However, the set of all probability distributions over an infinite sample space is unmanageable. To deal with this, the DP restricts the class of distributions under consideration to a more manageable set: discrete probability distributions over the infinite sample space that can be written as an infinite sum of weighted indicator functions (see Defn 3). For some set B of the infinite sample space, a realization of the DP will assign probability $P\left(B\right) = \sum_{h=1}^{\infty} \pi_h \delta_{y_h}\left(B\right)$ for weights π_h summing to 1 and $y_h \overset{\text{iid}}{\sim} G_0$. Ferguson (1973) proves that $P\left(\cdot\right)$ is indeed a DP in Theorem 2 of his paper.

A Dirichlet process is a stochastic process, or collection of random variables, indexed over the σ-field $\mathscr{B}\left(\mathbb{R}\right)$. Any set $B \in \mathscr{B}\left(\mathbb{R}\right)$ has a corresponding random variable $\tilde{P}\left(B\right) \in \left[0,1\right]$. B and B^c form a partition of $\mathbb{R}$, and the random vector $\left(\tilde{P}\left(B\right), \tilde{P}\left(B^c\right)\right)$ has a Dirichlet distribution. A realization of the Dirichlet process is a probability measure $P : \mathscr{B}\left(\mathbb{R}\right) \mapsto \left[0,1\right]$. Even if the base distribution is continuous, the distributions drawn from the Dirichlet process are almost surely discrete (Ferguson, 1973). The concentration parameter specifies how strong this discretization is: in the limit of $\alpha \to 0$, the realizations are all concentrated at a single value, while in the limit of $\alpha \to \infty$, the realizations become continuous. Between the two extremes, the realizations are discrete distributions with less and less concentration as α increases.

DP Defn 1 (Ferguson Distribution)

When α is a scalar, $G \sim \text{DP}\left(\alpha, G_0\right)$ iff for every partition $\bigcup_{i=1}^{K} A_i = \mathbb{R}$, we have

$$\left(G\left(A_1\right), \ldots, G\left(A_K\right)\right) \sim \text{Dirichlet}\left(\alpha G_0\left(A_1\right), \ldots, \alpha G_0\left(A_K\right)\right)$$

That is, any finite partition of the sample space of a DP will have a Dirichlet distribution.

The more general definition specifies

$$\left(G\left(A_1\right), \ldots, G\left(A_K\right)\right) \sim \text{Dirichlet}\left(\alpha\left(A_1\right), \ldots, \alpha\left(A_K\right)\right)$$

for a general measure $\alpha(\cdot)$. In the simplified (first) case, we happen to set $\alpha(\cdot) = \alpha G_0(\cdot)$, where the α on the right side is a scalar (sorry for confusing notation).

DP Defn 2 (Gamma):

$G \sim \text{DP}(\alpha, G_0)$ iff for every partition $\bigcup_{i=1}^{K} A_i = \mathbb{R}$, we have

$$\begin{aligned}(G(A_1), \ldots, G(A_K)) &\sim \left(\frac{\Gamma(A_1)}{\sum_{i=1}^{K}\Gamma(A_i)}, \ldots, \frac{\Gamma(A_K)}{\sum_{i=1}^{K}\Gamma(A_i)}\right) \\ &\sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K)),\end{aligned}$$

where $\Gamma(A_i)$ are independent $\text{Gamma}(\alpha G_0(A_i), 1)$ random variables. The notation used here by Lorenzo can be a bit confusing. $\Gamma(A_i)$ does not represent the Gamma function here, and is instead just a letter representing a random variable. The (A_i) shows the dependence of Γ on A_i. This equivalence is a direct consequence of Proposition 1.

DP Defn 3 (Stick Breaking)

Let $V_1, V_2, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), w_h = v_h \prod_{\ell=1}^{m-1}(1 - v_\ell), \theta_h \overset{\text{iid}}{\sim} G_0$. Then a realization from the DP is defined as

$$G(A) = \sum_{h:\theta_h \in A} w_h$$

or equivalently, we can construct the realization as $P(x) = \sum_{h=1}^{\infty} w_h \mathbb{I}(x = \theta_h)$. The proof of this is provided by Ferguson (1973). This topic was covered in Lab 5.

DP Defn 4 (Poisson)

Lorenzo's Let $\{T_i\}$ be a non-homogeneous Poisson process on $[0, 1]$ with mean measure density

$$\lambda(t) = m \exp(-t)/t, t > 0, m > 0$$

Then a realization from the DP is defined as

$$G\left(A\right)=\sum_{h:\theta_h\in A}\frac{T_h}{\sum_{i=1}^{\infty}T_i}$$

Not much information on this definition, and it was not covered in great detail in class.

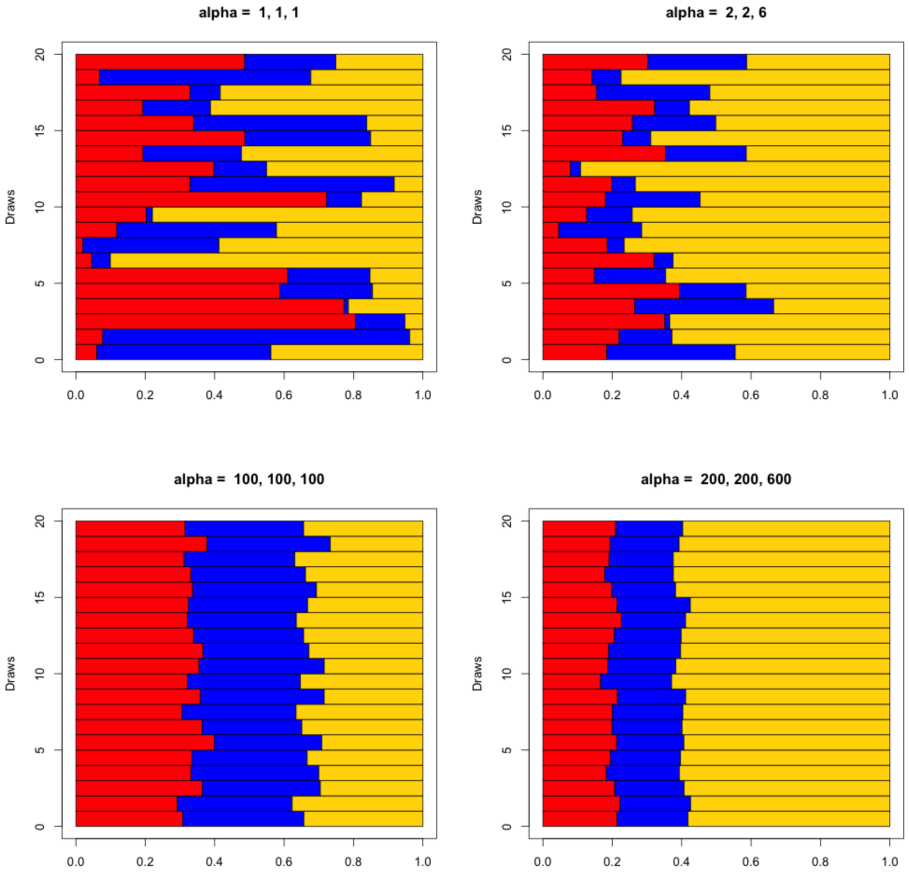
DP Defn 5 (Pólya Urn)

Pólya Urn to Generate from Dirichlet Distribution

Suppose we want to generate a realization of $(Q_1,\ldots,Q_K)\sim\text{Dirichlet}\left(\alpha_1,\ldots,\alpha_K\right)$. To start, put α_i balls of colour i for $i=1,\ldots,K$ in the urn, where the α_i do not need to be integers. At each iteration, draw one ball uniformly at random from the urn, and then place it back into the urn along with an additional ball of the same colour. As we iterate this procedure more and more times, the proportion of balls of each colour will converge to a pmf that is a sample from $\text{Dirichlet}\left(\alpha_1,\ldots,\alpha_K\right)$.

The Pólya urn method is the least efficient (compared to stick breaking or generating Gammas) because it depends on a convergence result. One needs to iterate the urn-drawing scheme many times to get good results, and for any finite number of iterations of the scheme, the resulting pmf is not perfectly accurate.

The plots below represent 20 draws of a Dirichlet using this method, for four different $\left(\alpha_1,\alpha_2,\alpha_3\right)$. The pmf is not perfectly accurate, as these are for 10000 iterations of the urn drawing method, but we can see that their results are close to what is expected.



Pólya Urn to Generate from DP

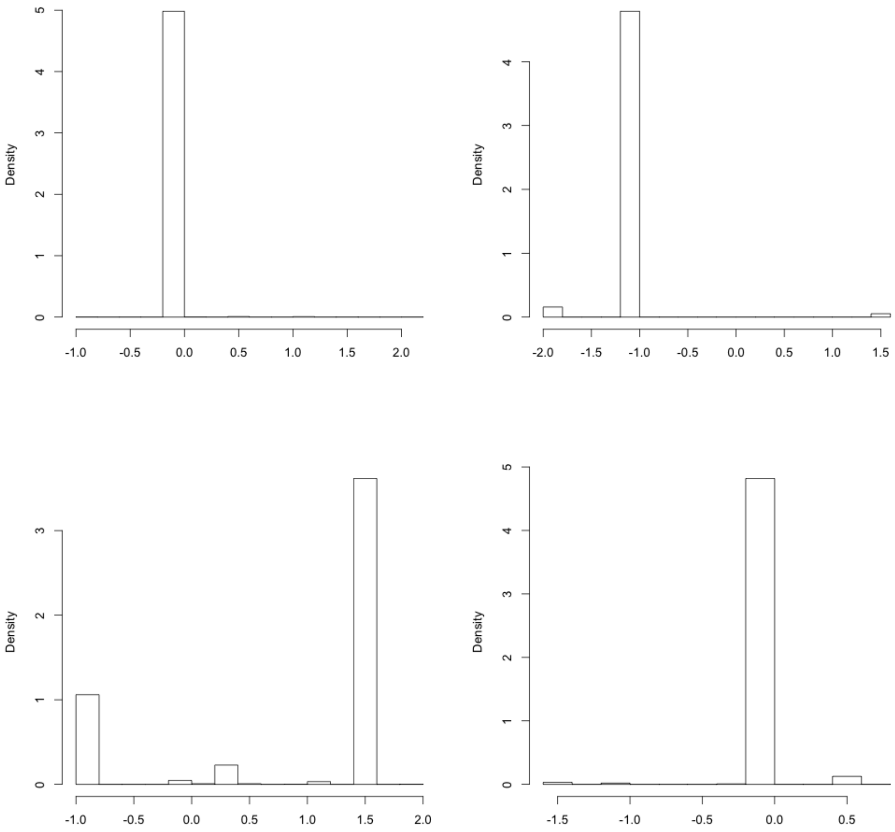
Like the stick-breaking procedure, we would like to generate sequences $\{\pi_h\}$ and $\{y_h\}$ to produce $P(x) = \sum_{h=1}^{\infty} w_h \mathbb{I}(x = \theta_h)$. $\{y_h\}$ is easy to generate from the base distribution. $\{\pi_h\}$ is problematic because they are dependent and sum to 1. The stick-breaking draws π_k exactly, but one at a time, and it takes an infinitely long time to generate one sample. If we stop early, we would have the first $k-1$ coefficients correct. In practice, depending on α, we eventually reach an h_0 such that $\pi_h = 0$ for all $h > h_0$ (equivalently $P(x)$ will have measure 0 for $\theta_h : h > h_0$), at which point we can stop.

For the DP, we have infinite colours. Here, a colour represents a draw y_h from G_0. Starting with an empty urn and $n = 1$, with probability $\frac{n}{n+\alpha}$, pick a ball out of the urn, and put it back with another ball of the same colour. Otherwise, with probability $\frac{\alpha}{n+\alpha}$, pick a new colour (from G_0), and add it to the urn. We will then have an infinite sequence of colours $(y_1, y_2, \dots)$, where the weight π_h is the proportion of balls in the urn of colour h. When α is small, there is a small chance of picking a new colour (choosing a new sample

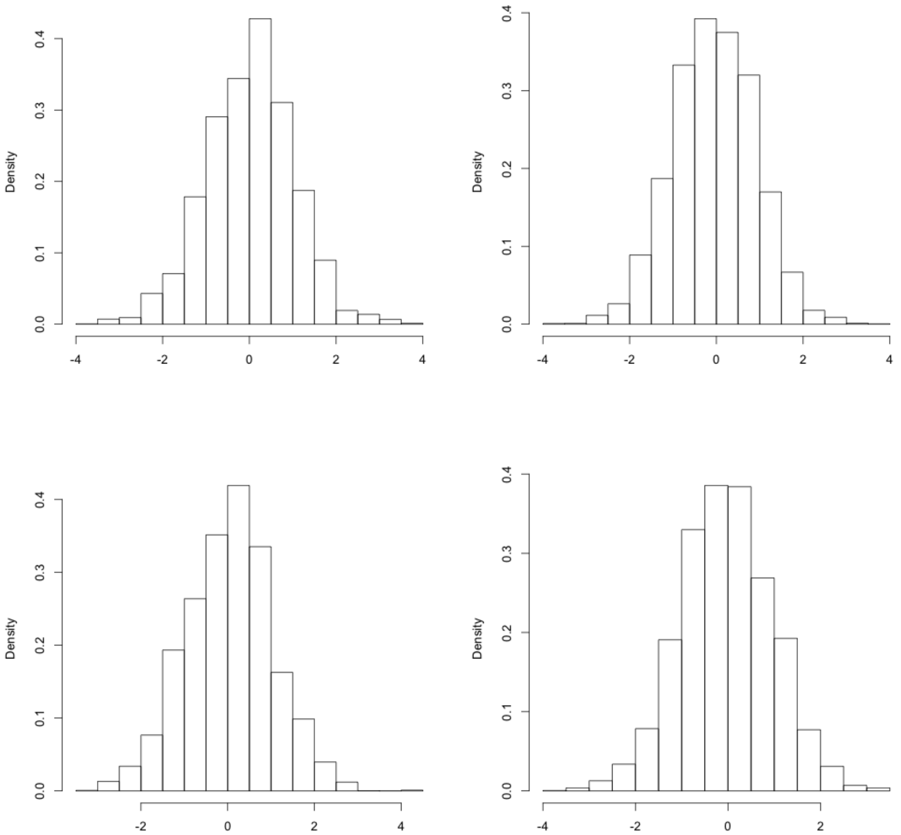
y_h) as n increases, so the first few ball colours tend to get repeatedly sampled, with a very small chance of getting a new ball colour (new y_h).

This is equivalent to the stick breaking example. We end up with an infinite sequence $\{\pi_h\}$ of proportions and an infinite sequence $\{y_h\}$ of colours.

Below are four draws of a DP $(\alpha = 1, \mathcal{N}(0,1))$. Instead of infinite colours, I used 10000.



Below are four draws of a DP $(\alpha = 1000, \mathcal{N}(0,1))$. Instead of infinite colours, I used 10000.



Blackwell and MacQueen 1973

In the nonparametric case, the parameter space $\boldsymbol{\Theta}$ is typically the set of all probability measures on $\mathcal{X}$. Denote the set of all probability measures on $\mathcal{X}$ by $M(\mathcal{X})$. We are interested when $\mathcal{X}$ is is a finite set, or when $\mathcal{X} = \mathbb{R}$. More generally, $\mathcal{X}$ is a complete separable metric space.

Let α be a finite measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. The general definition of DP is $G \sim \mathrm{DP}(\alpha)$ iff for every partition $\bigcup_{i=1}^{K} A_i = \mathbb{R}$, we have

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dirichlet}(\alpha(A_1), \ldots, \alpha(A_K))$$

We can normalize α to a probability measure $\frac{\alpha}{\alpha(\mathcal{X})}$ (this is our G_0 in the simplified notation).

Consider a Pólya urn with $\alpha(\mathcal{X})$ balls of which $\alpha(j)$ are colour $j, j = 1, \ldots, K$. Let $X_i = j$ if ball i is colour

j. Then

$$\begin{aligned} P\left(X_1 = j\right) &= \frac{\alpha\left(j\right)}{\alpha\left(\mathcal{X}\right)} \\ P\left(X_2 = j | X_1\right) &= \frac{\alpha\left(j\right) + \delta_{X_1}\left(j\right)}{\alpha\left(\mathcal{X}\right) + 1} \\ &\dots \\ P\left(X_{n+1} = j | X_1, X_2, \dots, X_n\right) &= \frac{\alpha\left(j\right) + \sum_{i=1}^{n} \delta_{X_i}\left(j\right)}{\alpha\left(\mathcal{X}\right) + n} = \frac{\alpha_n}{\alpha_n\left(\mathcal{X}\right)} \end{aligned}$$

Intuitively, for the $n+1$ ball to be colour j, we have $\alpha\left(\mathcal{X}\right) + n$ balls to draw from, and $\alpha\left(j\right)$ at the beginning of colour j, and $\sum_{i=1}^{n} \delta_{X_i}\left(j\right)$ added from the previous balls. Define $\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{X_i}$. Then $\{X_n, n \geq 1\}$ define a Pólya sequence with parameter α as defined in the paper. The authors show the equivalence between the DP and this Pólya urn scheme.

Proposition 3: The joint distribution of $X_1, X_2, \dots$ is exchangeable. First note that

$$P\left(X_1 = x_1, \dots, X_n = x_n\right) = \frac{\alpha\left(x_1\right)}{\alpha\left(\mathcal{X}\right)} \frac{\alpha\left(x_2\right) + \delta_{x_1}\left(x_2\right)}{\alpha\left(\mathcal{X}\right) + 1} \frac{\alpha\left(x_3\right) + \delta_{x_1}\left(x_3\right) + \delta_{x_2}\left(x_3\right)}{\alpha\left(\mathcal{X}\right) + 2} \dots$$

As we are dealing with balls in an urn, let n_j be the number of $X_i = j$ (number of balls of colour j). Additionally, define $m^{[n]} = m\left(m+1\right)\dots\left(m+n-1\right)$.

We can simplify the numerator to

$$\left[\alpha\left(1\right)\left(\alpha\left(1\right)+1\right)\dots\left(\alpha\left(1\right)+n_1-1\right)\right]\dots\left[\alpha\left(K\right)\left(\alpha\left(K\right)+1\right)\dots\left(\alpha\left(K\right)+n_K-1\right)\right]$$

or $\left[\alpha\left(1\right)\right]^{[n_1]}\dots\left[\alpha\left(K\right)\right]^{[n_K]}$. To see why, try an example where we have $K = 2$ colours and $n = 5$.

In the denominator, we have

$$\alpha\left(\mathcal{X}\right)\left(\alpha\left(\mathcal{X}\right)+1\right)\dots\left(\alpha\left(\mathcal{X}\right)+n-1\right) = \left[\alpha\left(\mathcal{X}\right)\right]^{[n]}$$

We can now conclude exchangeability by the commutative property of multiplication.

Theorem (From the Paper): Let $\{X_n\}$ be a Pólya sequence with parameter α. Then

(a) The sequence of measures $m_n = \frac{\alpha_n}{\alpha_n(\mathcal{X})}$ converges with probability 1 as $n \to \infty$ to a limiting discrete measure G.

(b) $G \sim \mathrm{DP}(\alpha)$.

(c) $X_1, X_2, \ldots | G$ are exchangeable.

Theorem (De Finetti): Let μ be a probability measure on $\mathbb{R}^\infty$. Then $X_1, X_2, \ldots | P$ is exchangeable iff there is a unique probability measure Π on $M(\mathbb{R})$ such that for all n and for any Borel sets $B_1, B_2, \ldots, B_n$,

$$\mu(X_1 \in B_1, \ldots, X_n \in B_n) = \int_{M(\mathbb{R})} \prod_{i=1}^{n} P(B_i)\, d\Pi(P)$$

The Pólya urn scheme described above leads to a sequence of exchangeable random variables, and the corresponding mixing measure Π coming out of De Finetti's theorem is precisely $\mathrm{DP}(\alpha)$.

Brief Summary

Ferguson (1973) showed that a realization of the DP characterized by Defn 1 can be written as $P(B) = \sum_{h=1}^{\infty} \pi_h \delta_{y_h}(B)$ for each $B \in \mathscr{B}(\mathbb{R})$, weights $\{\pi_h\}$, and atoms $\{y_h\}$. Blackwell and MacQueen provided an alternative and equivalent definition using Pólya urns. The interpretations are as follows

	Stick Breaking	Pólya Urn
π	parts of the stick	proportions of colours
y	from G_0	the generated colours from G_0

BIO 249 Lab 7: Dirichlet Process Mixture

November 7, 2016

Recommended reading: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems (Antoniak, 1974), Markov Chain Sampling Methods for Dirichlet Process Mixture Models (Neal, 2000).

Dirac Measure

Let $(\Omega, \mathscr{F})$ be a measurable space. Let $\omega \in \Omega$ be any point in Ω, and A any set in $\mathscr{F}$. The Dirac measure at ω, denoted δ_ω, is the measure defined by

$$\delta_\omega : \mathscr{F} \to \mathbb{R}, \delta_\omega(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Theorem 1 The Dirac measure is a (probability) measure.

Proof: We need to verify (i) $0 \leq \delta_\omega(A) \leq 1$ for all $A \in \mathscr{F}$, (ii) $\delta_\omega(\emptyset) = 0, \delta_\omega(\Omega) = 1$, and (iii) countable additivity.

Axioms (i) and (ii) are clear by construction of the Dirac measure.

For axiom (iii), let $A_1, A_2, \ldots$ be a disjoint sequence of $\mathscr{F}$-sets with $\bigcup_{k=1}^{\infty} A_k \in \mathscr{F}$.

Case 1: $\omega \in A_m$ for some m.

This means that $\omega \in A_m \subseteq \bigcup_{k=1}^{\infty} A_k$, so that

$$\delta_\omega\left(\bigcup_{k=1}^{\infty} A_k\right) = 1$$

Additionally, $\omega \notin A_k$ iff $k \neq m$, so that

$$\sum_{k=1}^{\infty} \delta_\omega(A_k) = \delta_\omega(A_m) + \sum_{k \neq m} \delta_\omega(A_k) = 1 + \sum_{k \neq m} 0 = 1$$

Then $\delta_\omega \left(\bigcup_{k=1}^\infty A_k\right) = \sum_{k=1}^\infty \delta_\omega \left(A_k\right)$, as required.

Case 2: $\omega \notin A_k$ for all k.

This means that $\omega \notin A_k$ for all k, and $\omega \notin \bigcup_{k=1}^\infty A_k$, so that

$$\delta_\omega \left(\bigcup_{k=1}^\infty A_k\right) = 0 = \sum_{k=1}^\infty \delta_\omega \left(A_k\right)$$

Then $\delta_\omega \left(\bigcup_{k=1}^\infty A_k\right) = \sum_{k=1}^\infty \delta_\omega \left(A_k\right)$, as required.

Theorem 2 The DP on $(\mathbb{R}, \mathscr{B})$ defined by $P(A) = \sum_{h=1}^\infty \pi_h \delta_{\theta_h}(A), A \in \mathscr{B}$, is a (probability) measure. Recall the condition that $\sum_{h=1}^\infty \pi_h = 1$.

Proof: We need to verify (i) $0 \leq P(A) \leq 1$ for all $A \in \mathscr{B}$, (ii) $P(\emptyset) = 0, P(\mathbb{R}) = 1$, and (iii) countable additivity.

Axiom (i)

By construction, $P(A) \geq 0$. Additionally, $P(A) = \sum_{h=1}^\infty \pi_h \delta_{\theta_h}(A) \leq \sum_{h=1}^\infty \pi_h (1) = 1$.

Axiom (ii)

$$P(\emptyset) = \sum_{h=1}^\infty \pi_h \delta_{\theta_h}(\emptyset) = \sum_{h=1}^\infty \pi_h (0) = 0$$
$$P(\mathbb{R}) = \sum_{h=1}^\infty \pi_h \delta_{\theta_h}(\mathbb{R}) = \sum_{h=1}^\infty \pi_h (1) = 1$$

Axiom (iii)

Let $A_1, A_2, \ldots$ be a disjoint sequence of $\mathscr{B}$-sets with $\bigcup_{k=1}^\infty A_k \in \mathscr{B}$.

$$\begin{aligned} P\left(\bigcup_{k=1}^\infty A_k\right) &= \sum_{h=1}^\infty \pi_h \delta_{\theta_h}\left(\bigcup_{k=1}^\infty A_k\right) \\ &= \sum_{h=1}^\infty \pi_h \sum_{k=1}^\infty \delta_{\theta_h}(A_k) \\ &= \sum_{h=1}^\infty \sum_{k=1}^\infty \pi_h \delta_{\theta_h}(A_k) \\ &= \sum_{k=1}^\infty \sum_{h=1}^\infty \pi_h \delta_{\theta_h}(A_k) \\ &= \sum_{k=1}^\infty P(A_k) \end{aligned}$$

We can switch the order of summation because $\sum_{h=1}^{\infty}\sum_{k=1}^{\infty}|\pi_h\delta_{\theta_h}(A_k)| \leq 1 < \infty$.

Mixture Models

We say that a distribution f is a (finite) mixture of K component distributions $f_1, \ldots, f_K$ if

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x)$$

with the λ_k being the mixing weights, $\lambda_k > 0, \sum_{k=1}^{K}\lambda_k = 1$. Generating data from $f(x)$ involves generating $z \sim \text{Discrete}(\lambda_1, \ldots, \lambda_K)$, followed by $x|z \sim f_z$. In practice, we typically deal with parametric mixture models, where f_k are from the same parametric family, but with different parameters. In that case, we write the mixture model as

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(x|\theta_k)$$

One way to view mixture models is through clusters defined by the component distributions. The idea is that all data points of the same type, belonging to the same cluster, are more or less equivalent and all come from the same distribution, and any differences between them are matters of chance. The classic estimation procedure for mixture models is the EM algorithm.

Dirichlet Process Mixture (DPM)

Consider a finite mixture model of the form $y_i \sim \sum_{k=1}^{K}\pi_k F(y|\theta_k)$. y is distributed as a mixture of distributions having the same parametric form F but differing in their parameters. Also let all the parameters θ_k be drawn from the same (base) distribution G_0. This mixture model can be expressed hierarchically as follows:

$$\begin{aligned} y_i|c_i, \mathbf{\Theta} &\sim F(y|\theta_{c_i}) \\ \theta_{c_i} &\sim G_0 \\ c_i|\pi_1, \ldots, \pi_K &\sim \text{Discrete}(\pi_1, \ldots, \pi_K) \\ (\pi_1, \ldots, \pi_K) &\sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \end{aligned}$$

The π_i are the mixture coefficients. They are the prior probability of a data point belonging to the k^{th} cluster. The latent variables c_i are labels assigning the y_i to a cluster from 1 to K with associated parameter

θ_{c_i}. We assume that the data belongs to K distinct clusters with means $\theta_1, \ldots, \theta_K$, since $c_i \in \{1, \ldots, K\}$. We assume no initial information distinguishing the clusters, which is indicated by the symmetric Dirichlet prior. We further assign independent and identical prior distributions G_0 to each of the cluster means θ_{c_i}.

Now consider the limit as $K \to \infty$. The Dirichlet distribution becomes a DP. This mixture model applies to data $y_1, \ldots, y_n$, which we regard as part of an indefinite exchangeable sequence, or equivalently, as being independently drawn from some unknown distribution. We model the distribution from which the y_i are drawn as a mixture of distributions of the form $F(\theta)$, with the mixing distribution over θ being G. We let the prior for this mixing distribution be a $\text{DP}(\alpha, G_0)$. This gives the following model (DP Mixture, or DPM):

$$\begin{aligned} y_i|\theta_i &\sim F(y|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

Clusters (tables, urn ball colours) θ_i are sampled from the distribution G, which is sampled from $\text{DP}(\alpha, G_0)$. y_i is generated from F defined by the cluster θ_i.

Fitting the model described above means finding the posterior distribution for $\boldsymbol{\theta}$.

Sampling using a DPM

The most direct approach to sampling for this model is to repeatedly draw values for each θ_i from its conditional distribution given both the data and the θ_j for $j \neq i$. This conditional distribution is obtained by combining the likelihood for θ_i that results from y_i having distribution $F(\theta_i)$, which will be written as $F(y_i, \theta_i)$, and the prior conditional on θ_{-i}, which is

$$\theta_i|\boldsymbol{\theta}_{-i} \sim \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} G_0$$

The conditional posteriors, given a data point y_i, can be calculated as

$$p(\theta_i|\boldsymbol{\theta}_{-i}, y_i) \propto \mathcal{L}(y_i|\theta_i)\, p(\theta_i|\boldsymbol{\theta}_{-i})$$

where $y_i|\theta_i \sim F(\theta_i)$ and $\theta_i|\boldsymbol{\theta}_{-i} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta(\theta_j) + \frac{\alpha}{n-1+\alpha}G_0$. This can be simulated with a Gibbs sampler.

Examples

- The Dirichlet Process Mixture (DPM) Model by Ananth Ranganathan
- Fast Food Application: Clustering the McDonald's Menu by Edwin Chen

BIO 249 Lab 8: Introduction to Decision Theory

November 14, 2016

Recommended reading: Berger Chapter 1-4

Notation and Background

The formal theory of statistical decision making is made up of four fundamental elements:

1. **State:** formalization of the underlying unknown reality that affects the decision process. This is done by considering all that is unknown but relevant for the decision maker. The *state of nature* can be represented by an entity s taking values on a state space $\mathcal{S}$. Typically, s will be a parameter $\boldsymbol{\theta}$ taking values on a parameter space $\boldsymbol{\Theta}$.

2. **Data:** formal model of the observations, a sample $\boldsymbol{x}$ taking values on a sample space $\mathcal{X}$. Typically, when experiments are performed to obtain information about $\boldsymbol{\theta}$, the experiments are designed so that the observations $\boldsymbol{x}$ are distributed according to some probability distribution with $\boldsymbol{\theta}$ as an unknown parameter.

3. **Action:** decisions are more commonly called *actions* in the literature. Based on the data $\boldsymbol{x}$, a decision rule has to choose an action $a = \delta\left(\boldsymbol{x}\right)$ among a set $\mathcal{A}$ of allowed actions. Formally, a decision rule is a function $\delta\left(\boldsymbol{x}\right) : \mathcal{X} \rightarrow \mathcal{A}$, specifying how actions are chosen, given $\boldsymbol{x}$. For estimation problems, usually $\mathcal{A} = \mathcal{S}$. In those cases, the decision rule will output estimates (guesses) of the true s.

4. **Loss:** formally expressed via a loss function $L\left(s, a\right) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, specifying the "cost" that is incurred when the true state of nature is s and the chosen decision is a. In other literature, the equivalent concept of utility $U\left(s, a\right) = -L\left(s, a\right)$ is used, specifying the "gain" that is obtained.

A statistical decision problem will be considered solved when a decision rule $\delta\left(\boldsymbol{x}\right)$ is chosen such that it achieves some sort of optimality criterion associated with the loss function. We will be involved with decision making in the presence of uncertainty. Hence the actual incurred loss $L\left(s, a\right)$ will never be known with certainty (at the time of decision making). A natural method method of proceeding in the face of this uncertainty is to consider the expected loss of making a decision, and then choose an optimal decision with respect to this expected loss.

Frequentist Decision Theory

Defn 1: The *frequentist risk* is defined as

$$R(\theta, \delta) = E_{\mathcal{X}}\left[L(\theta, \delta(\boldsymbol{x}))\,|\theta\right] = \int_{\mathcal{X}} L(\theta, \delta(\boldsymbol{x}))\, f(\boldsymbol{x}|\theta)\, dx$$

To a frequentist, it is desirable to use a decision rule δ with small (or optimal) $R(\theta, \delta)$. However, θ is unknown, so there is a problem in saying what "small" means. Moreover, each θ may lead to a different optimal decision rule.

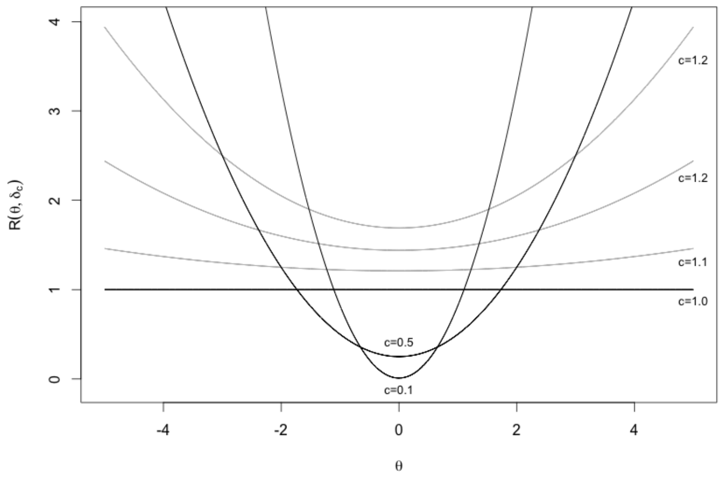
Defn 2: A decision rule δ_1 is *dominated* by another rule δ_0 if $R(\theta, \delta_0) \leq R(\theta, \delta_1)$ for all $\theta \in \Theta$, with strict inequality for at least one $\theta \in \Theta$.

Defn 3: A decision rule δ is said to be *admissible* if there exists no other rule that dominates it. Decision rules that are inadmissible should not be considered.

Ex 1: Consider an estimation problem where $X \sim \mathcal{N}(\theta, 1)$. Here, $\mathcal{A} = \mathcal{S} = \Theta = \mathbb{R}$, and $\mathcal{X} = \mathbb{R}$. Using the squared-error loss $L(\theta, \delta(x)) = (\theta - \delta(x))^2$, consider the decision rules $\delta_c(x) = cx, c \geq 0$. We are trying to estimate θ using one data point X, and the estimator cX for some scalar c. The frequentist risk is

$$\begin{aligned}
R(\theta, \delta_c) &= E_{\mathcal{X}}\left[L(\theta, \delta_c(X))\,|\theta\right] \\
&= E_{\mathcal{X}}\left[(\theta - cX)^2\,|\theta\right] \\
&= E_{\mathcal{X}}\left(\left[c(\theta - X) + (1-c)\theta\right]^2|\theta\right) \\
&= c^2 E_{\mathcal{X}}\left[(\theta - X)^2\,|\theta\right] + 2c(1-c)\,E_{\mathcal{X}}\left[(\theta - X)\,\theta|\theta\right] + (1-c)^2\, E_{\mathcal{X}}\left(\theta^2|\theta\right) \\
&= c^2 + (1-c)^2\,\theta^2
\end{aligned}$$

Each δ_c is a different decision rule (function over the true unknown value of θ).



For $c > 1, R(\theta, \delta_1) = 1 < R(\theta, \delta_c)$, as demonstrated above. Then δ_c is inadmissible for $c > 1$. Informally, the estimators cX for $c > 1$ perform worse (have higher squared-error loss) than X $(c = 1)$, for all $\theta \in \Theta$. On the other hand, we cannot make the same conclusion for $0 \leq c \leq 1$. In fact, for $0 \leq c \leq 1, \delta_c$ is admissible. Even a silly decision rule (estimator) $\delta_0(x) = 0$ is admissible. While admissibility may be a desirable property for a decision rule, it gives no assurance that the decision rule is reasonable.

Bayesian Decision Theory

The Bayesian approach to decision theory introduces another element: a priori knowledge concerning the state of nature θ, in the form of a probability function, usually referred to as the prior. A fundamental disadvantage of the frequentist risk is that it cannot be used to derive an optimal decision rule.

Defn 4: If $\pi(\theta)$ is the believed probability distribution of θ at the time of decision making (prior distribution), the (posterior) *Bayesian expected loss* is

$$\rho(\pi, \delta(\boldsymbol{x})) = E_\Theta[L(\theta, \delta(\boldsymbol{x}))|\boldsymbol{x}] = \int_\Theta L(\theta, \delta(\boldsymbol{x}))\, p(\theta|\boldsymbol{x})\, d\theta$$

Here $p(\theta|\boldsymbol{x})$ is the posterior distribution of θ given the data $\boldsymbol{x}$.

Defn 5: The *Bayes action* $\delta^*(\boldsymbol{x}) = \arg\min_{\delta \in \mathcal{A}} \rho(\pi, \delta(\boldsymbol{x}))$ for any fixed $\boldsymbol{x}$ is the decision that minimizes the Bayesian expected loss for each particular observation $\boldsymbol{x} \in \mathcal{X}$.

Ex 2: Consider $L(\theta, \delta(\boldsymbol{x})) = (\theta - \delta(\boldsymbol{x}))^2$. Then

$$\begin{aligned}\rho(\pi, \delta(\boldsymbol{x})) &= \int_\Theta (\theta - \delta(\boldsymbol{x}))^2\, p(\theta|\boldsymbol{x})\, d\theta \\ &= \delta^2(\boldsymbol{x}) - 2\delta(\boldsymbol{x}) \int_\Theta \theta p(\theta|\boldsymbol{x})\, d\theta + \int_\Theta \theta^2 p(\theta|\boldsymbol{x})\, d\theta\end{aligned}$$

Setting the first derivative to zero gives

$$\begin{aligned}\frac{\partial \rho(\pi, \delta(\boldsymbol{x}))}{\partial \delta(\boldsymbol{x})} &= 2\delta(\boldsymbol{x}) - 2\int_\Theta \theta p(\theta|\boldsymbol{x})\, d\theta = 0 \\ \delta^*(\boldsymbol{x}) &= \int_\Theta \theta p(\theta|\boldsymbol{x})\, d\theta \text{ the posterior mean}\end{aligned}$$

Ex 3: Consider $L(\theta, \delta(\boldsymbol{x})) = |\theta - \delta(\boldsymbol{x})|$. Then

$$\begin{aligned}\rho(\pi, \delta(\boldsymbol{x})) &= \int_\Theta |\theta - \delta(\boldsymbol{x})|\, p(\theta|\boldsymbol{x})\, d\theta \\ &= \int_{-\infty}^{\delta(\boldsymbol{x})} (\delta(\boldsymbol{x}) - \theta)\, p(\theta|\boldsymbol{x})\, d\theta + \int_{\delta(\boldsymbol{x})}^{\infty} (\theta - \delta(\boldsymbol{x}))\, p(\theta|\boldsymbol{x})\, d\theta\end{aligned}$$

Setting the first derivative to zero gives

$$\begin{aligned}\frac{\partial\rho\left(\pi,\delta\left(\boldsymbol{x}\right)\right)}{\partial\delta\left(\boldsymbol{x}\right)} &= \int_{-\infty}^{\delta(\boldsymbol{x})} p\left(\theta|\boldsymbol{x}\right)d\theta - \int_{\delta(\boldsymbol{x})}^{\infty} p\left(\theta|\boldsymbol{x}\right)d\theta = 0\\ \delta^*\left(\boldsymbol{x}\right) &= \text{the posterior median}\end{aligned}$$

Unifying these Ideas

Defn 6: The *Bayes risk* of a decision rule δ, with respect to a prior distribution π on Θ, is defined as

$$\begin{aligned}r\left(\pi,\delta\right) &= E_\theta\left[R\left(\theta,\delta\right)\right]\\ &= \int_\Theta R\left(\theta,\delta\right)\pi\left(\theta\right)d\theta\end{aligned}$$

This is the average of the frequentist risk over θ. It can be equivalently viewed as

$$\begin{aligned}r\left(\pi,\delta\right) &= \int_\Theta R\left(\theta,\delta\right)\pi\left(\theta\right)d\theta\\ &= \int_\Theta\int_\mathcal{X} L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)f\left(\boldsymbol{x}|\theta\right)dx\pi\left(\theta\right)d\theta\\ &= \int_\Theta\int_\mathcal{X} L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)f\left(\boldsymbol{x},\theta\right)dxd\theta\\ &= \int_\Theta\int_\mathcal{X} L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)p\left(\theta|\boldsymbol{x}\right)f\left(\boldsymbol{x}\right)dxd\theta\\ &= \int_\mathcal{X}\int_\Theta L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)p\left(\theta|\boldsymbol{x}\right)d\theta f\left(\boldsymbol{x}\right)dx\ \text{(Fubini)}\\ &= \int_\mathcal{X}\rho\left(\pi,\delta\left(\boldsymbol{x}\right)\right)f\left(\boldsymbol{x}\right)dx\\ &= E_\mathcal{X}\left[\rho\left(\pi,\delta\left(\boldsymbol{x}\right)\right)\right]\end{aligned}$$

Using the Bayes risk and the Bayesian expected loss lead to the same decision rule. The definitions are summarized in the table below.

<table>
<tr><td colspan="2">Loss $L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)$</td></tr>
<tr><td>Bayesian (condition on $\boldsymbol{x}$)</td><td>Frequentist (condition on θ)</td></tr>
<tr><td>Bayesian Expected Loss
$\rho\left(\pi,\delta\left(\boldsymbol{x}\right)\right) = E_\Theta\left[L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)|\boldsymbol{x}\right]$</td><td>Frequentist Risk
$R\left(\theta,\delta\right) = E_\mathcal{X}\left[L\left(\theta,\delta\left(\boldsymbol{x}\right)\right)|\theta\right]$</td></tr>
<tr><td>integrate out $\boldsymbol{x}$</td><td>integrate out θ</td></tr>
<tr><td colspan="2">Bayes Risk
$r\left(\pi,\delta\right) = E_\mathcal{X}\left[\rho\left(\pi,\delta\left(\boldsymbol{x}\right)\right)\right] = E_\theta\left[R\left(\theta,\delta\right)\right]$</td></tr>
</table>

Ex 4: Consider $L(\theta, \delta(\boldsymbol{x})) = (\theta - \delta(\boldsymbol{x}))^2$. Let $\hat{\theta} = \delta(\boldsymbol{x})$. The frequentist risk is

$$\begin{aligned} R(\theta, \delta_c) &= E_{\mathcal{X}}\left[L\left(\theta, \hat{\theta}\right)|\theta\right] \\ &= E_{\mathcal{X}}\left[\left(\theta - \hat{\theta}\right)^2 |\theta\right] \end{aligned}$$

This is the mean squared error of the decision rule (estimator) $\hat{\theta}$.

Recall from Ex 2 that the Bayes action for this loss function is the posterior mean $\tilde{\theta} = \int_\Theta \theta p(\theta|\boldsymbol{x})\, d\theta$. The Bayesian expected loss using that Bayes action is

$$\begin{aligned} \rho(\pi, \delta(\boldsymbol{x})) &= E_\Theta\left[L\left(\theta, \tilde{\theta}\right)|\boldsymbol{x}\right] \\ & E_\Theta\left[\left(\theta - \tilde{\theta}\right)^2 |\boldsymbol{x}\right] \\ &= \int_\Theta \left(\theta - \tilde{\theta}\right)^2 p(\theta|\boldsymbol{x})\, d\theta \text{ the posterior variance} \end{aligned}$$

Ex 5: Consider Ex 1 again, now knowing that $\pi(\theta) \sim \mathcal{N}(0, \tau^2)$. Then, for decision rule δ_c,

$$\begin{aligned} r(\pi, \delta) &= E_\Theta[R(\theta, \delta)] \\ &= E_\Theta\left[c^2 + (1-c)^2\theta^2\right] \\ &= c^2 + (1-c)^2\tau^2 \end{aligned}$$

We would like to minimize this Bayes risk with respect to δ (c in this case), so setting the first derivative to zero gives

$$\frac{\partial r}{\partial c} = 2c - 2(1-c)\tau^2 = 0$$
$$c = \frac{\tau^2}{1+\tau^2}$$

Then

$$c = \frac{1}{1+\tau^{-2}}$$

We can easily check that $c \to 0$ as $\tau \to 0$ and $c \to 1$ as $\tau \to \infty$. We will show that we get the same optimal decision when we try to minimize the Bayesian expected loss.

To find the Bayesian expected loss, we need $p(\theta|x) \sim \mathcal{N}\left(\frac{x}{1+\tau^{-2}}, \frac{1}{1+\tau^{-2}}\right)$ by conjugacy. Then

$$\begin{aligned} \rho(\pi, \delta_c(x)) &= E_\Theta[L(\theta, \delta_c(x))|x] \\ &= E_\Theta\left[(\theta - cx)^2|x\right] \\ &= E_\Theta\left[\left(\theta^2 - 2cx\theta + c^2x^2\right)|x\right] \\ &= E_\Theta\left(\theta^2|x\right) - 2cxE(\theta|x) + c^2x^2 \\ &= \frac{1}{1+\tau^{-2}} + \left(\frac{x}{1+\tau^{-2}}\right)^2 - 2cx\frac{x}{1+\tau^{-2}} + c^2x^2 \end{aligned}$$

We would like to minimize this Bayes risk with respect to δ (c in this case), so setting the first derivative to zero gives

$$\begin{aligned} \frac{\partial \rho}{\partial c} = -2\frac{x^2}{1+\tau^{-2}} + 2x^2 c = 0 \\ c = \frac{1}{1+\tau^{-2}} \end{aligned}$$

Once again, we can easily check that $c \to 0$ as $\tau \to 0$, and $c \to 1$ as $\tau \to \infty$. The interpretation is that as the prior for θ gets more concentrated at its mean 0 $(\tau \to 0)$, the optimal decision approaches the rule $\delta_0(x) = 0$. As the prior for θ gets more vague, the optimal decision approaches the rule $\delta_1(x) = x$, which dominates the decision rules $\{\delta_c(x)\}_{c>1}$, as shown in Ex 1.

BIO 249 Lab 9: Bayesian Sequential Analysis

November 28, 2016

Recommended reading: Berger Chapter 7

Background on Sequential Analysis

Experimental design is the choice of experiment. Being as this choice must usually be made before the data (and hence the posterior distribution) can be obtained, the subject is frequently called preposterior analysis by Bayesians. The general problem of experimental design is very complicated. We will content ourselves with a study of the simplest design problem, that of deciding when to stop sampling.

Defn: Assume random variables $X_1, X_2, \ldots$ are available for observation, and define $\boldsymbol{X}^j = (X_1, \ldots, X_j)$. Typically, the X_i are independent observations from a common density $f(x|\theta)$, so that $f_j(\boldsymbol{x}^j) = \prod_{i=1}^{j} f(x_i|\theta)$. Then we say that $X_1, X_2, \ldots$ is a *sequential sample* from the density $f(x|\theta)$.

Under this scenario, the observations can be taken in stages, or sequentially. After observing some $\boldsymbol{X}^j$, the experimenter has the option of either making an immediate decision or taking further observations. In addition to the loss associated with a decision (Lab 8), we now have an additional experimental cost, which is the cost of taking observations in this simplified setting.

We restrict ourselves to studying the two most common methods of taking observations. The first is the fixed sample size method, in which one preselects a sample size n, observes $\boldsymbol{X}^n$, and makes a decision. The second method is that of sequential analysis, in which the observations are taken one at a time, with a decision being made, after each observation, to either cease sampling (and choose an action) or take another observation.

Fixed Sample Size Example

Assume that $X_1, X_2, \ldots$ is a sequential sample from a $\mathcal{N}(\theta, 1)$ density, and that it is desired to estimate θ under squared error loss. The parameter θ is thought to have a $\mathcal{N}(0, \tau^2)$ prior density where τ^2 is known. If $\boldsymbol{X}^n$ is to be observed, then

$$\pi(\theta|\boldsymbol{x}^n) \sim \mathcal{N}\left(\frac{\sum_{i=1}^n x_i}{\frac{1}{\tau^2}+n}, \left(\frac{1}{\tau^2}+n\right)^{-1}\right)$$

From Lab 8, the Bayes action is the posterior mean, and the Bayes risk is the posterior variance, or

$$r(\pi, E(\theta|\boldsymbol{x}^n)) = \left(\frac{1}{\tau^2}+n\right)^{-1}$$

In the sequential problem, we have the additional component of cost per observation. Consider two cases.

Case 1: Each observation costs c, so that $C(n) = nc$. Then our Bayes risk is

$$r(\pi, E(\theta|\boldsymbol{x}^n)) + C(n) = \left(\frac{1}{\tau^2}+n\right)^{-1} + nc$$

Setting the derivative equal to zero gives

$$c = \left(\frac{1}{\tau^2}+n\right)^{-2}, n^* = \frac{1}{\sqrt{c}} - \frac{1}{\tau^2}$$

For an uninformative prior, the optimal n^* is roughly inversely proportional to $\sqrt{c}$. This makes sense as a higher cost per observation should lower the number we would want in our optimal design. Note that $n^* < 0$ corresponds to making a decision without taking observations. The smallest attainable Bayes risk is then

$$\begin{aligned} r\left(\pi, E\left(\theta|\boldsymbol{x}^{n^*}\right)\right) + C(n^*) &= \left(\frac{1}{\tau^2}+n^*\right)^{-1} + n^*c \\ &= \sqrt{c} + c\left(\frac{1}{\sqrt{c}} - \frac{1}{\tau^2}\right) \\ &= 2\sqrt{c} - \frac{c}{\tau^2} \end{aligned}$$

Case 2: $C(n) = \log(1+n)$, which makes it more affordable to take larger samples. Then our Bayes risk is

$$r\left(\pi, E\left(\theta|\boldsymbol{x}^n\right)\right) + C\left(n\right) = \left(\frac{1}{\tau^2} + n\right)^{-1} + \log\left(1+n\right)$$

Setting the derivative equal to zero gives

$$\left(\frac{1}{\tau^2} + n\right)^{-2} = \frac{1}{1+n}, n^* = \frac{1}{2} - \frac{1}{\tau^2} + \sqrt{5 - \frac{4}{\tau^2}}$$

We can analytically see that $n^* \to \frac{1}{2} + \sqrt{5} \doteq 2.736$ as $\tau^2 \to \infty$.

Toy Sequential Example

A manufacturing firm is trying to decide whether to build a new plant in Ohio (action a_0) or in Alabama (action a_1). The plant would cost \$1,000,000 less to build at the site in Alabama, but there is perhaps a lack of skilled labor in the area, whereas the site in Ohio has an abundance of skilled labor. A total of 700 skilled workers are needed, and the company feels that the size of the available skilled labor force near the site in Alabama, has a $\theta \sim \mathcal{N}\left(350, 100^2\right)$ prior density (we treat θ as a continuous variable for convenience). The company will have to train workers if skilled workers are not available, at a cost of \$3,500 each. Assuming the company has an approximately linear utility function for money, the decision loss can be written as

$$L\left(\theta, a\right) = \begin{cases} 10^6 & \text{if } a = a_0 \\ 3500\left(700 - \theta\right) & \text{if } a = a_1, 0 \leq \theta \leq 700 \\ 0 & \text{if } a = a_1, \theta > 700 \end{cases}$$

We can think of this as a sequential analysis before we collect any data $\boldsymbol{x}$ that provides more information on θ. In this scenario, the action space $\mathcal{A} = \{a_0, a_1\}$ is discrete, and we would like to find the action a that minimizes our Bayes risk. For a_0 this Bayes risk is

$r(\pi, a_0) = 10^6$. For a_1,

$$\begin{aligned} r(\pi, a_1) &= \int_\Theta L(\theta, a_1)\, \pi(\theta)\, d\theta \\ &= 3500 \int_0^{700} (700 - \theta)\, \pi(\theta)\, d\theta \\ &\doteq 3500 \int_{-\infty}^{-\infty} (700 - \theta)\, \pi(\theta)\, d\theta \\ &= 3500\,[700 - E(\theta)] \\ &= 1.225 \times 10^6 \end{aligned}$$

Here, $L(\theta, a_1)$ has non-zero mass only for $0 \leq \theta \leq 700$, which explains the bounds of integration. Based on $r(\pi, a_0) < r(\pi, a_1)$, if we were to make an immediate decision, we would choose to build in Ohio (a_0), with a Bayes risk of $1,000,000$.

Suppose now that the company can either make an immediate decision, or can commission a survey to be conducted (at a cost of \$20,000), the result of which would be an estimate, X, of θ. It is known that the accuracy of the survey would be such that $X \sim \mathcal{N}(\theta, 30^2)$. The problem is to decide whether to commission the survey (i.e., to decide whether to make an immediate decision, or to take the observation and then make a decision).

By conjugacy, the posterior distribution is

$$\begin{aligned} \pi_x = \pi(\theta|x) &\sim \mathcal{N}\left(\frac{\frac{350}{100^2} + \frac{x}{30^2}}{\frac{1}{100^2} + \frac{1}{30^2}}, \left(\frac{1}{100^2} + \frac{1}{30^2}\right)^{-1}\right) \\ &\dot{\sim} \mathcal{N}\left(28.8991 + 0.9174x, 28.7348^2\right) \end{aligned}$$

This time, for a_1, the Bayesian expected loss is (including the survey cost)

$$\begin{aligned} \rho(\pi_x, a_1) &= \int_\Theta L(\theta, a_1)\, \pi(\theta|x)\, d\theta + 20000 \\ &= 3500 \int_0^{700} (700 - \theta)\, \pi(\theta|x)\, d\theta + 20000 \\ &\doteq 3500 \int_{-\infty}^{-\infty} (700 - \theta)\, \pi(\theta|x)\, d\theta + 20000 \\ &= 3500\,[700 - E(\theta|x)] + 20000 \\ &= 3500\,(671.1009 - 0.9174x) + 20000 \\ &= 3500\,(676.8152 - 0.9174x) \end{aligned}$$

Compared to $\rho\left(\pi_x, a_0\right) = 10^6 + 20000$, the break-even point is $x = 420.0857$. If we actually go ahead with the survey and get a known value for x, we would select the action with the lower Bayesian expected loss. However, we are doing this analysis before collecting such data, and need to average over the distribution of x. To calculate the Bayes risk, we need the marginal distribution of x, or

$$f\left(x\right) = \int_{\Theta} f\left(x|\theta\right) \pi\left(\theta\right) d\theta \sim \mathcal{N}\left(350, 100^2 + 30^2\right)$$

Finally, the Bayes risk is approximately

$$\begin{aligned} & E_{\mathcal{X}}\left[\min\left\{\rho\left(\pi_x, a_0\right), \rho\left(\pi_x, a_1\right)\right\}\right] \\ =& \int_{-\infty}^{420.0857} \left(10^6 + 20000\right) f\left(x\right) dx + \int_{420.0857}^{\infty} 3500\left(676.8152 - 0.9174x\right) f\left(x\right) dx \\ =& 969,725 \end{aligned}$$

This amount is less than the Bayes risk of an immediate decision, so it would be worth conducting the survey.

Another Example

Upper management of the Oklahoma City Thunder are concerned about attendance for the upcoming year after the departure of their star player Dion Waiters. They must decide whether or not to implement a half-million dollar promotional campaign (a_1 to implement, a_0 to not implement). If the team is a contender, they feel that \$4 million in attendance revenues will be earned (regardless of whether or not the promotional campaign is implemented). Letting θ denote the team's proportion of wins, they feel the team will be a contender if $\theta \geq 0.6$. If $\theta < 0.6$, they feel their attendance revenues will be $1 + 5\theta$ million dollars without the promotional campaign, and $2 + \frac{10}{3}\theta$ million dollars with the promotional campaign. It is felt that θ has a $\mathrm{U}\left(0, 1\right)$ distribution.

The utility functions for our two decisions can be written as

$$U(\theta, a_0) = \begin{cases} 4 & \text{if } \theta \geq 0.6 \\ 1 + 5\theta & \text{if } \theta < 0.6 \end{cases} = \min\{4, 1 + 5\theta\}$$

$$U(\theta, a_1) = \begin{cases} 3.5 & \text{if } \theta \geq 0.6 \\ 1.5 + \frac{10}{3}\theta & \text{if } \theta < 0.6 \end{cases} = \min\left\{3.5, 1.5 + \frac{10}{3}\theta\right\}$$

We will do this problem using utility instead of loss. For a_0, our Bayes utility is

$$\begin{aligned} r(\pi, a_0) &= \int_\Theta U(\theta, a_0)\, d\theta \\ &= \left[\int_0^{0.6} (1 + 5\theta)\, d\theta + \int_{0.6}^1 4 d\theta\right] \\ &= 1.5 + 1.6 = 3.1 \end{aligned}$$

For a_1, our Bayes utility is

$$\begin{aligned} r(\pi, a_1) &= \int_\Theta U(\theta, a_1)\, d\theta \\ &= \left[\int_0^{0.6} \left(1.5 + \frac{10}{3}\theta\right) d\theta + \int_{0.6}^1 3.5 d\theta\right] \\ &= 1.5 + 1.4 = 2.9 \end{aligned}$$

If we were to make an immediate decision, we would choose not to implement the campaign.

Assume now that the Thunder acquire Chad, a PhD student who can help in the prediction of the team's winning proportion θ. Due to lack of funding, Chad is only able to report one data point $X \sim \text{Bern}(\theta)$. How much money would management expect Chad's report to be worth? Note that we are ignoring the cost of analyzing results (Chad's time is worth nothing).

By conjugacy, the posterior distribution is

$$\pi_x = \pi(\theta|x) \sim \text{Beta}(x + 1, 2 - x)$$

$$\begin{aligned} \pi(\theta|x) &= \frac{\Gamma(3)}{\Gamma(x+1)\Gamma(2-x)} \theta^x (1-\theta)^{1-x} \\ &= 2\theta^x (1-\theta)^{1-x}, x = 0, 1 \end{aligned}$$

To calculate the Bayes utility, we need the marginal distribution of x, or

$$\begin{aligned}
f(x) &= \int_{\Theta} f(x|\theta)\,\pi(\theta)\,d\theta \\
&= \int_{\Theta} \theta^{x}(1-\theta)^{1-x}\,d\theta \\
&= \frac{\Gamma(x+1)\,\Gamma(2-x)}{\Gamma(3)} \\
&= \frac{1}{2} \text{ for } x = 0, 1
\end{aligned}$$

This time, for a_0, the Bayesian expected utility is

$$\begin{aligned}
\rho(\pi_x, a_0) &= \int_{\Theta} U(\theta, a_0)\,\pi(\theta|x)\,d\theta \\
&= \int_0^{0.6} (1+5\theta)\,\pi(\theta|x)\,d\theta + \int_{0.6}^{1} 4\pi(\theta|x)\,d\theta \\
&= \int_0^{0.6} 2\theta^{x}(1-\theta)^{1-x}\,d\theta + 5\int_0^{0.6} 2\theta^{x+1}(1-\theta)^{1-x}\,d\theta + 4\int_{0.6}^{1} 2\theta^{x}(1-\theta)^{1-x}\,d\theta \\
&= \begin{cases} 0.36 + 0.72 + 2.56 = 3.64 & \text{if } x = 1 \\ 0.84 + 1.08 + 0.64 = 2.56 & \text{if } x = 0 \end{cases}
\end{aligned}$$

For a_1, the Bayesian expected utility is

$$\begin{aligned}
\rho(\pi_x, a_1) &= \int_{\Theta} U(\theta, a_1)\,\pi(\theta|x)\,d\theta \\
&= \left[\int_0^{0.6} \left(1.5 + \frac{10}{3}\theta\right)\pi(\theta|x)\,d\theta + \int_{0.6}^{1} 3.5\pi(\theta|x)\,d\theta\right] \\
&= \left[1.5\int_0^{0.6} 2\theta^{x}(1-\theta)^{1-x}\,d\theta + \frac{10}{3}\int_0^{0.6} 2\theta^{x+1}(1-\theta)^{1-x}\,d\theta + 3.5\int_{0.6}^{1} 2\theta^{x}(1-\theta)^{1-x}\,d\theta\right] \\
&= \begin{cases} 0.54 + 0.48 + 2.24 = 3.26 & \text{if } x = 1 \\ 1.26 + 0.72 + 0.56 = 2.54 & \text{if } x = 0 \end{cases}
\end{aligned}$$

If $x = 1$, our Bayes expected utility is 3.64; if $x = 0$, our Bayes expected utility is 2.56. Averaging over the distribution of x yields a Bayes utility of 3.1. This is the same as our Bayes utility for an immediate decision. This is because the data point x changes us from a flat prior (rectangle on $[0,1] \times [0,1]$) to either a posterior of 2θ for $x = 1$ or $2(1-\theta)$ for $x = 0$, each with a 50/50 chance of occurring. By the symmetry of these distributions, and the symmetry of the marginal distribution of x, there is no additional information provided by our point x.

BIO 249 Lab 10: Survival Analysis

December 5, 2016

Recommended reading: Notes 17, Chapter 1-3 of Bayesian Survival Analysis by Ibrahim, Chen, and Sinha.

Notation

Let T be a continuous non-negative random variable representing the survival times of an event of interest. Survival data are often right censored - the times T are known for only a portion of the individuals under study, and the remainder of the survival times are known only to exceed certain values. Let the censoring time be C. A typical key assumption is that the censoring mechanism is non-informative. That is, T and C are independent, possibly conditional on a set of covariates $\boldsymbol{X}$.

The observable data is $Y = \min\{T, C\}$, and we use the indicator $\delta = \mathbb{I}(T \leq C)$ to identify if T or C is being observed. That is, the data in this framework can be represented by the variables $(Y_i, \delta_i, \boldsymbol{x}_i)$ for some covariates $\boldsymbol{x}_i$.

Motivation for Bayes

1. Survival models are generally quite hard to fit, especially in the presence of complex censoring schemes. With the use of the Gibbs sampler and other MCMC techniques, fitting complex survival models is fairly straightforward.

2. MCMC sampling enables us to make exact inference for any sample size without resorting to asymptotic calculations. In the frequentist paradigm, variance estimates, for example, usually require asymptotic arguments which can be quite complicated to derive and in some models are simply not available. There is always the issue of whether the sample size is large enough for the asymptotic approximation to be valid.

3. The Bayesian paradigm enables us to incorporate prior information in a natural way. For example, when we have historical data, as is the case with most clinical trials, we can use priors to formally incorporate this data into the current analysis. For many models, frequentist inference can be obtained as a special case of Bayesian inference with many types of non-informative priors.

For more, see Section 1.8 of the Ibrahim textbook.

Background

To characterize the distribution of T, we consider four functions, each of which fully characterizes the underlying distribution of T.

1. Probability density function $f\left(t\right)$
2. Survival function $S\left(t\right) = 1 - F\left(t\right)$
3. Hazard function $\lambda\left(t\right) = \frac{f(t)}{S(t)}$
4. Cumulative hazard function $\Lambda\left(t\right) = \int_0^t \lambda\left(s\right) ds$

Given observables $\left(Y_i, \delta_i\right)$, we can formulate the likelihood as

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{y}|\boldsymbol{\theta}\right) &= \prod_{i=1}^{n} f\left(y_i|\boldsymbol{\theta}\right)^{\delta_i} S\left(y_i|\boldsymbol{\theta}\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda\left(y_i|\boldsymbol{\theta}\right)^{\delta_i} S\left(y_i|\boldsymbol{\theta}\right) \end{aligned}$$

Exponential Model

The Exponential distribution with parameter θ is defined as

$$\begin{aligned} f\left(t\right) &= \theta \exp\left(-\theta t\right) \\ S\left(t\right) &= \exp\left(-\theta t\right) \\ \lambda\left(t\right) &= \theta \\ \Lambda\left(t\right) &= \theta t \end{aligned}$$

The likelihood is

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{y}|\theta\right) &= \prod_{i=1}^{n} \lambda\left(y_i|\theta\right)^{\delta_i} S\left(y_i|\theta\right) \\
&= \theta^{\sum_{i=1}^{n}\delta_i} \exp\left(-\theta\sum_{i=1}^{n} y_i\right)
\end{aligned}$$

Consider a Gamma $\left(\alpha_0, \theta_0\right)$ prior. Then

$$\begin{aligned}
p\left(\theta|\boldsymbol{y}\right) &\propto \mathcal{L}\left(\boldsymbol{y}|\theta\right)\pi\left(\theta\right) \\
&\propto \theta^{\sum_{i=1}^{n}\delta_i} \exp\left(-\theta\sum_{i=1}^{n} y_i\right)\theta^{\alpha_0-1}\exp\left(-\theta_0\theta\right) \\
&= \theta^{\alpha_0+\sum_{i=1}^{n}\delta_i-1}\exp\left[-\left(\theta_0+\sum_{i=1}^{n} y_i\right)\theta\right] \\
&\sim \text{Gamma}\left(\alpha_0+\sum_{i=1}^{n}\delta_i, \theta_0+n\bar{y}\right)
\end{aligned}$$

And the posterior predictive distribution of a single future failure time $\tilde{y}$ is given by

$$\begin{aligned}
p\left(\tilde{y}|\boldsymbol{y}\right) &= \int \mathcal{L}\left(\tilde{y}|\theta\right)p\left(\theta|\boldsymbol{y}\right)d\theta \\
&\propto \int \theta\exp\left(-\theta\tilde{y}\right)\theta^{\alpha_0+\sum_{i=1}^{n}\delta_i-1}\exp\left[-\left(\beta_0+n\bar{y}\right)\theta\right]d\theta \\
&= \int \theta^{\alpha_0+\sum_{i=1}^{n}\delta_i}\exp\left[-\left(\tilde{y}+\beta_0+n\bar{y}\right)\theta\right]d\theta \\
&= \frac{\Gamma\left(\alpha_0+\sum_{i=1}^{n}\delta_i+1\right)}{\left(\beta_0+n\bar{y}+\tilde{y}\right)^{\alpha_0+\sum_{i=1}^{n}\delta_i+1}} \\
&\propto \left(\beta_0+n\bar{y}+\tilde{y}\right)^{-\left(\alpha_0+\sum_{i=1}^{n}\delta_i+1\right)} \\
&\sim \text{Inverse Beta}
\end{aligned}$$

To build a regression model, we introduce covariates $\boldsymbol{x}_i$ through θ, and write $\theta_i = \varphi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$, where $\varphi\left(\cdot\right)$ is commonly $\exp\left(\cdot\right)$. In that case, the likelihood function is

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{y}|\boldsymbol{\beta}\right) &= \prod_{i=1}^{n} \lambda\left(y_i|\theta\right)^{\delta_i} S\left(y_i|\theta\right) \\
&= \exp\left(\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{\beta}\delta_i\right)\exp\left[-\sum_{i=1}^{n}\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)y_i\right]
\end{aligned}$$

Common prior distributions for $\boldsymbol{\beta}$ include the uniform improper prior $\pi(\boldsymbol{\beta}) \propto 1$, and a normal prior. In general, closed forms for the posterior distribution of $\boldsymbol{\beta}$ are not available in the regression setting. However, MCMC techniques can be used to fit these types of models.

Weibull Model

The Weibull distribution with parameters α and γ is defined as

$$\begin{aligned}
&f(t) = \alpha\gamma t^{\alpha-1}\exp\left(-\gamma t^{\alpha}\right), t > 0, \alpha > 0, \gamma > 0 \\
&S(t) = \exp\left(-\gamma t^{\alpha}\right) \\
&\lambda(t) = \gamma\alpha t^{\alpha-1} \\
&\Lambda(t) = \gamma t^{\alpha}
\end{aligned}$$

Setting $\alpha = 1$ gives the Exponential distribution. It is often more convenient to write the model in terms of the parameterization $\theta = \log(\gamma)$, leading to

$$\begin{aligned}
f(t|\alpha,\theta) &= \alpha t^{\alpha-1}\exp\left(\theta - \exp(\theta)\,t^{\alpha}\right) \\
S(t|\alpha,\theta) &= \exp\left(-\exp(\theta)\,t^{\alpha}\right)
\end{aligned}$$

The likelihood is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}|\alpha,\theta) &= \prod_{i=1}^{n}\lambda(y_i|\theta)^{\delta_i} S(y_i|\theta) \\
&= \left[\exp(\theta)\,\alpha\right]^{\sum_{i=1}^{n}\delta_i}\prod_{i=1}^{n}\left(y_i^{\alpha-1}\right)^{\delta_i}\exp\left(-\sum_{i=1}^{n}\left[\exp(\theta)\,y_i^{\alpha}\right]\right) \\
&= \alpha^{\sum_{i=1}^{n}\delta_i}\exp\left(\theta\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\left[\delta_i(\alpha-1)\log(y_i)\right] - \exp(\theta)\sum_{i=1}^{n}y_i^{\alpha}\right) \\
&= \alpha^{\sum_{i=1}^{n}\delta_i}\exp\left(\log(\gamma)\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\left[\delta_i(\alpha-1)\log(y_i)\right] - \gamma\sum_{i=1}^{n}y_i^{\alpha}\right)
\end{aligned}$$

If α is known, the conjugate prior for γ is the Gamma distribution. Consider a Gamma (α_0, γ_0) prior for γ. Then

$$\begin{aligned}
p(\gamma|\boldsymbol{y}) &\propto \mathcal{L}(\boldsymbol{y}|\gamma)\,\pi(\gamma) \\
&\propto \exp\left[\log(\gamma)\sum_{i=1}^{n}\delta_i - \gamma\sum_{i=1}^{n} y_i^{\alpha}\right]\gamma^{\alpha_0-1}\exp(-\gamma_0\gamma) \\
&= \gamma^{\alpha_0+\sum_{i=1}^{n}\delta_i-1}\exp\left[-\left(\gamma_0+\sum_{i=1}^{n} y_i^{\alpha}\right)\gamma\right] \\
&\sim \text{Gamma}\left(\alpha_0+\sum_{i=1}^{n}\delta_i, \gamma_0+\sum_{i=1}^{n} y_i^{\alpha}\right)
\end{aligned}$$

No joint conjugate prior is available when (α, θ) are both assumed unknown. In this case, a typical joint prior specification is to take α and θ to be independent, where $\alpha \sim$ Gamma (α_0, κ_0) and $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$. In that case,

$$\begin{aligned}
p(\alpha,\theta|\boldsymbol{y}) &\propto \mathcal{L}(\boldsymbol{y}|\alpha,\theta)\,\pi(\alpha)\,\pi(\theta) \\
&\propto \alpha^{\sum_{i=1}^{n}\delta_i}\exp\left(\theta\sum_{i=1}^{n}\delta_i+\sum_{i=1}^{n}\left[\delta_i(\alpha-1)\log(y_i)\right]-\sum_{i=1}^{n}\left[\exp(\theta)\,y_i^{\alpha}\right]\right) \\
&\quad \alpha^{\alpha_0-1}\exp(-\kappa_0\alpha)\exp\left[-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right] \\
&\propto \alpha^{\alpha_0+\sum_{i=1}^{n}\delta_i-1} \\
&\quad \exp\left[\theta\sum_{i=1}^{n}\delta_i+(\alpha-1)\sum_{i=1}^{n}\delta_i\log(y_i)-\exp(\theta)\sum_{i=1}^{n}y_i^{\alpha}-\kappa_0\alpha-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right]
\end{aligned}$$

The joint posterior does not have a closed form, but this is a scenario where Gibbs sampling is simple to perform.

To build the Weibull regression model, we introduces covariates $\boldsymbol{x}_i$ through θ, and write $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Common prior distributions for $\boldsymbol{\beta}$ include the uniform improper prior, and a normal prior.

Simple Numerical Example of DP Prior

I saw this example in the Ibrahim text and thought it would be good to go over as a potential numeric problem on an exam.

Let F have non-negative support. Suppose $y_1, \ldots, y_5 \overset{\text{iid}}{\sim} F$ where $F \sim \text{DP}\left(\alpha, G_0\right), \alpha = 0.1$, and $G_0 \sim \text{Exp}\left(1\right)$ so that $G_0\left(y\right) = 1 - e^{-y}$. Further, suppose $y_1 = 1, y_2 = 0.7, y_3 = 0.8, y_4 = 1.2, y_5 = 1.3$, and let $B_1 = \left(0, 1\right], B_2 = \left(1, 1.25\right]$, and $B_3 = \left(1.25, \infty\right)$. Then we have three unknown values $p_i = F\left(B_i\right)$ for $i = 1, 2, 3$, where $p_1 + p_2 + p_3 = 1$, with prior

$$\begin{aligned}\left(p_1, p_2, p_3\right) &\sim \text{Dirichlet}\left(\alpha G_0\left(B_1\right), \alpha G_0\left(B_1\right), \alpha G_0\left(B_1\right)\right) \\ &\sim \text{Dirichlet}\left(0.0632, 0.0081, 0.0287\right)\end{aligned}$$

Recall that the posterior distribution of F is

$$\begin{aligned}F|\boldsymbol{y} &\sim \text{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{y_i}}{n}\right) \\ &\sim \text{DP}\left(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{y_i}}{\alpha + n}\right)\end{aligned}$$

where

$$\left(F\left(B_1\right), F\left(B_2\right), F\left(B_3\right)\right)|\boldsymbol{y} \sim \text{Dirichlet}\left(\alpha G_0\left(B_1\right) + \sum_{i=1}^{n} \delta_{y_i}\left(B_1\right), \ldots, \alpha G_0\left(B_3\right) + \sum_{i=1}^{n} \delta_{y_i}\left(B_3\right)\right)$$

Then the posterior distribution is

$$\begin{aligned}\left(p_1, p_2, p_3\right)|\boldsymbol{y} &\sim \text{Dirichlet}\left(\alpha G_0\left(B_1\right) + 3, \alpha G_0\left(B_2\right) + 1, \alpha G_0\left(B_3\right) + 1\right) \\ &\sim \text{Dirichlet}\left(3.0632, 1.0081, 1.0287\right)\end{aligned}$$

Note that this answer depends on the partition of $\left(0, \infty\right)$.

Lab Discussion: Dirichlet Process in Survival Analysis

Let T be a positive random variable with distribution F and let C be independent of T with distribution G. Our interest is in the posterior distribution of T given $\left(Y_i, \delta_i\right)$ where $Y = \min\left\{T, C\right\}$. To do this, we need a prior distribution for F. The posterior can be represented in many ways. An elegant approach comes from viewing a Dirichlet process as a neutral to the right prior (Doksum, 1974).

Neutral to the right priors (Ghosh and Ramamoorthi Chapter 10) are a class of non-parametric priors that were introduced by Doksum in 1974. Doksum showed that if a prior Π is neutral to the right, then the posterior given n observations is also neutral to the right.

Defn: A prior Π is said to be *neutral to the right* if, under Π, for all $k \geq 1$ and all $0 < t_1 < \cdots < t_k$,

$$S\left(t_1\right), \frac{S\left(t_2\right)}{S\left(t_1\right)}, \ldots, \cdots \frac{S\left(t_k\right)}{S\left(t_{k-1}\right)}$$

are independent. The DP is neutral to the right.

Recommended Reading: Ghosh and Ramamoorthi, Chapter 9-10

BIO 249 Lab 11: Survival Analysis 2: Exponential Boogaloo

December 12, 2016

Recommended reading: Dec 6 Lecture Notes by Lorenzo.

Notation

Let W_i be the time to event, with censoring time C_i. We let $W_i|F \overset{\text{iid}}{\sim} F$ and $C_i|H \overset{\text{iid}}{\sim} H$. We make the assumption of non-informative censoring, or $F \perp\!\!\!\perp G$. Let δ_i be the censoring indicator such that $\delta_i = \mathbb{I}\left(T_i \geq C_i\right)$. That is, $\delta_i = 1$ for censored observations and $\delta_i = 0$ for uncensored observations. This is the opposite of the Lab 10 (textbook) notation, but isn't very consequential. Let the observed time be $Y_i = \min\left(W_i, C_i\right)$, so that the observable data are $\left(Y_i, \delta_i\right), 0 \leq i \leq n$. Of interest is the posterior distribution of F given the data.

Discrete Example

For simplicity, we deal with $t \in \{1, 2, \dots\}$. The more general form would be $t \in \{t_1, t_2, \dots\}$. Then

$$\begin{aligned} f(t) &= \mathbb{P}(W = t) \\ F(t) &= \mathbb{P}(W \leq t) = \sum_{x=1}^{t} f(x) \\ S(t) &= \mathbb{P}(W \geq t) = \sum_{x=t}^{\infty} f(x) = 1 - F(t-1) \end{aligned}$$

Notationally, we write

$$F(\{t\}) = f(t) = \mathbb{P}(W = t) = F(t) - F(t-1) = S(t-1) - S(t)$$

Define the "time risk" (hazard function) at time j as

$$R_j = \frac{f(j)}{S(j)} = \frac{F(\{j\})}{1 - F(j-1)} = \mathbb{P}(W = j | W \geq j)$$

so that

$$\begin{aligned} S(j) &= \prod_{x=1}^{j-1}(1-R_x) \\ F(\{j\}) = R_j S(j) &= R_j \prod_{x=1}^{j-1}(1-R_x) \end{aligned}$$

Consider two individuals. The first has failure time at 3, and the second has a censored observation at time 2. That is, $C_1 > 3, C_2 = 2$, with $(Y_1 = 3, \delta_1 = 0), (Y_2 = 2, \delta_2 = 1)$. We can formulate the likelihood of our two data points as

$$\begin{aligned} \mathcal{L}(y_1, y_2, \delta_1, \delta_2 | F) = \mathbb{P}(W_1 = 3, W_2 \geq 2) &= F(\{3\}) \times [1 - F(2)] \\ \mathcal{L}(y_1, y_2, \delta_1, \delta_2 | R_1, R_2, R_3) &= (1-R_1)(1-R_2)R_3 \times (1-R_1) \\ &= (1-R_1)^2(1-R_2)R_3 \end{aligned}$$

We can either put a prior on F, or directly on R_j, which completely characterizes F. It is more convenient to put a prior on R_j. Because R_j is a probability in the discrete case, we can put a Beta prior on it, let it be $\text{Beta}(1,2)$. If we use the same prior for each j, then the prior mean for each hazard point is centered on same value (analogous to the constant hazard corresponding to the exponential distribution). However, this use of iid priors does not make sense in practice.

Let $\mathscr{D} = (y_1, y_2, \delta_1, \delta_2)$. We formulate the posterior distribution as

$$\begin{aligned} p(R_1|\mathscr{D}) &\propto (1-R_1)^2\left[R_1^{1-1}(1-R_1)^{2-1}\right] \\ &= R_1^{1-1}(1-R_1)^{4-1} \sim \text{Beta}(1,4) \\ p(R_2|\mathscr{D}) &\propto (1-R_2)\left[R_2^{1-1}(1-R_2)^{2-1}\right] \\ &= R_2^{1-1}(1-R_2)^{3-1} \sim \text{Beta}(1,3) \\ p(R_3|\mathscr{D}) &\propto R_3\left[R_3^{1-1}(1-R_3)^{2-1}\right] \\ &= R_3^{2-1}(1-R_3)^{2-1} \sim \text{Beta}(2,2) \end{aligned}$$

From these posterior distributions, we can estimate

1. $\mathbb{E}(R_1|\mathscr{D}) = \frac{1}{5}$

2. $\mathbb{E}\left[F(t)\,|\mathscr{D}\right] = 1 - \mathbb{E}\left[S(t+1)\,|\mathscr{D}\right] = 1 - \mathbb{E}\left[\prod_{x=1}^{t}(1-R_x)\,|\mathscr{D}\right] = 1 - \prod_{x=1}^{t}\mathbb{E}\left[(1-R_x)\,|\mathscr{D}\right],$ where the last equality is due to the fact that R_j has a neutral to the right prior, and will thus have a neutral to the right posterior.

3. $\mathbb{V}\left[F(t)\,|\mathscr{D}\right] = \mathbb{E}\left[F^2(t)\,|\mathscr{D}\right] - \mathbb{E}\left[F(t)\,|\mathscr{D}\right]^2,$ where $\mathbb{E}\left[F(t)\,|\mathscr{D}\right]^2$ is computed in the previous example. Here we also need

$$\begin{aligned}\mathbb{E}\left[F^2(t)\,|\mathscr{D}\right] &= \mathbb{E}\left(\left[1-F(t)\right]^2|\mathscr{D}\right) + 2\mathbb{E}\left[F(t)\,|\mathscr{D}\right] - 1\\ &= \mathbb{E}\left[S^2(t+1)\,|\mathscr{D}\right] + 2\mathbb{E}\left[F(t)\,|\mathscr{D}\right] - 1\\ &= \mathbb{E}\left(\left[\prod_{x=1}^{t}(1-R_x)\right]^2|\mathscr{D}\right) + 2\mathbb{E}\left[F(t)\,|\mathscr{D}\right] - 1\\ &= \mathbb{E}\left[\prod_{x=1}^{t}(1-R_x)^2\,|\mathscr{D}\right] + 2\mathbb{E}\left[F(t)\,|\mathscr{D}\right] - 1\end{aligned}$$

Continuous Example

Recommended reading: Section 9.2 from the Ghosh textbook.

Let the data be $(1, 4+, 3, 2+)$, or $(Y_1 = 1, \delta_1 = 0), (Y_2 = 4, \delta_2 = 1), (Y_3 = 3, \delta_1 = 0), (Y_4 = 2, \delta_2 = 1)$. Like before, let $W_i|F \overset{\text{iid}}{\sim} F$ and $C_i|H \overset{\text{iid}}{\sim} H$. In this scenario, we would like to place a DP prior on $F \sim \text{DP}(m, G_0)$ where $G_0 \sim \text{Exp}(\tau)$ for some known τ to facilitate calculations. Once again we would like to compute $\mathbb{E}\left[F(t)\,|\mathscr{D}\right]$ and $\mathbb{V}\left[F(t)\,|\mathscr{D}\right]$.

Begin with calculating $\mathbb{P}(W_1 = 1, W_2 > 4, W_3 = 3, W_4 > 2)$. To evaluate the posterior given the data, first look at the posterior given all uncensored observations.

We will describe the DP using the Polya Urn representation. $\mathbb{P}(W_1 = 1, W_3 = 3) = g_0(1)\frac{m}{m+1}g_0(4)$. Here we know that W_1 and W_3 are different draws from the Polya Urn (1 and 3 are different "colours"). Recall (Lab 6) that the probability of these being different (probability of a new colour) is $\frac{m}{m+1}$. In this scenario, 1 is the existing colour in the urn, and 3 is the new colour, which is drawn with probability $\frac{m}{m+1}$.

By the properties of DP, $F|W_1, W_3 \sim \text{DP}\left(m+2, \frac{mG_0 + \sum_{i=1,3}\delta_{W_i}}{m+2}\right) \sim \text{DP}(m', G_0')$. Then

$$P(W_2 > 4, W_4 > 2|W_1 = 1, W_3 = 4) = \left[1 - G_0'(4)\right]\frac{m'}{m'+1}\left[1 - G'(2)\right] + \frac{1}{m'+1}\left[1 - G_0'(4)\right]$$

Once again we have a Polya Urn interpretation. Due to the censoring, we cannot know if W_2 and W_4 are the same. Note in this example, we have updated our parameters with W_1 and W_3 and starting a "new" Polya Urn with these updated parameters. With probability $\frac{m'}{m'+1}$, W_2 will be different from W_4, and with probability $\frac{1}{m'+1}$, they will be the same. If W_2 and W_4 are the same, we are essentially asking the probability of drawing the one (censored above 4) colour from the urn, which occurs with probability $1 - G_0'(4)$.

This example from class seems rather incomplete. For a more general example, see pages 237-240 of Ghosh.

Neutral to the Right Priors

Recommended reading: Chapter 10 from the Ghosh textbook.

Defn (Neutral to the Right): Define arbitrary time points $0 \leq t_0 < t_1 < \ldots$ Let $F(t)$ denote a random distribution function (with a DP prior) we wish to estimate. For $t > s$, let $F(t|s) = \mathbb{P}(T \leq t | T \geq s) = \frac{F(t)-F(s)}{1-F(s)} = 1 - \frac{S(t)}{S(s)}$. We say F is *neutral to the right* if

$$F(t_1|t_0), F(t_2|t_1), \ldots$$

are independent random variables (or equivalently, if $\frac{S(t_1)}{S(t_0)}, \frac{S(t_2)}{S(t_1)}, \ldots$ are independent random variables). Doksum showed that the posterior distribution of F is also neutral to the right. Our simple discrete model has the neutral to the right property by the independence of the R_j.

Ex. (Gamma Process): Consider a Gamma process. For any t_0, t_1, t_2, the increment in the first region is a Gamma$(1, t_1 - t_0)$, which is independent of the increment in the second region Gamma$(1, t_2 - t_1)$. Let $R(t)$ be the random function representing the cumulative hazard. Define $S(t) = \exp[-R(t)]$. Then $S(t)$ is a valid survival function (decreasing, between 0 and 1, approaches 0). Additionally, $S(t)$ is neutral to the right. To see this, let Δ_j be the increment between t_j and t_{j-1}. Then

$$F(t_1|t_0) = 1 - \frac{S(t_1)}{S(t_0)} = 1 - \exp(-[R(t_1) - R(t_0)]) = 1 - \exp(-\Delta_1)$$
$$F(t_2|t_1) = 1 - \exp(-\Delta_2)$$

and so on. We get independence of $F(t_1|t_0), F(t_2|t_1), \ldots$ through the independence of the Δ_j.

Supplementary Reading

- For a formal proof of Doksum, see pages 256-257 of Ghosh (or Doksum's paper).
- For the distinction between tail-free and neutral to the right, see pages 70-71 and 90-91 of Ghosh.