QCB 408 / 508 – Notes on Week 6

Taishi Nakase

2020-03-23

Summary

In Week 6 of Foundations of Statistical Genomics, we shifted gears from frequentist inference to Bayesian inference. We introduced the fundamentals of Bayesian inference including estimation, classification, priors and empirical bayes which serve as Bayesian analogs to the frequentist techniques covered in previous weeks. Next, we proceeded with numerical methods for likelihood, where we covered latent variable models, the EM algorithm and Markov Chain Monte Carlo. Detailed proofs and additional examples are included to provide mathematical justification to important results and to illustrate possible applications of the techniques presented.

- Introduction to Bayesian Inference
- Bayesian Estimation
- Bayesian Classification
- Priors
- Empirical Bayes
- Latent Variable Models
- EM Algorithm
- Markov Chain Monte Carlo

Introduction to Bayesian Inference

Frequentist Inference vs. Bayesian Inference: Definitions

So far we have used a *frequentist* interpretation of probability. Frequentist inference only uses the likelihood, which is a conditional distribution of data given specific hypotheses. We assume that there exists some true hypothesis from which the observed data is being sampled and we measure the uncertainty of our conclusions in terms of theoretical repetitions of the sampling scheme used to generate the original observed data.

Bayesian inference assigns subjective probabilities to hypotheses, while acknowledging that there may exist some true hypothesis. As such, Bayesian inference involves a posterior probability, which is the likelihood times the prior subjective probabilities. Given that there is no single method for assigning such probabilities, it is inherently subjective. This subjectivity means that different people with different subjective beliefs may produce different conclusions.

Frequentist Inference vs. Bayesian Inference: Intuition

Suppose we have a simple random sample of n data points collected such that the following model of the data is reasonable: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Normal(\mu, \sigma^2)$. We assume that σ^2 is available and our goal is to do inference on μ.

Frequentist Approach
"I know μ is fixed, so I cannot assign probabilities to μ. I will use sample data and the likelihood function to come up with an estimate of μ, such as the MLE."

Bayesian Approach
"μ may be fixed, but I can still use probabilities to represent my subjective uncertainty. I will assign a probability distribution over μ (prior) and use the collected data to refine my initial beliefs (posterior) and approach the *truth*.

citation: probabilisticworld.com (this has a great document comparing frequentist and bayesian approaches)

Important Definitions

Prior probability distribution is the probability distribution of the unknown parameter that reflects one's subjective beliefs about its possible values. Note that the chosen prior distribution can influence the conclusions drawn from the sample data.

Posterior probability distribution is the probability distribution of the unknown parameter given the observed data. It represents our uncertainty in the parameter after combining the observed data (the likelihood) with what we assumed before we collected the data (the prior). We use this posterior distribution to obtain Bayesian analogs of frequentist confidence intervals and hypothesis tests.

General Framework

Suppose we model $(X_1, X_2, ..., X_n)|\theta \overset{iid}{\sim} F_\theta$ with prior distribution $\theta \sim F_\tau$.
The goal is to determine the posterior distribution of $\theta|\mathbf{X}$ through Bayes theorem:

$$f(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta) \cdot f(\theta)}{f(\mathbf{X})} = \frac{f(\mathbf{X}|\theta) \cdot f(\theta)}{\int f(\mathbf{X}|\theta^*) f(\theta^*) d\theta^*}$$

- $f(\theta)$ denotes the prior probability.
- $f(\theta|\mathbf{X})$ denotes the posterior probability.

Since $f(\mathbf{x})$ is not a function of θ it is often easier to use $f(\theta|\mathbf{x}) \propto L(\theta; \mathbf{x}) f(\theta)$ in computations.

Note that if there is indeed a "true" θ, then if certain regularity conditions are met $f(\theta|\mathbf{X})$ concentrates around the true θ as $n \to \infty$.

Example 1: Bayesian Inference for the Binomial Distribution

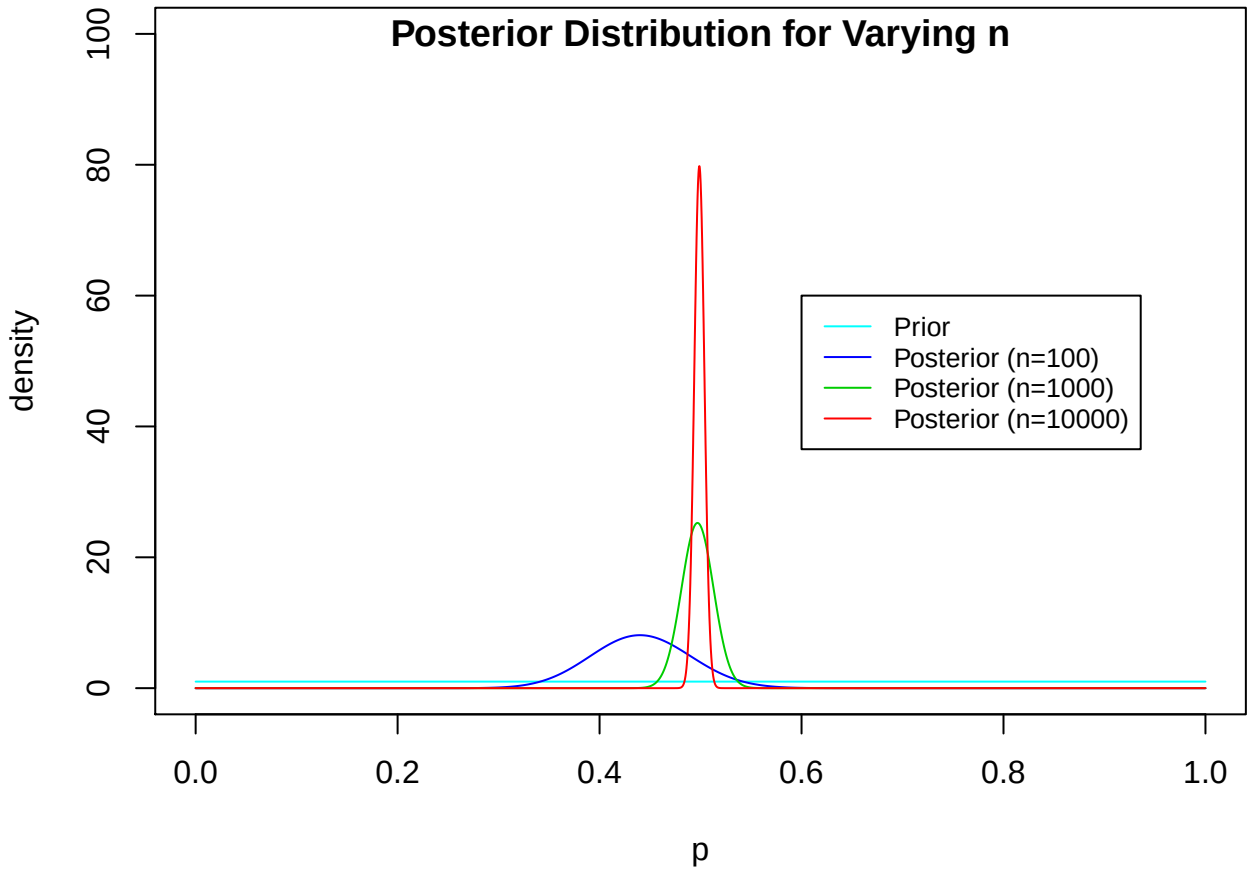
Prior Distribution: $P \sim Uniform(0, 1)$.
Data generating distribution: $X|P = p \sim Binomial(n, p)$.
Posterior pdf:

$$\begin{aligned}
f(p|X = x) &= \frac{Pr(X = x|P = p) \cdot f(p)}{Pr(X = x)} \\
&= \frac{Pr(X = x|P = p) \cdot f(p)}{\int_0^1 Pr(X = x|P = p*) f(p^*) dp^*} \\
&= \frac{\binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \cdot 1}{\int_0^1 \binom{n}{x} \cdot (p^*)^x \cdot (1-p^*)^{n-x} \cdot 1 dp^*} \\
&= \frac{p^x \cdot (1-p)^{n-x}}{\int_0^1 (p^*)^x \cdot (1-p^*)^{n-x} dp^*} \\
&= \frac{p^{(x+1)-1} \cdot (1-p)^{(n-x+1)-1}}{\int_0^1 (p^*)^{(x+1)-1} \cdot (1-p^*)^{(n-x+1)-1} dp^*} \\
&= \frac{p^{\alpha-1} \cdot (1-p)^{\beta-1}}{\int_0^1 (p^*)^{\alpha-1} \cdot (1-p^*)^{\beta-1} dp^*}
\end{aligned}$$

Hence, we have that $f(p|X=x) = \frac{p^{\alpha-1}\cdot(1-p)^{\beta-1}}{\int_0^1 (p^*)^{\alpha-1}\cdot(1-p^*)^{\beta-1}dp^*}$, where $\alpha = x+1$ and $\beta = n-x+1$.
As such, the posterior distribution is given as follows: $P|X=x \sim \text{Beta}(1+\text{x}, 1+\text{n-x})$.

Simulation
We claimed above that if there exists a "true" parameter p, then $f(p|X)$ concentrates around the true p as $n \to \infty$. We will verify this claim through simulation.
Suppose that there is a true p given by $p = 0.5 \implies$ the data generating distribution is $X \sim \text{Binomial}(\text{n}, 0.5)$. We will simulate the posterior distribution, $P|X=x \sim \text{Beta}(1+\text{x}, 1+\text{n-x})$, for increasing n to see whether $f(p|X)$ starts to concentrate around $p = 0.5$.

```
> set.seed(10)
> p = seq(0, 1, length=10000)
> plot(p, dbeta(p, 1, 1), xlab = "p", ylab = "density",
+      type="l", col=5, ylim = c(0, 100))
> title(main = "Posterior Distribution for Varying n", line=-1)
> n = c(100, 1000, 10000)
> x1 = rbinom(1, n[1], 0.5)
> lines(p, dbeta(p, x1+1,n[1]-x1+1), col=4)
> x2 = rbinom(1, n[2], 0.5)
> lines(p, dbeta(p, x2+1,n[2]-x2+1), col=3)
> x3 = rbinom(1, n[3], 0.5)
> lines(p, dbeta(p, x3+1,n[3]-x3+1), col=2)
> legend(0.6, 60,
+        c("Prior", "Posterior (n=100)","Posterior (n=1000)","Posterior (n=10000)"),
+        lty=c(1,1,1,1), col=c(5,4,3,2), cex = 0.8)
```



As expected, when $p = 0.5$ we see that $f(p|X)$ concentrates around the true p as $n \to \infty$.

Example 2: Bayesian Inference for the Normal Distribution

Consider n independent observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ such that $X_i|\mu \sim N(\mu, \sigma^2)$, where σ^2 is known.
Prior: $\mu \sim N(\mu_0, \sigma_0^2)$, where $f(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}$.
Posterior pdf:

$$\begin{aligned} f(\mu|\mathbf{x}) &= \frac{f(\mathbf{x}|\mu) \cdot f(\mu)}{f(\mathbf{x})} \\ &= \frac{\prod_{i=1}^n f(x_i|\mu) \cdot f(\mu)}{\int f(\mathbf{x}|\mu^*) \cdot f(\mu^*) d\mu^*} \\ &= \frac{\prod_{i=1}^n f(x_i|\mu) \cdot f(\mu)}{\int \prod_{i=1}^n f(x_i|\mu^*) \cdot f(\mu^*) d\mu^*} \end{aligned}$$

In this example it is significantly more difficult to calculate the posterior distribution.
The challenge typically lies in calculating marginal distribution $f(\mathbf{X})$ in Bayes theorem: $f(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta) \cdot f(\theta)}{f(\mathbf{X})}$.
Often it is actually impossible to calculate the posterior distribution, so Bayesian inference often requires intensive calculations to obtain numerical approximations of the posterior probability distribution.

Bayesian Estimation

Posterior Expectation

The posterior expected value is a common point estimate of θ.

$$\begin{aligned} E[\theta|\mathbf{x}] &= \int \theta f(\theta|\mathbf{x}) d\theta \\ &= \frac{\int \theta L(\theta; \mathbf{x}) f(\theta) d\theta}{\int L(\theta; \mathbf{x}) f(\theta) d\theta} \end{aligned}$$

Example: Consider the data generating distribution $X|\theta \sim \text{Binomial}(n, \theta)$ with prior $\theta \sim \text{Uniform}[0, 1]$.
We showed above that $\theta|X = x \sim \text{Beta}(1 + \text{x}, 1 + \text{n-x})$. (posterior distibution)
As such, the posterior expected value of θ is $E[\theta|x] = \frac{1+x}{1+x+1+n-x} = \frac{1+x}{2+n} \implies \hat{\theta} = \frac{1+x}{2+n}$.

Posterior Interval (Credible Interval)

The Bayesian analog of the frequentist confidence interval is the $1 - \alpha$ posterior interval.
We specify the level of confidence $1 - \alpha$ we want to achieve and find an interval (C_l, C_u) that achieves that level of confidence.

$$1 - \alpha = Pr\{C_l \leq \theta \leq C_u|\mathbf{x}\}$$

Note that the interval need not be symmetric.

Maximum *A Posterior* Probability

The maximum *a posterior* probability (MAP) is the Bayesian analog of the frequentist MLE.
Formally, the MAP is the value of θ that maximizes the posterior pdf or pmf.

$$\begin{aligned} \hat{\theta}_{MAP} &= \text{argmax}_\theta \, f(\theta|x) \\ &= \text{argmax}_\theta \, L(\theta; \mathbf{x}) f(\theta) \end{aligned}$$

It is worth noting that $L(\theta; \mathbf{x}) f(\theta)$ represents a weighted likelihood, where greater weight is placed on θ values with higher prior probabilities. Assuming the existent of a fixed θ, some may wonder whether a poorly

selected prior probability distribution can aversely affect the quality of the MAP estimator. For small data sets, this is certainly a possibility and this emphasizes the need to think carefully about the selected prior and to try different priors to see how sensitive the results are to the choice of prior. With enough data, however, $\hat{\theta}_{MAP}$ will converge to the fixed θ because $L(\theta; \mathbf{x}) \to \infty$ causing the $f(\theta)$ priors to become irrelevant.

Loss Functions

A loss function, $L(\theta, \hat{\theta})$ is a function of the true parameter θ and an estimate of the parameter $\hat{\theta}$.
Loss functions measure how *bad* the current estimate is, so the larger the loss the worse the estimate is according to the selected loss function.
There are many examples of loss functions.

- Squared-error loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.
- Absolute loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$.
- Asymmetric squared-error loss:

$$L(\theta, \hat{\theta}) = \begin{cases} (\theta - \hat{\theta})^2 & \hat{\theta} < \theta \\ c(\theta - \hat{\theta})^2 & \hat{\theta} \geq \theta, 0 < c < 1 \end{cases}$$

  The asymmetric squared-error loss function penalizes underestimates of the true estimate more than overestimates of the true estimate.

Note that the true parameter θ is unknown, so we cannot compute the true loss associated with using $\hat{\theta}$ as an estimate. In the Bayesian framework, however, we have the posterior distribution which characterizes the distribution of the unknown parameter we are trying to estimate given the collected data. As such, we are able to compute the *expected loss* given an estimate with respect to the posterior distribution; this expected loss is known as *Bayes risk*.

Bayes Risk

The Bayes risk, $R(\theta, \hat{\theta})$, is the expected loss with respect to the posterior:

$$E[L(\theta, \hat{\theta})|\mathbf{x}] = \int L(\theta, \hat{\theta}) f(\theta|\mathbf{x}) d\theta$$

Bayes Estimators

The *Bayes estimator* minimizes the Bayes risk.

1. The posterior expectation $E[\theta|\mathbf{x}]$ minimizes the Bayes risk of $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.
   Proof:
   The goal is to minimize $E[(\theta - \hat{\theta})^2|\mathbf{X} = \mathbf{x}]$.

$$\begin{aligned} E[(\theta - \hat{\theta})^2|\mathbf{X} = \mathbf{x}] &= E[(\theta - E(\theta|\mathbf{x}) + E(\theta|\mathbf{x}) - \hat{\theta})^2|\mathbf{X} = \mathbf{x}] \\ &= E[(\theta - E(\theta|\mathbf{x}))^2|\mathbf{X} = \mathbf{x}] + E[(E(\theta|\mathbf{x}) - \hat{\theta})^2|\mathbf{X} = \mathbf{x}] - 2(E(\theta|\mathbf{x}) - \hat{\theta})E[(\theta - E[\theta|\mathbf{x}])|\mathbf{X} = \mathbf{x}] \\ &= E[(\theta - E(\theta|\mathbf{x}))^2|\mathbf{X} = \mathbf{x}] + E[(E(\theta|\mathbf{x}) - \hat{\theta})^2|\mathbf{X} = \mathbf{x}] - 2(E(\theta|\mathbf{x}) - \hat{\theta})(E[\theta|\mathbf{x}] - E[\theta|\mathbf{x}]) \\ &= E[(\theta - E(\theta|\mathbf{x}))^2|\mathbf{X} = \mathbf{x}] + E[(E(\theta|\mathbf{x}) - \hat{\theta})^2|\mathbf{X} = \mathbf{x}] \\ &= E[(\theta - E(\theta|\mathbf{x}))^2|\mathbf{X} = \mathbf{x}] + (E(\theta|\mathbf{x}) - \hat{\theta})^2 \end{aligned}$$

   We see that the Bayes risk under the squared error loss function, $E[(\theta - \hat{\theta})^2|\mathbf{X} = \mathbf{x}]$, is minimized when $(E(\theta|\mathbf{x}) - \hat{\theta})^2$ is as small as possible. As such, $\hat{\theta} = E(\theta|\mathbf{x})$ is the Bayes estimator.

2. The median of $f(\theta|\mathbf{x})$, calculated by $F^{-1}_{\theta|\mathbf{x}}(\frac{1}{2})$ minimizes the Bayes risk of $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$.

Example: Consider the data generating distribution $X|\theta \sim \text{Binomial}(n, \theta)$ with prior $\theta \sim \text{Beta}(\alpha, \beta)$.
Suppose we want the Bayes estimator with respect to the squared error loss.
As shown above, $E(\theta|\mathbf{x})$ is the Bayes estimator under the squared loss function.

$$
\begin{aligned}
E[\theta|x] &= \int_0^1 \theta f(\theta|x) d\theta \\
&= \frac{\int_0^1 \theta L(\theta; \mathbf{x}) f(\theta) d\theta}{\int_0^1 L(\theta^*; \mathbf{x}) f(\theta^*) d\theta^*} \\
&= \frac{\int_0^1 \theta \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}{\int_0^1 \binom{n}{x} (\theta^*)^x (1-\theta^*)^{n-x} \frac{1}{B(\alpha,\beta)} (\theta^*)^{\alpha-1}(1-\theta^*)^{\beta-1} d\theta^*} \quad \text{where } B(\alpha,\beta) \text{ is the beta function.} \\
&= \frac{\int_0^1 \theta \cdot \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}{\int_0^1 (\theta^*)^x (1-\theta^*)^{n-x} (\theta^*)^{\alpha-1}(1-\theta^*)^{\beta-1} d\theta^*} \\
&= \frac{\int_0^1 \theta \cdot \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} d\theta}{\int_0^1 (\theta^*)^{x+\alpha-1}(1-\theta^*)^{n-x+\beta-1} d\theta^*} \\
&= \int_0^1 \theta \cdot g(\theta) d\theta \quad \text{where } g(\theta) \text{ is the beta pdf with parameters } x+\alpha \text{ and } n-x+\beta \\
&= \frac{x+\alpha}{x+\alpha+n-x+\beta} \\
&= \frac{\alpha+x}{\alpha+\beta+n}
\end{aligned}
$$

Hence, the Bayes estimator of θ is $\frac{\alpha+x}{\alpha+\beta+n}$.

Bayesian Classification

Assumptions

Let $(X_1, X_2, ..., X_n)|\theta \overset{iid}{\sim} F_\theta$ where $\theta \in \Theta$ and $\theta \sim F_\tau$.
Let $\Theta_0, \Theta_1 \subseteq \Theta$ so that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$.
Θ_0 is the null hypothesis and Θ_1 is the alternative hypothesis.

Bayesian analog of hypothesis testing: Given observed data $\mathbf{x}$, we wish to clarify whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$.

Prior Probability on H

Let H be a random variable such that $H = 0$ when $\theta \in \Theta_0$ and $H = 1$ when $\theta \in \Theta_1$.
From the prior distribution on θ, we can calculate

$$
\begin{aligned}
Pr\{H = 0\} &= \int_{\theta \in \Theta_0} f(\theta) d\theta \\
Pr\{H = 1\} &= 1 - Pr\{H = 0\}
\end{aligned}
$$

The prior probability on H is a measure of our subjective belief that the null hypothesis is true.

Posterior Probability

Using Bayes Theorem, we have that

$$\begin{aligned} Pr\{H=0|\mathbf{x}\} &= \frac{f(\mathbf{x}|H=0)Pr\{H=0\}}{f(\mathbf{x})} \\ &= \frac{\int_{\theta\in\Theta_0} f(\mathbf{x}|\theta)f(\theta)d\theta}{\int_{\theta\in\Theta} f(\mathbf{x}|\theta)f(\theta)d\theta} \end{aligned}$$

Note that $Pr\{H=1|\mathbf{x}\} = 1 - Pr\{H=0|\mathbf{x}\}$.
We update our subjective belief that the null hypothesis or the alternative hypothesis is true after observing some data from a reasonable data generating distribution.

Loss Function

Let $L(\hat{H}, H)$ be such that:

$$\begin{aligned} L(\hat{H}=1, H=0) &= C_I \\ L(\hat{H}=0, H=1) &= C_{II} \end{aligned}$$

for some $C_I, C_{II} > 0$.
We can interpret C_I and C_{II} as the type I error penalty and type II error penalty respectively.

Bayes Risk

The Bayes risk, $R(\hat{H}, H)$, is

$$\begin{aligned} E[L(\theta,\hat{\theta})|\mathbf{x}] &= C_I Pr\{\hat{H}=1, H=0\} + C_{II} Pr\{\hat{H}=0, H=1\} \\ &= C_I Pr\{\hat{H}=1|H=0\}Pr\{H=0\} + C_{II} Pr\{\hat{H}=0|H=1\}Pr\{H=1\} \end{aligned}$$

where $Pr\{\hat{H}=1|H=0\}$ represents the type I error rate and $Pr\{\hat{H}=0|H=1\}$ represents the type II error rate. As such, we can adjust the type I (C_I) and type II (C_{II}) penalities to selectively control the "loss" we place on type I and type II errors respectively.

Example

Suppose we are trying to estimate the proportion of Princeton's student body θ infected with COVID-19.
Prior Distribution: $\theta \sim Beta(2,5)$.
Data generating distribution: $X|\theta \sim Binomial(n,\theta)$.
Suppose we are interesting in testing the following hypothesis:

- $\Theta_0 = (0, 0.4)$ and $\Theta_1 = (0.4, 1)$ where $\Theta = (0,1)$.

Prior Probability on H

$$\begin{aligned} Pr\{H=0\} &= \int_0^{0.4} f(\theta)d\theta = 0.76672 \\ Pr\{H=1\} &= 1 - Pr\{H=0\} = 0.23328 \end{aligned}$$

Posterior Probability on H

$$\begin{aligned} Pr\{H=0|x\} &= \frac{\int_{\theta\in\Theta_0} f(x|\theta)f(\theta)d\theta}{\int_{\theta\in\Theta} f(x|\theta)f(\theta)d\theta} \\ &= \frac{\int_0^{0.4} \theta^x(1-\theta)^{n-x}\theta^{2-1}(1-\theta)^{5-1}d\theta}{\int_0^1 \theta^x(1-\theta)^{n-x}\theta^{2-1}(1-\theta)^{5-1}d\theta} \\ &= \frac{\int_0^{0.4} \theta^{(x+2)-1}(1-\theta)^{(n-x+5)-1}d\theta}{\int_0^1 \theta^{(x+2)-1}(1-\theta)^{(n-x+5)-1}d\theta} \\ &= Pr\{B\leq 0.4\} \quad \text{where } B\sim Beta(x+2, n-x+5) \end{aligned}$$

Furthermore, $Pr\{H=1|x\} = 1 - Pr\{H=0|x\} = Pr\{B > 0.4\}$ where $B\sim Beta(x+2, n-x+5)$.

Bayes Rule
Given the serious nature of the COVID-19 pandemic, we would naturally prefer to have a false positive to a false negative; that is, we would rather conclude that the pandemic is worse than it actually is and have an overproportionate response than conclude that it is better than it actually is and have an insufficient response. As such we select the penalities C_{II} and C_I such that $C_{II} > C_I$ to ensure that we assign a larger penalty to false negatives.
For illustration, suppose $C_I = 1$ and $C_{II} = 2$.
By Bayes rule, we assign $\hat{H} = 1$ when $Pr\{H=1|x\} \geq \frac{1}{3}$.
Suppose we collect the following sample: $n = 20$ and $x = 8$.

$$Pr\{H=1|x\} = 1 - Pr\{B\leq 0.5\} = 0.3641828 \geq 0.3333333$$

Since $Pr\{H=1|x\} \geq \frac{1}{3}$, we have that $\hat{H} = 1$.
We see that when we impose a large penalty on false negatives, we conclude that the alternative hypothesis is true even for small $\frac{x}{n}$. Note that this is a simplified example used to demonstrate how one would approach Bayesian classification problems and as such I have purposely exaggerated certain parameter values.

Priors

So far we have neglected to address the reasoning behind our selection of prior distributions. Though we are permitted to select any prior distribution we seem fit, there are some prior distributions with useful properties that make our calculations of the posterior distribution easier.

Conjugate Priors

A *conjugate prior* is a prior distribution for a data generating distribution so that the posterior distribution is of the same type as the prior distribution. Conjugate priors provide a closed-form expression for the posterior without the need to resort to numerical integration. Moreover, conjugate priors clearly show how a likelihood function updates a prior distribution.

Note that all exponential family distributions have conjugate priors.

Example: Beta-Bernoulli

Suppose $\mathbf{X}|\theta \overset{iid}{\sim} Bernoulli(p)$ and suppose that $\theta \sim Beta(\alpha, \beta)$.
Recall

- $f(x_i;\theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$ for $x_i \in \{0,1\}$.
- $f(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$ where B is a normalization constant and $\theta \in (0,1)$.

Claim: $\theta|\mathbf{X}=\mathbf{x} \sim Beta(\alpha+\sum_{i=1}^n x_i, \beta+\sum_{i=1}^n(1-x_i))$.
Proof:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto L(\theta;\mathbf{x})f(\theta) \\ &\propto \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+\sum x_i-1}(1-\theta)^{\beta+\sum(1-x_i)-1} \\ &\propto Beta(\alpha+\sum x_i, \beta+\sum(1-x_i)) \end{aligned}$$

Hence, $\theta|\mathbf{X}=\mathbf{x} \sim \text{Beta}(\alpha+\sum x_i, \beta+\sum(1-x_i))$.
We see that the prior has a beta distribution with parameters α and β and the posterior has a beta distribution with parameters $\alpha+\sum x_i$ and $\beta+\sum(1-x_i)$.

We can easily obtain the point estimate $E[\theta|\mathbf{x}] = \frac{\alpha+\sum x_i}{\alpha+\beta+n}$.
It is worth noting that $n \to \infty$, $E[\theta|\mathbf{x}] \to \overline{x} = \hat{\theta}_{MLE}$.
That is, as we collect more data our Bayesian estimator approaches the frequentist estimator.

Example: Normal-Normal

Suppose $\mathbf{X}|\mu \overset{iid}{\sim} Normal(\mu, \sigma^2)$ where σ^2 is known and suppose that $\mu \sim Normal(a, b^2)$
Claim: $\mu|\mathbf{X}=\mathbf{x} \sim Normal(\frac{b^2}{\frac{\sigma^2}{n}+b^2}\overline{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}a, \frac{b^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2})$
Proof:

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto L(\mu;\mathbf{x})f(\mu) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_i-\mu)^2}{2\sigma^2}\} \cdot \frac{1}{\sqrt{2\pi b^2}}\exp\{-\frac{(\mu-a)^2}{2b^2}\} \\ &\propto \exp\{-\frac{(\mu-a)^2}{2b^2} - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\} \\ &= \exp\{-\frac{(\mu-a)^2\sigma^2 + b^2\sum_{i=1}^n(x_i-\mu)^2}{2b^2\sigma^2}\} \\ &= \exp\{\frac{-\mu^2(\sigma^2+nb^2) + 2\mu(a\sigma^2+b^2\sum_{i=1}^n x_i) - (a^2\sigma^2+b^2\sum_{i=1}^n x_i^2)}{2b^2\sigma^2}\} \\ &= \exp\{\frac{-\mu^2 + 2\mu\frac{(a\sigma^2+b^2\sum_{i=1}^n x_i)}{(\sigma^2+nb^2)} - \frac{(a^2\sigma^2+b^2\sum_{i=1}^n x_i^2)}{(\sigma^2+nb^2)}}{\frac{2b^2\sigma^2}{(\sigma^2+nb^2)}}\} \\ &\propto \exp\{\frac{-\mu^2 + 2\mu\frac{(a\sigma^2+b^2\sum_{i=1}^n x_i)}{(\sigma^2+nb^2)} - (\frac{a\sigma^2+b^2\sum_{i=1}^n x_i}{\sigma^2+nb^2})^2}{\frac{2b^2\sigma^2}{(\sigma^2+nb^2)}}\} \cdot \exp\{-\frac{a^2\sigma^2+b^2\sum_{i=1}^n x_i^2}{2b^2\sigma^2}\} \\ &\propto \exp\{-\frac{(\mu - \frac{a\sigma^2+b^2\sum_{i=1}^n x_i}{\sigma^2+nb^2})^2}{\frac{2b^2\sigma^2}{(\sigma^2+nb^2)}}\} \\ &\propto \text{Normal}(\frac{a\sigma^2+b^2\sum_{i=1}^n x_i}{\sigma^2+nb^2}, \frac{b^2\sigma^2}{\sigma^2+nb^2}) \end{aligned}$$

Hence, we have that$\mu|\mathbf{X}=\mathbf{x} \sim \text{Normal}(\frac{a\sigma^2+b^2\sum_{i=1}^n x_i}{\sigma^2+nb^2}, \frac{b^2\sigma^2}{\sigma^2+nb^2}) = Normal(\frac{b^2}{\frac{\sigma^2}{n}+b^2}\overline{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}a, \frac{b^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2})$

We can also obtain the following properties of $\mu|\mathbf{X} = \mathbf{x}$.

- $E[\mu|\mathbf{x}] = \frac{b^2}{\frac{\sigma^2}{n}+b^2}\overline{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}a.$
- $Var(\mu|\mathbf{x}) = \frac{b^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+b^2}.$

Example: Beta-Geometric

Suppose $\mathbf{X}|\theta \overset{iid}{\sim}$ Geometric(θ) with $\theta \sim$ Beta(α, β).
Recall:

- $f(x_i;\theta) = \theta(1-\theta)^{x_i}$ for $x_i \in \{0, 1, ..., \}$.
- $f(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$ where B is a normalization constant and $\theta \in (0, 1)$.

Claim: $\theta|\mathbf{X} = \mathbf{x} \sim Beta(\alpha + n, \beta + \sum_{i=1}^n x_i)$.
Proof:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto L(\theta;\mathbf{x})f(\theta) \\ &\propto \prod_{i=1}^{n}(1-\theta)^{x_i}\theta \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+n-1}(1-\theta)^{\beta+\sum_{i=1}^n x_i-1} \\ &\propto Beta(\alpha+n, \beta+\sum_{i=1}^{n} x_i) \end{aligned}$$

Hence, $\theta|\mathbf{X} = \mathbf{x} \sim$ Beta$(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

Jeffreys Prior

Jeffreys prior is proportional to the square root of the Fisher information:

$$f(\theta) \propto \sqrt{I(\theta)}$$

It has the key property that it is invariant to transformations of θ.

It is worth noting that Jeffreys prior is an uninformative prior in the sense that it gives minimal information about the parameter, θ. As such, the Jeffreys posterior reflects the information about θ that is gathered solely from the data. It is typically used when a suitable prior distribution is not available or if we don't know what θ should be and we want the "data to speak for itself".

Example: Jeffreys Prior for the Bernoulli

Suppose $X \sim Ber(\theta)$ where $f(x;\theta) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0, 1\}$.
The log-likelihood is

$$log(f(X;\theta)) = Xlog(\theta) + (1-X)log(1-\theta)$$

The score function is

$$s(\theta) = \frac{d}{d\theta}log(f(X;\theta)) = \frac{X}{\theta} - \frac{1-X}{1-\theta}$$

Hence, the Fisher information is given by

$$\begin{aligned} I(\theta) &= -E[\frac{d}{d\theta}s(\theta)] \\ &= -E[\frac{-X}{\theta^2} - \frac{1-X}{(1-\theta)^2}] \\ &= \frac{E(X)}{\theta^2} + \frac{1-E(X)}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Hence, Jeffreys prior is $f(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \propto \text{Beta}(\frac{1}{2}, \frac{1}{2})$.

Other Jeffreys prior examples

- Normal(μ, σ^2), σ^2 is known: $f(\mu) \propto 1$.
- Normal(μ, σ^2), μ is known: $f(\sigma) \propto \frac{1}{\sigma}$.
- Poisson(λ): $f(\lambda) \propto \frac{1}{\sqrt{\lambda}}$.
- Bernoulli(p): $f(p) \propto \frac{1}{\sqrt{p(1-p)}}$.

Improper Prior

An improper prior is a prior such that $\int f(\theta)d\theta = \infty$.
Sometimes it may still be the case that $f(\theta|\mathbf{x}) \propto L(\theta; \mathbf{x})f(\theta)$ yields a probability distribution.

Consider $\mathbf{X}|\mu \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ where σ^2 is known and $f(\mu) \propto 1$.
Note that $\int f(\mu)d\mu = \infty$.

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto L(\mu; \mathbf{x})f(\mu) \\ &\propto \exp\{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\} \\ &\propto \exp\{-\frac{(n\mu^2 - 2\mu\sum_{i=1}^{n} x_i)}{2\sigma^2}\} \\ &= \exp\{-\frac{(\mu - 2\mu\overline{x})}{\frac{2\sigma^2}{n}}\} \\ &\propto \exp\{-\frac{(\mu - \overline{x})^2}{\frac{2\sigma^2}{n}}\} \\ &\propto \text{Normal}(\overline{x}, \frac{\sigma^2}{n}) \end{aligned}$$

Hence, even though $\int f(\mu)d\mu = \infty$, we have that $\mu|\mathbf{X} = \mathbf{x} \sim \text{Normal}(\overline{x}, \frac{\sigma^2}{n})$.

Empirical Bayes

Introduction

Consider $\mathbf{X}|\theta \overset{iid}{\sim} F_\theta$ with prior probability $\theta \sim F_\tau$.
We want to determine values for τ using the complete set of empirical measurements.
Empirical Bayes approach uses the observed data to estimate the prior parameter(s) τ.
Note that this is useful for high dimensional data when parameters are simultaneously drawn from a prior with multiple observations drawn per parameter realisation.

Approach

1. Integrate out the parameter to obtain: $f(\mathbf{x};\tau) = \int f(\mathbf{x}|\theta)f(\theta;\mathbf{x})d\theta$
2. An estimation method such as MLE is then applied to estimate τ.
3. The Bayesian inference proceeds as usual under the assumption that $\theta \sim f(\theta;\hat{\tau})$.

Example

Suppose that $X_i|\mu_i \sim \text{Normal}(\mu_i, 1)$ for $i = 1, 2, ...n$ where these rv's are independent.
Also, suppose that $\mu_i \overset{iid}{\sim} \text{Normal}(a, b^2)$.

$$\begin{aligned}
f(x_i;a,b) &= \int f(x_i|\mu_i)f(\mu_i;a,b)d\mu_i \sim \text{Normal}(a, 1+b^2) \\
&\implies \hat{a} = \overline{x}, 1+\hat{b}^2 = \frac{\sum_{k=1}^n (x_k - \overline{x})}{n} \quad \text{(MLE)} \\
E[\mu_i|x_i] &= \frac{1}{1+b^2}a + \frac{b^2}{1+b^2}x_i \implies \\
E[\hat{\mu_i}|x_i] &= \frac{1}{1+\hat{b}^2}\hat{a}^2 + \frac{\hat{b}^2}{1+\hat{b}^2}x_i \\
&= \frac{n}{\sum_{k=1}^n (x_k - \overline{x})^2}\overline{x} + (1 - \frac{n}{\sum_{k=1}^n (x_k - \overline{x})^2})x_i
\end{aligned}$$

Introduction to Numerical Methods for Likelihood

Motivation

- Frequentist model: $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$.
- Bayesian model: $X_1, \ldots, X_n|\theta \overset{iid}{\sim} F_\theta$ and $\theta \sim F_\tau$.

Ideally we would like to work directly with likelihood (frequentist) or the posterior probability (Bayesian) to analytically derive closed form estimates for $\hat{\theta}_{MLE}$, $\hat{\theta}_{MAP}$, $E[\theta|\mathbf{x}]$, or $f(\theta|\mathbf{x})$. Sometimes it is not possible to find formula for such estimates, and we have to resort to numerical methods.

Approaches

We will consider the following numerical approaches to likelihood based inferences:

- Expectation-maximization (EM) algorithm
- Variational inference
- Markov Chain Monte Carlo (MCMC)
  - Metropolis Sampling
  - Metropolis-Hastings Sampling
  - Gibbs Sampling

Latent Variable Models

Latent variables are random variables that are present in the probabilistic model, but are unobserved.
The probabilistic model involves pairs of observed and hidden random variables,

$$(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n) \overset{iid}{\sim} F_\theta$$

where $\mathbf{Z}$ are the latent variables and $\mathbf{X}$ are the observed variables.

The EM algorithm and variational inference are set up using latent variables.

Empirical Bayes Revisited

Note that *Bayesian models* are special cases of latent variable models, where the unobserved random parameters simulated by the prior distribution are the latent variables.
Suppose $X_i|\mu_i \sim \text{Normal}(\mu_i, 1)$ for $i = 1, \ldots, n$ where the rvs are independent with $\mu_i \overset{iid}{\sim} \text{Normal}(\text{a}, \text{b}^2)$. The unobserved parametes $\mu_1, \mu_2, \ldots, \mu_n$ are latent variables with $\boldsymbol{\theta} = (a, b^2)$.

Normal Mixture Model

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$ where
$\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$ with pdf

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x_1 - \mu_k)^2}{2\sigma_k^2}\}$$

π_k denotes the probability of drawing from the k^{th} normal distribution with parameters μ_k and σ_k^2.
Informally, you have K underlying normal distributions and you choose one of those K with probability π_k and then you draw your observed random variable from the normal distribution you randomly chose.
We must resort to numerical methods because there are no closed form solutions for the MLEs.

There is a *latent variable model* that produces the same marginal distribution and likelihood function.
Let $\mathbf{Z_1}, \ldots, \mathbf{Z_n} \overset{iid}{\sim} \text{Multinomial}_K(1, \boldsymbol{\pi})$ where $\pi = (\pi_1, \ldots, \pi_K)$.
Since $\mathbf{Z}_i$ is a multinomial distribution with parameters 1 and π we have that $Z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} Z_{ik} = 1$.
Let $[X_i|Z_{ik} = 1] \sim \text{Normal}(\mu_k, \sigma_k^2)$ where $\{X_i|\mathbf{Z_i}\}_{i=1}^{n}$ are jointly independent.
The joint pdf is

$$f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\}]^{z_{ik}}$$

If we work with this pdf, we can derive analytical solutions to all the maximum likelihood estimates.

It can be verified that $f(\mathbf{x}; \boldsymbol{\theta})$ is the marginal distribution of this latent variable model.
Proof:

$$\begin{aligned}
\prod_{i=1}^{n} \sum_{\mathbf{z_i}} f(x_i, \mathbf{z_i}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \sum_{\mathbf{z_i}} \prod_{k=1}^{K} [\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\}]^{z_{ik}} \\
&= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\} \quad (1) \\
&= \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\
&= f(\mathbf{x}; \theta)
\end{aligned}$$

(1) follows from the fact that each possible $\mathbf{z_i}$ has the property that $z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} z_{ik} = 1$.

Bernoulli Mixture Model

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, p_1, \ldots, p_k)$.
Note that we now have K underlying Bernoulli distributions.
The pmf is given by

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k p_k^{x_i} (1 - p_k)^{1 - x_i}$$

There is a also a latent variable model that produces the same marginal distribution and likelihood function.

Let $\mathbf{Z_1}, \ldots, \mathbf{Z_n} \overset{iid}{\sim} \text{Multinomial}_K(1, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$.
Since $\mathbf{Z_i}$ is a multinomial distribution with parameters 1 and π we have that $Z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} Z_{ik} = 1$.
Let $[X_i | Z_{ik} = 1] \sim \text{Bernoulli}(p_k)$ where $\{X_i | \mathbf{Z_i}\}_{i=1}^{n}$ are jointly independent.
The joint pmf is

$$f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k p_k^{x_i} (1 - p_k)^{1 - x_i})^{z_{ik}}$$

EM Algorithm

Introduction

In the EM algorithm, we introduce a latent variable model to make the likelihood more tractable. This allows us to derive analytical formulas for the MLE and the MAP, which would be otherwise impossible if we were to use the initial intractable marginal distribution. This process effectively treats the latent variables as missing data; in fact, when we actually have missing data, we can apply the EM algorithm.

Requirements

The EM algorithm allows us to calculate MLEs and MAPs when certain geometric properties are satisfied in the probabilisitc mode.

In order for the EM algorithm to be a practical approach, we should have a latent variable model $f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ that marginally gives the original pdf/pmf.

Note that some $(\mathbf{x}, \mathbf{z})$ is called *complete data* and $\mathbf{x}$ is called the *observed data* when we are using the EM as a method for dealing with missing data.

The Algorithm

1. Set up latent variable model so that the complete data likelihood is tractable.

2. Choose initial value $\boldsymbol{\theta}^{(0)}$.
3. Calculate $f(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})$.
4. Calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)}]$
5. Set $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$
6. Iterate until convergence and set $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(\boldsymbol{\infty})}$.

$\boldsymbol{\theta}^{(\mathbf{0})}$	Initial guess for the parameters.
$f(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})$	Conditional distribution of the latent variables given the observed data $\mathbf{x}$ and the current parameter estimate $\boldsymbol{\theta}^{(\boldsymbol{t})}$.
$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$	Function of both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(t)}$, where we integrate out the latent variable in terms of the conditional distribution $\mathbf{Z}|\mathbf{X} = \mathbf{x}$ under the current parameter guess, $\boldsymbol{\theta}^{(t)}$. As such, we handle the unknown $\mathbf{Z}$ by replacing it with $E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[\cdot]$.

Continuous $\mathbf{Z}$:

- $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \int log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)}) d\mathbf{z}$.

Discrete $\mathbf{Z}$:

- $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{\mathbf{z}} log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$.

EM for MAP

In the frequentist framework, we want to maximize the likelihood. In the Bayesian framework, however, we want to maximize the posterior probaility. As such, we must replace the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ shown above with

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[logf(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}); \boldsymbol{\theta}^{(t)}] + logf(\boldsymbol{\theta})$$

where $logf(\boldsymbol{\theta})$ is the prior distribution on $\boldsymbol{\theta}$.

EM Example: Normal Mixture Model

Step 1: Set up latent variable model

$$\log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) + z_{ik} \log(\phi(x_i; \mu_k, \sigma_k^2))$$

where

$$\phi(x_i; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\}$$

Step 2: Calculate $f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$.

$$\begin{aligned} f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)}) &= \prod_{i=1}^{n} f(\mathbf{z_i}|x_i; \boldsymbol{\theta}^{(t)}) \\ &= \prod_{i=1}^{n} \frac{f(\mathbf{z_i}, x_i; \boldsymbol{\theta}^{(t)})}{f(x_i; \boldsymbol{\theta^{(t)}})} \\ &= \prod_{i=1}^{n} \frac{\prod_{k=1}^{K} [\pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, \sigma_k^{2,(t)})]^{z_{ik}}}{\sum_{k=1}^{K} \pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, \sigma_k^{2,(t)})} \end{aligned}$$

Step 3: Calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
Since $\log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ is linear in z_{ik}, we only to need to know $E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x}; \boldsymbol{\theta}^{(t)}]$ to calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

$$\begin{aligned} E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x}; \boldsymbol{\theta}^{(t)}] &= \sum_{\mathbf{z}} z_{ik} \cdot f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)}) \\ &= \sum_{\mathbf{z}} z_{ik} \cdot \prod_{i=1}^{n} \frac{f(\mathbf{z_i}, x_i; \boldsymbol{\theta}^{(t)})}{f(x_i; \boldsymbol{\theta}^{(t)})} \\ &= \sum_{\mathbf{z}_i} z_{ik} \cdot \frac{f(\mathbf{z_i}, x_i; \boldsymbol{\theta}^{(t)})}{f(x_i; \boldsymbol{\theta}^{(t)})} \\ &= \frac{1}{f(x_i; \boldsymbol{\theta}^{(t)})} \sum_{\mathbf{z}_i} z_{ik} f(\mathbf{z_i}, x_i; \boldsymbol{\theta}^{(t)}) \\ &= \frac{1}{f(x_i; \boldsymbol{\theta}^{(t)})} \sum_{\mathbf{z}_i} z_{ik} \prod_{k=1}^{K} [\pi_k^{(t)} \frac{1}{\sqrt{2\pi\sigma_k^{2,(t)}}} \exp\{-\frac{(x_i - \mu_k^{(t)})^2}{2\sigma_k^{2,(t)}}\}]^{z_{ik}} \\ &= \frac{1}{f(x_i; \boldsymbol{\theta}^{(t)})} [\pi_k^{(t)} \frac{1}{\sqrt{2\pi\sigma_k^{2,(t)}}} \exp\{-\frac{(x_i - \mu_k^{(t)})^2}{2\sigma_k^{2,(t)}}\}] \\ &= \frac{\pi_k^{(t)} \phi(x_i; \mu_k^{(t)}, \sigma_k^{2,(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi(x_i; \mu_j^{(t)}, \sigma_j^{2,(t)})} \end{aligned}$$

Define $\hat{z}_{ik}^{(t)} = E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x};\boldsymbol{\theta}^{(t)}] = \frac{\pi_k^{(t)}\phi(x_i;\mu_k^{(t)},\sigma_k^{2,(t)})}{\sum_{j=1}^{K}\pi_j^{(t)}\phi(x_i;\mu_j^{(t)},\sigma_j^{2,(t)})}$.

Therefore,

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}^{(t)}log(\pi_k) + \hat{z}_{ik}^{(t)}log(\phi(x_i;\mu_k,\sigma_k^2))$$

Step 4: Calculate $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)})$.

- $\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}$.
- $\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}x_i}{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}}$.
- $\sigma_k^{2,(t+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}(x_i-\mu_k^{(t+1)})^2}{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}}$.

Derivation 1: $\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}$.
Consider the following lagrangian: $L = Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) - \lambda(\sum_{k=1}^{K}\pi_k - 1)$.
We have the following KKT optimality conditions:

1. $\sum_{i=1}^{K}\hat{z}_{ik}^{(t)}\cdot\frac{1}{\pi_k} - \lambda = 0$ for all $k = 1,\ldots,K$.
2. $\sum_{k=1}^{K}\pi_k = 1$ and $\lambda > 0$. (We enforce the constraint to be binding to ensure the probabilities add to 1).

Note that we can rearrange condition 1 as follows: $\frac{1}{\lambda}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} = \pi_k$.
If we sum condition 1 for all $k \in \{1,\ldots,K\}$, we have that $\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} = \sum_{k=1}^{K}\pi_k = 1$.
$\implies \lambda = \sum_{k=1}^{K}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} = \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}^{(t)} = \sum_{i=1}^{n}1 = n$.
Hence, $\pi_k^{(t+1)} = \frac{1}{\lambda}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}$.

Derivation 2: $\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}x_i}{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}}$.

$$\begin{aligned}\frac{\partial}{\partial\mu_k}Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}\frac{\partial}{\partial\mu_k}log(\phi(x_i;\mu_k,\sigma_k^2)) \\ &= \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}\frac{\partial}{\partial\mu_k}[log(\frac{1}{\sqrt{2\pi\sigma_k^2}}) - \frac{(x_i-\mu_k)^2}{2\sigma_k^2}] \\ &= \sum_{i=1}^{n}\hat{z}_{ik}^{(t)}\frac{(x_i-\mu_k)}{\sigma_k^2}\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{\partial}{\partial\mu_k}Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) = 0 \\ \implies &\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}x_i - \mu_k\sum_{i=1}^{n}\hat{z}_{ik}^{(t)} = 0 \\ \implies &\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}x_i}{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}}.\end{aligned}$$

The derivation for $\sigma_k^{2,(t+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}(x_i-\mu_k^{(t+1)})^2}{\sum_{i=1}^{n}\hat{z}_{ik}^{(t)}}$ is very similar to the one provided above.

EM Example: Bernoulli Mixture Model

Step 1: Set up latent variable model

$$\log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) + z_{ik} \log(p_k^{x_i}(1-p_k)^{1-x_i})$$

Step 2: Calculate $f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$.

$$\begin{aligned} f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(t)}) &= \prod_{i=1}^{n} f(\mathbf{z_i}|x_i; \boldsymbol{\theta}^{(t)}) \\ &= \prod_{i=1}^{n} \frac{f(\mathbf{z_i}, x_i; \boldsymbol{\theta}^{(t)})}{f(x_i; \boldsymbol{\theta^{(t)}})} \\ &= \prod_{i=1}^{n} \frac{\prod_{k=1}^{K} [\pi_k p_k^{x_i}(1-p_k)^{1-x_i}]^{z_{ik}}}{\sum_{k=1}^{K} \pi_k p_k^{x_i}(1-p_k)^{1-x_i}} \end{aligned}$$

Step 3: Calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
Since $\log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ is linear in z_{ik}, we only to need to know $E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x}; \boldsymbol{\theta}^{(t)}]$ to calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

$$E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x}; \boldsymbol{\theta}^{(t)}] = \frac{\pi_k^{(t)} p_k^{x_i,(t)}(1-p_k)^{1-x_i,(t)}}{\sum_{j=1}^{K} \pi_j^{(t)} p_j^{x_i,(t)}(1-p_j)^{1-x_i,(t)}}$$

Define $\hat{z}_{ik}^{(t)} = E_{\mathbf{Z}|\mathbf{X}=\mathbf{x}}[Z_{ik}|\mathbf{x}; \boldsymbol{\theta}^{(t)}] = \frac{\pi_k^{(t)} p_k^{x_i,(t)}(1-p_k)^{1-x_i,(t)}}{\sum_{j=1}^{K} \pi_j^{(t)} p_j^{x_i,(t)}(1-p_j)^{1-x_i,(t)}}$.

Therefore,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik}^{(t)} \log(\pi_k) + \hat{z}_{ik}^{(t)} \log(p_k^{x_i}(1-p_k)^{1-x_i})$$

Step 4: Calculate $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

- $\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)}$.
- $p_k^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_i}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}}$.

The derivation of these results is analagous to the process shown above for the Normal mixture model.

Simulation of EM algorithm on Normal Mixture Model

Suppose that we have collected gene expression data for gene Y from four distinct cell populations. Unfortunately, the data we collected was not labelled and as such, we do not know from which of the four populations a given data point was collected. We will use the EM algorithm to estimate the parameters of the original four data generating distributions.

Generation of Simulated Data

```
> set.seed(500)
> n = 1000
> # mixture components
> mu.values <- c(5, 18, 12, 25)
> sigma.values <- c(0.5, 1.8, 1.5, 0.5)
>
> # determine Z_i
> p.values <- c(0.1, 0.3, 0.5, 0.1)
> Z <- rmultinom(n, 1, p.values)
>
> # sample from mixture model
> mu.vector <- apply(Z, 2, function(x) (sum(x*mu.values)))
> sd.vector <- apply(Z, 2, function(x) (sum(x*sigma.values)))
> d <- rnorm(n, mean = mu.vector, sd=sqrt(sd.vector))
> hist(d, breaks = 50, main = "", xlab = "Gene Expression", freq = FALSE, ylim = c(0, 0.3))
> title(main = "Histogram of Gene Expression Data", line=-1)
```

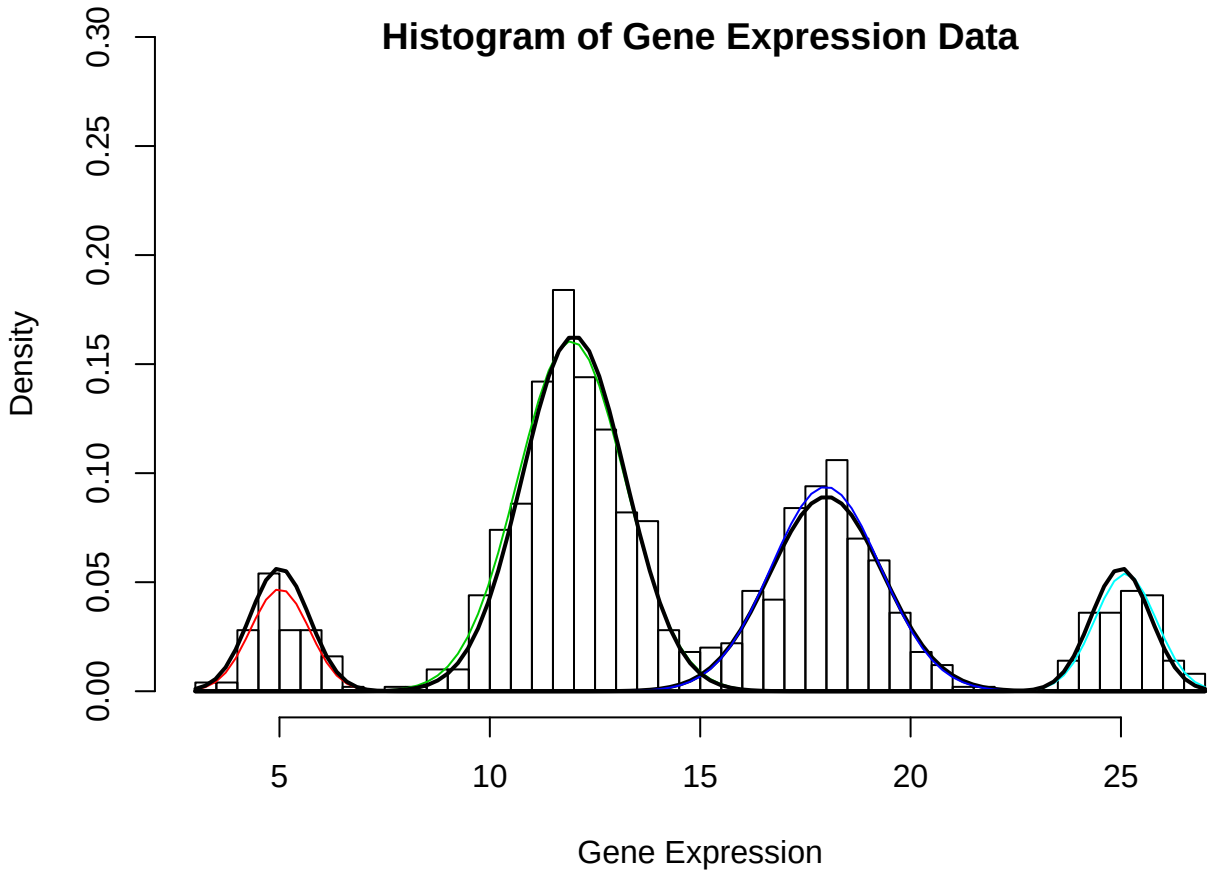
We observe the histogram of the collected data and note that the data from each of the four populations

appears to be normally distributed. Hence, it is suitable for us to proceed with a normal mixture model.

Implementation of EM Algorithm on Simulated Data

```
> # Initialization of the values
> B <- 100
> K <- 4
> p <- matrix(0,B,K)
> mu <- matrix(0,B,K)
> s <- matrix(0,B,K)
> p[1,] <- rep(0.25, K)
> mu[1,] <- c(mean(sort(d)[1:250]),mean(sort(d)[251:500]),
+             mean(sort(d)[501:750]),mean(sort(d)[751:1000]))
> s[1,] <- c(var(sort(d)[1:250]),var(sort(d)[251:500]),
+             var(sort(d)[501:750]),var(sort(d)[751:1000]))
> z <- matrix(0, K, n)
>
> # EM Algorithm
> for (i in 2:B) {
+   denom <- rep(0, n)
+   for (w in 1:K) {
+     denom <- denom + p[i-1, w] * dnorm(d, mean=mu[i-1,w], sd=sqrt(s[i-1,w]))
+   }
+   for (j in 1:K) {
+     z[j, ] <- p[i-1,j] * dnorm(d, mean=mu[i-1,j], sd=sqrt(s[i-1,j])) / denom
+     mu[i, j] <- sum(z[j, ]*d) / sum(z[j,])
+     s[i, j] <- sum(z[j, ]* (d - mu[i, j])^2) / sum(z[j,])
+     p[i, j] <- sum(z[j, ]) / length(z[j, ])
+   }
+ }
```

```
> hist(d, breaks = 50, main = "", xlab = "Gene Expression", freq = FALSE, ylim = c(0, 0.3))
> title(main = "Histogram of Gene Expression Data", line=-1)
>
> for (i in 1:K) {
+   x<-seq(-2*sqrt(s[B,i]) + mu[B, i],2*sqrt(s[B,i]) + mu[B, i],by=0.001)
+   curve(dnorm(x,mean=mu[B,i],sd=sqrt(s[B,i]))*p[B,i], add=TRUE, col = i+1)
+   curve(dnorm(x,mean=mu.values[i],sd=sqrt(sigma.values[i]))*p.values[i],
+         add=TRUE, col = 1, lwd = 2)
+ }
```



The normal distribution fits when I do not know the truth (colored lines) closely approximate the normal distributions when all parameters of the normal mixture model are known (bold black lines). This is an idealized example, so in practice we cannot expect the normal distribution fits when we are missing data to approximate the true normal mixture model to this great of an extent.

Markov Chain Monte Carlo

Motivation

Please note that this method is very slow, it does not scale well to high dimensional data sets and there are tuning steps that are tricky.

In Bayesian inference it is often possible to calculate

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto L(\boldsymbol{\theta};\mathbf{x})f(\boldsymbol{\theta})$$

but it is typically much more difficult to calculate

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\boldsymbol{\theta};\mathbf{x})f(\boldsymbol{\theta})}{f(\mathbf{x})}$$

because $f(\mathbf{x})$ tends to be very difficult to compute.

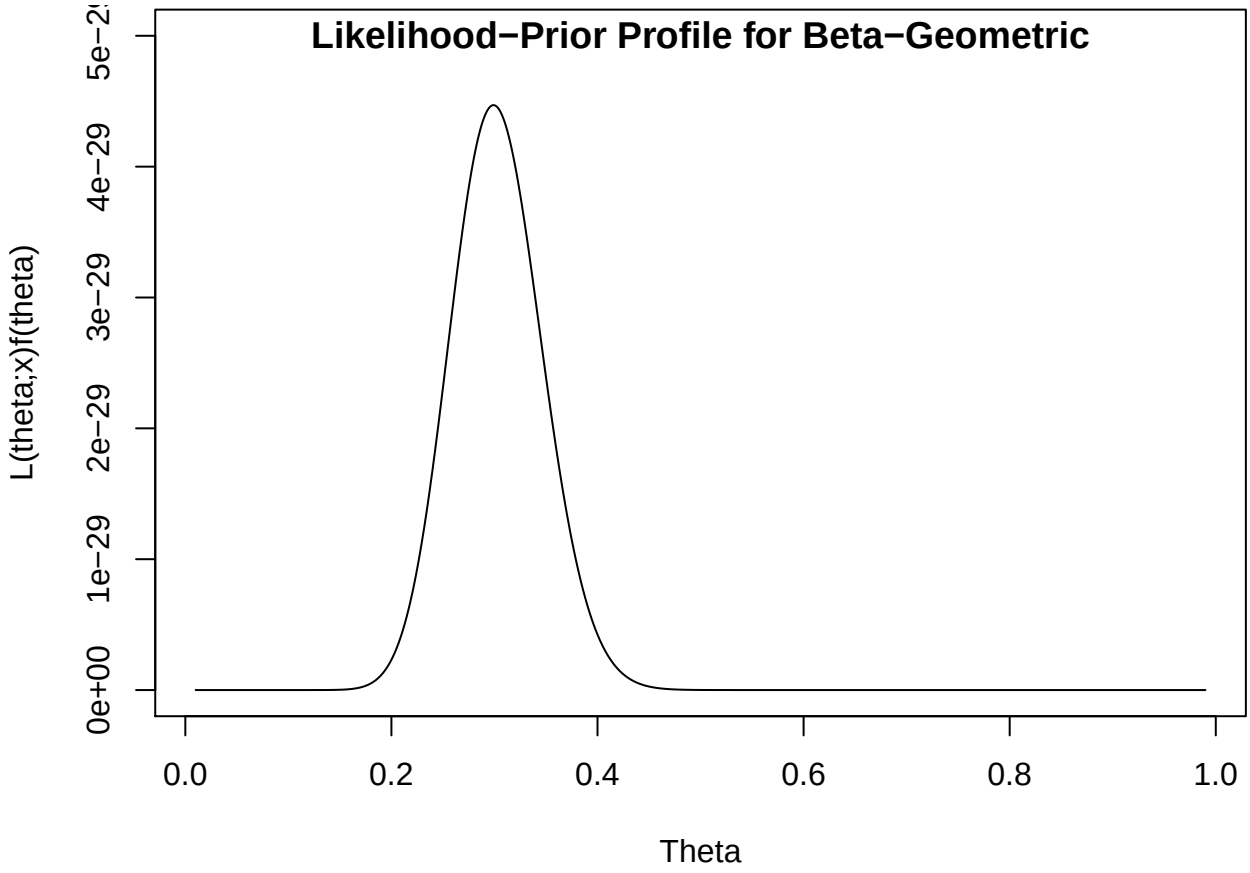
Given that $L(\boldsymbol{\theta};\mathbf{x})f(\boldsymbol{\theta})$ is easier to compute, MCMC works directly with this and neglects to explicitly compute the posterior probability. It is a method for simulating data from $f(\boldsymbol{\theta}|\mathbf{x})$ with knowlegde of only $L(\boldsymbol{\theta};\mathbf{x})f(\theta)$.

To illustrate the essence of the MCMC method, consider $\mathbf{X}|\theta \overset{iid}{\sim}$ Geometric(θ) with $\theta \sim$ Beta(α,β).

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto L(\theta;\mathbf{x})f(\theta) \\ &\propto \prod_{i=1}^{n}(1-\theta)^{x_i}\theta \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+n-1}(1-\theta)^{\beta+\sum_{i=1}^{n}x_i-1} \end{aligned}$$

Note that $L(\theta;\mathbf{x})f(\theta) \propto \theta^{\alpha+n-1}(1-\theta)^{\beta+\sum_{i=1}^{n}x_i-1}$ has the following profile over different realization of θ.

```
> set.seed(10)
> n <- 30
> x <- rgeom(n, prob = 0.3)
> theta.values <- seq(0.01,0.99,by=0.001)
> alpha <- 3
> beta <- 2
> y <- sapply(theta.values, function(z) z^(alpha + n - 1)*(1 - z)^(beta + sum(x) - 1))
> plot(theta.values, y, type = "l", ylab = "L(theta;x)f(theta)", xlab = "Theta",
+      ylim = c(0, 5*10^(-29)))
> title(main = "Likelihood-Prior Profile for Beta-Geometric", line=-1)
```



It is clear that $L(\theta;\mathbf{x})f(\theta)$ is very high in the range $(0.2, 0.4)$ and very low elsewhere. This indicates that these "high" areas are areas of high posterior probability and the "low" areas are areas of low posterior

probability. To convert $L(\theta;\mathbf{x})f(\theta)$ into a probability distribution, we can use Monte Carlo integration which involves taking random points along $L(\theta;\mathbf{x})f(\theta)$ and then summing them to get a numerical integration of the profile. This is easy to do when θ is a single parameter as in the above example. If $\boldsymbol{\theta}$ is a vector of many parameters, we have higher dimensional integration which makes this Monte Carlo integration approach computationally infeasible. Markov Chain Monte Carlo resolves this issue by spending more time in high probability areas such that the numerical integration is taken where the function $L(\theta;\mathbf{x})f(\theta)$ has high values thereby ensuring computational feasibility.

Key Properties of MCMC

- We draw a Markov chain of $\boldsymbol{\theta}$ values so that asymptotically they are equivalent to iid draws from $f(\boldsymbol{\theta}|\mathbf{x})$.
- The draws are done *competitively* so that the next draw of a realization of $\boldsymbol{\theta}$ dependes on the current value.
- The Markov chain is set up so that it only dependes on $L(\boldsymbol{\theta};\mathbf{x})f(\boldsymbol{\theta})$ and $f(\mathbf{x})$ does not need to be known.

Metropolis Hastings Algorithm

1. Initialize $\boldsymbol{\theta}^{(0)}$.
2. Generate $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ for some pdf or pmf $q(\cdot|\cdot)$.
3. With probability
$$A(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(b)}) = \min(1, \frac{L(\boldsymbol{\theta}^*;\mathbf{x})f(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(b)};\mathbf{x})f(\boldsymbol{\theta}^{(b)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(b)})})$$
set $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^*$. Otherwise, set $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^{(b)}$.
4. Continue for $b = 1, 2, \ldots, B$ iterations and carefully select which $\boldsymbol{\theta}^{(b)}$ are utilized to approximate iid observations from $f(\boldsymbol{\theta}|\mathbf{x})$.

Additional Remarks

- The probability distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ must selected such that there is some probability of generating new parameter values given current parameter values. It may be necessary to test out different $q(\cdot|\cdot)$. Furthermore, $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ should be chosen such that new values of $\boldsymbol{\theta}^{(b+1)}$ are generated within a reasonable number of iterations.
- Taking the ratio of $L(\boldsymbol{\theta}^*;\mathbf{x})f(\boldsymbol{\theta}^*)$ to $L(\boldsymbol{\theta}^{(b)};\mathbf{x})f(\boldsymbol{\theta}^{(b)})$ means that we tend to choose new $\boldsymbol{\theta}$ values with a higher posterior probability. We do not want to do this determinstically, however, because we might get trapped in local maxima. As such, we must multiply this ratio by the random component, $\frac{q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(b)})}$, which introduces enough stochastic behaviour in our selection of the new $\boldsymbol{\theta}$ so that we do not get trapped in local maxima.

Metropolis Algorithm

The Metropolis algorithm restricts $q(\cdot|\cdot)$ to be symmetric so that $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(b)}) = q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^*)$ and

$$A(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(b)}) = \min(1, \frac{L(\boldsymbol{\theta}^*;\mathbf{x})f(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(b)};\mathbf{x})f(\boldsymbol{\theta}^{(b)})})$$

Utilizing MCMC Output

There are two common uses for the output from MCMC.

1. $E[f(\boldsymbol{\theta}|\mathbf{x})]$ is approximated by

$$\hat{E}[f(\boldsymbol{\theta})|\mathbf{x}] = \frac{1}{B}\sum_{b=1}^{B} f(\boldsymbol{\theta}^{(b)}) \overset{B\to\infty}{\longrightarrow} E[f(\boldsymbol{\theta})|\mathbf{x}]$$

   Note that $f(\cdot)$ can be the indicator function so that we have an point estimate for $\boldsymbol{\theta}$.
2. Some subsequence $\boldsymbol{\theta}^{(b_1)}, \boldsymbol{\theta}^{(b_2)}, \ldots, \boldsymbol{\theta}^{(b_m)}$ from $\{\boldsymbol{\theta}^{(b)}\}_{b=1}^{B}$ is utilized as an empirical approximation to iid draws from $f(\boldsymbol{\theta}|\mathbf{x})$.

Note that the algorithm must be run for a certain number of iterations before observed $\boldsymbol{\theta}^{(b)}$ can be utilized.

Session Information

```
> sessionInfo()
R version 3.6.2 (2019-12-12)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS Catalina 10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gapminder_0.3.0 forcats_0.4.0   stringr_1.4.0   dplyr_0.8.4
 [5] purrr_0.3.3     readr_1.3.1     tidyr_1.0.2     tibble_2.1.3
 [9] ggplot2_3.2.1   tidyverse_1.3.0 knitr_1.28

loaded via a namespace (and not attached):
 [1] tidyselect_1.0.0 xfun_0.12        haven_2.2.0      lattice_0.20-38
 [5] colorspace_1.4-1 vctrs_0.2.3      generics_0.0.2   htmltools_0.4.0
 [9] yaml_2.2.1       rlang_0.4.4      pillar_1.4.3     withr_2.1.2
[13] glue_1.3.1       DBI_1.1.0        dbplyr_1.4.2     modelr_0.1.5
[17] readxl_1.3.1     lifecycle_0.1.0  munsell_0.5.0    gtable_0.3.0
[21] cellranger_1.1.0 rvest_0.3.5      evaluate_0.14    fansi_0.4.1
[25] broom_0.5.4      Rcpp_1.0.3       scales_1.1.0     backports_1.1.5
[29] jsonlite_1.6.1   fs_1.3.1         hms_0.5.3        digest_0.6.24
[33] stringi_1.4.6    grid_3.6.2       cli_2.0.1        tools_3.6.2
[37] magrittr_1.5     lazyeval_0.2.2   crayon_1.3.4     pkgconfig_2.0.3
[41] xml2_1.2.2       reprex_0.3.0     lubridate_1.7.4  assertthat_0.2.1
[45] rmarkdown_2.1    httr_1.4.1       rstudioapi_0.11  R6_2.4.1
[49] nlme_3.1-142     compiler_3.6.2
```