ECON6300/7320/8300 Advanced Microeconometrics Bayesian Methods

Christiern Rose

[1]University of Queensland

Lecture 9

A Thought Experiment:

- Experiment: select a person randomly on St. Lucia campus
- Event A := a randomly selected person is taller than 190cm. Then,

$$\Pr(A) = ?$$

- Event B := a randomly selected person is female.
- A female is selected. Then, what are the odds that she is 190cm or taller?

$$\Pr(A|B) = ?$$

Subjective Probability

- Each of one us can say what we believe about $\Pr(A)$, although it might be wrong
- When we learn that a female is selected, we may revise our belief about the person being taller than 190cm. Mathematically, we write it $\Pr(A|B)$
- Your belief before the additional information, $\Pr(A)$, is called the prior.
- Your (revised) belief after the information, $\Pr(A|B)$, is the posterior.

Bayes Theorem

- Recall that

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \text{ and similarly } \Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

- So, $\Pr(A \cap B) = \Pr(A)\Pr(B|A)$.

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$$

- This is a version of **Bayes theorem**.
- Bayes theorem shows how the prior is updated to the posterior when there is additional information.

Bayes Theorem for PDF

- Reprint Bayes theorem:

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$$

- Let $\pi(x)$ be the marginal PDF of X.
- Let $p(y)$ be the marginal PDF of Y.

- Let $f(y|x)$ be the conditional PDF of Y given $X = x$.
- By Bayes theorem, we write

$$\pi(x|y) = \frac{\pi(x)f(y|x)}{p(y)}$$

- $\pi(x)$ is the prior on X and $\pi(x|y)$ is the posterior on X given $Y = y$

Bayes Theorem for PDF

- If we use θ (the parameter) instead of x:

$$\pi(\theta|y) = \frac{\pi(\theta) f(y|\theta)}{p(y)}.$$

- $\pi(\theta)$ is the prior on θ and $\pi(\theta|y)$ is the posterior on θ given $Y = y$.
- The parameter (θ) may not be random. But, it is uncertain.

- Our subjective belief about θ is represented as $\pi(\theta)$ (prior belief, before data) and $\pi(\theta|y)$ (posterior belief, after data).

Bernouilli Example

- A coin is tossed and $Y = 1$ if the outcome is Head and otherwise $Y = 0$.
- Let $\theta := \Pr(Y = 1)$, i.e., $Y \sim \mathit{Bernouilli}(\theta)$. Then,

$$f(y|\theta) = \theta^y (1-\theta)^{1-y} \mathbb{1}(y \in \{0, 1\})$$

- Suppose you believe every $\theta \in (0, 1)$ is equally likely. Then, the prior is

$$\theta \sim \mathrm{Uniform}(0, 1), \text{ i.e., } \pi(\theta) = \mathbb{1}(\theta \in (0, 1))$$

- Then, the posterior is

$$\pi(\theta|y) = \underbrace{\mathbb{1}(\theta \in (0, 1))}_{\pi(\theta)} \cdot \underbrace{\theta^y (1-\theta)^{1-y}}_{f(y|\theta)} / p(y)$$

- If $y = 1$,

$$\begin{aligned} \pi(\theta|y = 1) &= \mathbb{1}(\theta \in (0, 1)) \cdot \theta / p(1) \\ &= 2\theta \cdot \mathbb{1}(\theta \in (0, 1)) \end{aligned}$$

- What is $\pi(\theta|y = 0)$?

Bernouilli Example

- To be more flexible, we consider the beta prior, i.e.,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1))$$

- Since $\theta \sim \textit{Beta}(\alpha, \beta)$, we know that

$$E[\theta] = \frac{\alpha}{\alpha+\beta}$$

$$V[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Suppose you believe that $\theta = 0.25$ on average, i.e., $E[\theta] = 1/4$, and your belief is quite strong, e.g., $V[\theta] = 0.01$. Then, $(\alpha, \beta) \approx (4, 13)$.
- As before, let's compute $\pi(\theta|y = 1)$.

Bernouilli Example

- To solve this problem, we need to learn a useful trick:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1))$$

$$= c \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1))$$

where c is the normalising constant (i.e. it does not depend on θ).

- Once we know

$$\pi(\theta) = c \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1)),$$

we also know $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ (Why?).

- So, even if we simply write

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1)),$$

we immediately know

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

Bernouilli Example

- Bayes theorem:

$$
\begin{aligned}
\pi(\theta|y) &= \pi(\theta)f(y|\theta)/p(y) \propto \pi(\theta)f(y|\theta) \\
&\propto \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}\cdot\mathbb{1}(\theta\in(0,1))}_{\propto\pi(\theta)}\cdot\underbrace{\theta^{y}(1-\theta)^{1-y}}_{f(y|\theta)} \\
&= \theta^{(\alpha+y)-1}(1-\theta)^{(\beta+1-y)-1}\cdot\mathbb{1}(\theta\in(0,1)) \\
&= \theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}\cdot\mathbb{1}(\theta\in(0,1))
\end{aligned}
$$

  where $\tilde{\alpha}=\alpha+y$ and $\tilde{\beta}=\beta+1-y$.

- The posterior density is

$$\pi(\theta|y) \propto \theta^{\tilde{\alpha}-1}(1-\theta)^{\tilde{\beta}-1}\cdot\mathbb{1}(\theta\in(0,1))$$

  which means

$$\theta|y \sim Beta(\tilde{\alpha},\tilde{\beta})$$

Bernouilli Example

- Since the posterior is *Beta*$(\tilde{\alpha}, \tilde{\beta})$, the posterior mean and variance when $y = 1$, are

$$E[\theta|y = 1] = \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} = \frac{\alpha + 1}{\alpha + \beta + 1},$$

$$V[\theta|y = 1] = \frac{\tilde{\alpha}\tilde{\beta}}{(\tilde{\alpha} + \tilde{\beta})^2(\tilde{\alpha} + \tilde{\beta} + 1)} = \frac{(\alpha + 1) \cdot \beta}{(\alpha + \beta + 1)^2(\alpha + \beta + 2)}$$

- Since $(\alpha, \beta) = (4, 13)$, we have

$$E[\theta|y = 1] = 0.29 > E[\theta] = 0.25$$
$$V[\theta|y = 1] = 0.0104 > V[\theta] = 0.01$$

- Since $y = 1$ observed, we believe $\theta = \Pr(y = 1)$ should be larger.
- Since the data contradicts our prior, uncertainty around our belief on θ increases. The variance increases.

Example: Normal Model

- When $X \sim \mathcal{N}(\mu, \sigma^2)$, the PDF of X is given as

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Suppose $y|\theta \sim \mathcal{N}(\theta, 1)$. The PDF is

$$f(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right)$$

- The prior on θ is given as $\mathcal{N}(0, 10^2)$.

$$\pi(\theta) = \frac{1}{\sqrt{2\pi 10^2}} \exp\left(-\frac{1}{2 \cdot 10^2}\theta^2\right)$$

Example: Normal Model

- Posterior density

$$\pi(\theta|y) \propto \underbrace{\exp\left(-\frac{1}{2\cdot 10^2}\theta^2\right)}_{\propto \pi(\theta)} \underbrace{\exp\left(-\frac{1}{2}(y-\theta)^2\right)}_{\propto f(y|\theta)}$$

$$= \exp\left\{-\frac{1}{2}\left(y^2 - 2y\theta + \theta^2 + \theta^2/100\right)\right\}$$

$$\propto \exp\left(-\frac{1}{2(100/101)}\left(\theta - (100/101)y\right)^2\right)$$

- This implies that the conditional distribution of θ given $Y = y$ is normal with mean $(100/101)y$ and variance $100/101$, i.e.,

$$\theta|y \sim \mathcal{N}\left(\frac{100y}{101}, \frac{100}{101}\right)$$

- Hence, $E[\theta|Y = y] = \left(\frac{100}{101}\right) y$ and $V[\theta|Y = y] = \frac{100}{101}$.

Conjugate Priors

- Previously, we considered an example where we observed Y given the unknown mean θ, and θ had a normal prior distribution. The posterior distribution turned out to be normal as well.
- This was an example of a conjugate prior distribution, which has the property that the posterior distribution is of the same family as the prior.
- Conjugacy is specific to prior and model combination. For example,
  - Normal prior is conjugate for Normal model for mean parameter
  - Beta prior is conjugate for Bernouilli model
  - But, for instance, Beta prior is not conjugate for Normal model.
- Conjugate priors are useful: the posterior is a standard distribution with known properties, e.g, posterior mean and posterior variance are easy to compute. (as above)

Conjugate Priors: an example of non-conjugate prior

- Consider Beta prior for the normal model above. The posterior is

$$\pi(\theta|y) \propto \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \mathbb{1}(\theta \in (0,1))}_{\propto Beta(\alpha,\beta)} \cdot \underbrace{\exp\left(-\frac{1}{2}(y-\theta)^2\right)}_{\propto \mathcal{N}(\theta,1)}$$

- Let $\pi^*(\theta|y)$ denote the RHS, i.e., $\pi(\theta|y) = c \cdot \pi^*(\theta|y)$ for some $c > 0$.
- There is no standard distribution proportional to $\pi^*(\theta|y)$.
- In order to obtain the posterior, we need to find c such that

$$\int_0^1 \pi(\theta|y)d\theta = \int_0^1 c \cdot \pi^*(\theta|y)d\theta = 1$$

Conjugate Priors: an example of non-conjugate prior

- Then, the posterior is

$$\pi(\theta|y) = \frac{\pi^*(\theta|y)}{\int_0^1 \pi^*(\tilde{\theta}|y)d\tilde{\theta}}$$

- Moreover, the posterior mean and posterior variance are

$$E[\theta|y] = \int_0^1 \theta \cdot \pi(\theta|y)d\theta = \int_0^1 \theta \left[\frac{\pi^*(\theta|y)}{\int_0^1 \pi^*(\tilde{\theta}|y)d\tilde{\theta}}\right] d\theta$$

$$V[\theta|y] = E\left[(\theta - E[\theta|y])^2 | y\right] = \int_0^1 (\theta - E[\theta|y])^2 \left[\frac{\pi^*(\theta|y)}{\int_0^1 \pi^*(\tilde{\theta}|y)d\tilde{\theta}}\right] d\theta$$

- The integrals need to be numerically evaluated. These days, computers are fast and algorithms are good. We will learn a few algorithms later.
- Still prior conjugacy is important: there are still limitations on what can be handled by computational algorithms.

Prior vs Sample

- Consider the model and prior as follows

$$Y|\theta \sim \mathcal{N}(\theta, \tau^{-1})$$

$$\theta \sim \mathcal{N}(d, h^{-1})$$

where τ and h are precisions (precision = inverse variance). Assume τ is known.

- Normal prior for normal mean is conjugate. Turns out the posterior is

$$\theta|Y \sim \mathcal{N}\left(\left[\frac{h}{h+\tau}\right] d + \left[\frac{\tau}{h+\tau}\right] Y, (h+\tau)^{-1}\right)$$

- The posterior mean is a weighted average of the prior information and the sample information
- The posterior variance is the sum of prior and sample precisions.

Prior vs Sample, and Improper prior

- Reprint the posterior distribution;

$$\theta|Y = y \sim \mathcal{N}\left(\left[\frac{h}{h+\tau}\right]d + \left[\frac{\tau}{h+\tau}\right]y, (h+\tau)^{-1}\right)$$

  where h is prior precision.
- If you are really sure about the value of θ before observing y. Then, your prior precision is high, i.e., $h \uparrow$. The weight on the prior mean is large and the weight on the observation is small. So, $E[\theta|y] \approx d = E[\theta]$
- On the other hand, if you are uncertain about θ, h would be small, and the weight on the observation would be relatively large.

Prior vs Sample, and Improper prior

- Reprint the posterior distribution;

$$\theta | Y = y \sim \mathcal{N}\left(\left[\frac{h}{h+\tau}\right] d + \left[\frac{\tau}{h+\tau}\right] y, (h+\tau)^{-1}\right)$$

  where h is prior precision.

- If $h \to 0$, the prior density gets flat, putting no weight on the prior mean.
- At the limit, $\pi(\theta) \propto 1 \Rightarrow$ often called uniform/flat prior, or even non informative prior.

- The flat prior is not a density because it does not integrate to 1. So, it is called an improper prior. But, the posterior is still a density.

More than one observation

- Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from $f(\cdot|\theta)$ and we have $\pi(\theta)$.
- When only $Y_1 = y_1$ is observed, the posterior is given as

$$\pi(\theta|y_1) \propto \pi(\theta) f(y_1|\theta)$$

- When $Y_2 = y_2$ is also observed, $\pi(\theta|y_1)$ is the prior to be updated

$$\pi(\theta|y_1, y_2) \propto \underbrace{\pi(\theta|y_1)}_{\text{new prior}} f(y_2|\theta) \propto \pi(\theta) f(y_1|\theta) f(y_2|\theta)$$

- When $Y_3 = y_3$ is observed, similarly

$$\pi(\theta|y_1, y_2, y_3) \propto \underbrace{\pi(\theta|y_1, y_2)}_{\text{new prior}} f(y_3|\theta) \propto \pi(\theta) f(y_1|\theta) f(y_2|\theta) f(y_3|\theta)$$

- So, the posterior is proportional to the prior times the likelihood

$$\pi(\theta|y_1, \ldots, y_n) \propto \pi(\theta) \cdot \prod_{i=1}^{n} f(y_i|\theta)$$

More than one observation

- Suppose we have a random sample from the normal model

$$Y_1 = y_1, \ldots, Y_n = y_n | \theta \overset{iid}{\sim} \mathcal{N}(\theta, \tau^{-1})$$

where τ is known and we have the normal prior $\theta \sim \mathcal{N}(d, h^{-1})$.

- Then, applying the conjugacy recursively, we have the posterior

$$\theta | y_1, ., y_n \sim \mathcal{N}\left(\left[\frac{h}{h + n\tau} \right] d + \left[\frac{n\tau}{h + n\tau} \right] \left[\frac{1}{n} \sum_{i=1}^{n} y_i \right], (h + n\tau)^{-1} \right)$$

- For all $(d, h) \in \mathbb{R} \times \mathbb{R}_+$, i.e., for all proper priors, as $n \to \infty$,
  - the posterior mean gets close to the sample mean
  - $\frac{1}{n} \sum_{i=1}^{n} y_i \xrightarrow{p} \theta_0$, the true population mean (LLN)
  - the posterior precision explodes
  - So, the posterior will be degenerate at θ_0 (Consistency)

Optimal Decision under Uncertainty

- Let $L(\theta, a)$ be the loss when you choose an action $a \in \mathcal{A}$ under $\theta \in \Theta$.
- If you knew the true θ, you would minimise $L(\theta, a)$
- But, θ is unknown.
- You have prior $\pi(\theta)$ and observe $y_1, \ldots, y_n \overset{iid}{\sim} f(y|\theta)$.
- Under axioms of Savage (1954), it is optimal to choose

$$
\begin{aligned}
a_B &:= \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a)\pi(\theta|y_1, \ldots, y_n)d\theta \\
&= \arg\min_{a \in \mathcal{A}} E[L(\theta, a)|y_1, \ldots, y_n]
\end{aligned}
$$

which is called the Bayes action. (Subjective Expected Utility Theory)

Estimation as a Decision Problem

- The econometrician would suffer the error square loss

$$L(\theta, a) = (\theta - a)^2$$

when she reports $a \in \Theta$ as her estimate when θ is the true.

- For this particular problem, it is optimal to choose

$$\hat{\theta}_B := \arg\min_{a \in \Theta} \int_{\Theta} (\theta - a)^2 \pi(\theta | y_1, \ldots, y_n) d\theta \underbrace{=}_{\text{turns out}} E[\theta | y_1, \ldots, y_n]$$

- If $L(\theta, a) = |\theta - a|$, it is optimal to report the posterior median.

- As seen here, the optimality depends on the loss function (the way you feel about the error you make)
- In the example of the normal model, the Bayesian estimate is the posterior mean under either of the loss functions above.

Summary of Uncertainty

- The posterior variance or the posterior standard deviation can be reported as a summary of uncertainty around the estimate.
- Alternatively, a 95% credible interval (CI) can also be reported, which is any interval $[a, b]$ such that

$$0.95 = \Pr(\theta \in [a, b] | y_1, \ldots, y_n) = \int_a^b \pi(\theta | y_1, \ldots, y_n) d\theta$$

- There could be many such intervals (not unique).
- The narrowest one is generally preferred and we use here.
- Interpretation is natural. (compare with 95% confidence interval!)
- For the normal model above, the 95% CI is

$$\hat{\theta}_B \pm 1.96\sqrt{V(\theta | y_1, \ldots, y_n)}$$

Linear Regression

- Consider a regression model

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i \\ &= x_i' \beta + e_i \end{aligned}$$

where we assume that $e_i | x_i, \beta \sim \mathcal{N}(0, 1/\tau)$, and β and x_i are $(K \times 1)$ vectors. Since $e_i = y_i - x_i'\beta$, we have

$$y_i - x_i' \beta \Big| x_i, \beta \sim \mathcal{N}\left(0, 1/\tau\right),$$

the likelihood is

$$\prod_{i=1}^{n} \underbrace{\sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2} e_i^2\right)}_{\text{PDF of } \mathcal{N}(0,1/\tau)} \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2\right)$$

$$= \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2} (y - X\beta)'(y - X\beta)\right)$$

Linear Regression

- Priors specification;

$$\tau \sim Gamma(\alpha_0/2, 2/\lambda_0)$$

$$\beta|\tau \sim \mathcal{N}\left(b_0, (\tau\Sigma_0)^{-1}\right)$$

where α_0 and λ_0 are positive scalars b_0 is a K vector, and Σ_0 is a $K \times K$ positive definite matrix.

- The prior is

$$\pi(\tau, \beta) \propto \underbrace{\tau^{\frac{\alpha_0}{2}-1} \exp\left(-\frac{\lambda_0}{2}\tau\right) \cdot \mathbb{1}(\tau > 0)}_{\propto \pi(\tau)} \times \underbrace{\tau^{\frac{K}{2}} \exp\left(-\frac{\tau}{2}(\beta - b_0)'\Sigma_0(\beta - b_0)\right)}_{\propto \pi(\beta|\tau)}$$

Linear Regression

- The posterior is the prior times the likelihood;

$$
\begin{aligned}
\pi(\tau, \beta | y, X) \propto & \tau^{\frac{\alpha_0}{2}-1} \exp\left(-\frac{\lambda_0}{2}\tau\right) \cdot \mathbb{1}(\tau > 0) \\
& \cdot \tau^{\frac{k}{2}} \exp\left(-\frac{\tau}{2}(\beta - b_0)'\Sigma_0(\beta - b_0)\right) \\
& \cdot \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}(y - X\beta)'(y - X\beta)\right)
\end{aligned}
$$

- To derive the posterior, it is useful to know

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta}_{LS})'(y - X\hat{\beta}_{LS}) + (\beta - \hat{\beta}_{LS})'(X'X)(\beta - \hat{\beta}_{LS})$$

where $\hat{\beta}_{LS}$ is the OLS estimate, i.e., $\hat{\beta}_{LS} = (X'X)^{-1}X'y$

Linear Regression

- Some additional algebra shows

$$\tau|y, X \sim Gamma(\alpha_n/2, 2/\lambda_n)$$

$$\beta|\tau, y, X \sim \mathcal{N}\left(b_n, (\tau\Sigma_n)^{-1}\right)$$

  where

$$\Sigma_n := (X'X + \Sigma_0)$$

$$b_n := \Sigma_n^{-1}(X'X\hat{\beta}_{LS} + \Sigma_0 b_0) = (X'X + \Sigma_0)^{-1}(X'y + \Sigma_0 b_0)$$

  and $\alpha_n := \alpha_0 + n$ and $\lambda_n := \lambda_0 + y'y + b_0'\Sigma_0 b_0 - b_n'\Sigma_n b_n$
- Note that the posterior mean of β is the weighted average of the OLS estimate $\hat{\beta}_{LS}$ and the prior b_0.
- Moreover, conditional on τ, the posterior precision of β is the sum of the OLS precision and the prior precision.

Linear Regression

- If the econometrician has the mean squared error, i.e.,

$$L(\theta, a) = (\theta - a)'(\theta - a), \text{ for } a \in \Theta$$

where $\theta = (\beta_1, \ldots, \beta_2, \tau)$ is a parameter vector, her optimal estimate is the posterior mean.

- Each β_j for $j = 1, \ldots, K$ is normally distributed under the posterior,

$$\beta_j | \tau, y, X \sim \mathcal{N}\left(b_{n,j}, V(\beta_j | \tau, y, X)\right)$$

where $V(\beta_j|\tau, y, X)$ is the (j, j) element of $(\tau\Sigma_n)^{-1}$.

- We often summarise the uncertainty about β_j by the posterior standard deviation $\sqrt{V(\beta_j|\tau, y, X)}$ or the (narrowest) 95% credible interval

$$b_{n,j} \pm 1.96\sqrt{V(\beta_j|\tau, y, X)}.$$

Bayesian Computation

- What if the prior is not conjugate? Then there is no analytic expression for the posterior...
- Consider a generic problem where θ is the parameter, $\pi(\theta)$ is the prior, and $f(y|\theta)$ is the density of Y given θ. Then, the posterior is

$$\pi(\theta|y) \propto \pi(\theta) f(y|\theta)$$

  where $y = (y_1, \ldots, y_n)$ is a sample and $f(y|\theta)$ is the density of y given θ.
- We are often interested in the posterior moment;

$$E[h(\theta)|y] = \int_\Theta h(\theta)\pi(\theta|y)d\theta$$

  where $h(\theta)$ is a measurable function of θ, e.g., $h(\theta) = \theta$ gives the posterior mean, $h(\theta) = (\theta - E[\theta|y])^2$ the posterior variance, and so on.

Bayesian Computation

- Suppose we can draw $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}$ from $\pi(\theta|y)$.
- If the sample has a good property (for example, i.i.d), then

$$\frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) \xrightarrow{p} E[h(\theta)|y]$$

- So, we may evaluate the integrals using a simulated sample from the posterior (Monte Carlo).
- For the Bayesian computation, especially, a Markov chain Monte Carlo approach is widely used.
- We outline the Metropolis-Hastings algorithm.

Bayesian Computation

- Let $q(\tilde{\theta}|\theta)$ be a conditional density, called the proposal density.
- At the s^{th} iteration of the Metropolis-Hastings:
  1. Draw a 'candidate'

     $$\tilde{\theta} \sim q(\tilde{\theta}|\theta^{(s-1)})$$

  2. Let $\theta^{(s)} := \tilde{\theta}$ with probability

     $$Q := \min\left\{1, \frac{\pi(\tilde{\theta}|y)}{\pi(\theta^{(s-1)}|y)} \cdot \frac{q(\theta^{(s-1)}|\tilde{\theta})}{q(\tilde{\theta}|\theta^{(s-1)})}\right\}$$

     and let $\theta^{(s)} := \theta^{(s-1)}$ with probability $1 - Q$.
- The normalising constants are cancelled out and the posterior ratio is computable.

Bayesian Computation

- Note that the sequence $\theta^{(1)}, \ldots, \theta^{(S)}$ is a Markov chain, i.e., the distribution of $\theta^{(s)}$ depends only on the previous value $\theta^{(s-1)}$.
- Theoretically, the chain $\theta^{(1)}, \ldots, \theta^{(S)}$ has a property good enough to approximate the posterior, i.e., regardless of the initial point $\theta^{(0)}$,

$$\frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) \xrightarrow{p} E[h(\theta)|y] \text{ as } S \to \infty.$$

- The Metropolis Hastings algorithm is an example of a Markov chain Monte Carlo (MCMC) method.

Bayesian Computation

- In practice, the quality of an MCMC outcome can heavily depend on the initial point $\theta^{(0)}$, the proposal density $q(\cdot|\cdot)$, and the model $f(\cdot|\theta)$.
- For example, if $\theta^{(0)}$ is far away from the posterior distribution, it may take a long time for the chain to reach to the support of the posterior.
- So, we exclude some early draws that seem to be a 'searching' phase rather than draws from the posterior, i.e., we discard the burn-in draws.
- If the proposal density is similar to the posterior density, the algorithm works well. But, in practice, it is hard to know the shape of the posterior.

Bayesian Computation

- The normal distribution is often used. In particular, $q(\tilde{\theta}|\theta)$ is the density of $\mathcal{N}(\theta, \Omega)$. Then,

$$q(\tilde{\theta}|\theta) \propto \exp\left[-\frac{1}{2}\left(\frac{\tilde{\theta}-\theta}{\sigma_\theta}\right)^2\right].$$

- Since $q(\tilde{\theta}|\theta) = q(\theta|\tilde{\theta})$, the acceptance rate simplifies to

$$Q = \min\left\{1, \frac{\pi(\tilde{\theta}|y)}{\pi(\theta^{(s-1)}|y)}\right\}.$$

- The algorithm is called the Gaussian Metropolis-Hastings algorithm.
- Often t distribution is used.

- Metropolis-Hastings algorithms require the researcher to tune up the proposal density. For the Gaussian MH, tuning parameter is σ_θ.
- If σ_θ is too big, Q would be small (as $\tilde{\theta}$ would usually be far from $\theta^{(s-1)}$) and the algorithm may not often update. That is, it can be $\theta^{(s)} = \theta^{(s+1)} = \theta^{(s+2)} = \cdots$ for a long time.
- If σ_θ is too small, Q is close to one accepting almost all candidates, but the chain moves very slowly. $\theta^{(s)} \approx \theta^{(s+1)} \approx \theta^{(s+2)} \approx \cdots$
- Either case, the chain does not effectively explore the posterior.

- It is important to choose σ_θ that enables the chain to explore the posterior efficiently, but it is not always easy.
- So, a number of adaptive methods have been proposed. For example, Haario, Saksman, and Tamminen (2001).
- If $\theta = (\theta_1, \ldots, \theta_K)$ is high dimension, it can be extremely hard to tune the proposal density. For the Gaussian MH, a *K* dimensional covariance matrix has to be chosen.
- For a high-dimensional problem, Gibbs sampler is widely used.
- Idea is to recursively update one component or a small dimensional sub-vector of θ.

Bayesian Computation

- Gibbs sampler: at each s^{th} iteration,
  1. draw $\theta_1^{(s)} \sim \pi(\theta_1 | \theta_2^{(s-1)}, \ldots, \theta_K^{(s-1)}, \text{data})$
  2. draw $\theta_2^{(s)} \sim \pi(\theta_2 | \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_K^{(s-1)}, \text{data})$
  3. draw $\theta_3^{(s)} \sim \pi(\theta_3 | \theta_1^{(s)}, \theta_2^{(s)}, \theta_4^{(s-1)}, \ldots, \theta_K^{(s-1)}, \text{data})$
  4. $\vdots$
  5. draw $\theta_K^{(s)} \sim \pi(\theta_K | \theta_1^{(s)}, \ldots, \theta_K^{(s)}, \text{data})$
- At each-substep, the Gibbs sampler updates each θ_k using either the (conditional) conjugacy or the MH algorithm (called MH within Gibbs).
- Each sub-step may update more than one component, i.e., it could update a block.

Model Selection

- We observe a sample $y := (y_1, \ldots, y_N)$ but don't know the true model.
- We may assume $y \sim f(y|\theta)$ with unknown $\theta \in \Theta$; Model 1.
- Or, we could assume $y \sim g(y|\psi)$ with unknown $\psi \in \Psi$; Model 2.

- We have two competing models (theories, assumptions), $\mathcal{M} := \{1, 2\}$.
- Suppose we have conditional priors $\pi(\theta|m = 1)$ and $\pi(\psi|m = 2)$.

Model Selection

- Conditional on $m = 1$, recall that the posterior of θ is

$$\pi(\theta|y, m = 1) = \frac{\pi(\theta|m = 1)f(y|\theta)}{\int \pi(\tilde{\theta}|m = 1)f(y|\tilde{\theta})d\tilde{\theta}} = \frac{\pi(\theta|m = 1)f(y|\theta)}{h(y|m = 1)}$$

  where $h(y|m = 1)$ is the **marginal likelihood** of model $m = 1$.

- Similarly, conditional on $m = 2$,

$$\pi(\psi|y, m = 2) = \frac{\pi(\psi|m = 2)g(y|\psi)}{\int \pi(\tilde{\psi}|m = 2)g(y|\tilde{\psi})d\tilde{\psi}} = \frac{\pi(\theta|m = 2)g(y|\psi)}{h(y|m = 2)}$$

  where $h(y|m = 2)$ is the **marginal likelihood** of model $m = 1$.

- Then, the (marginal) likelihood ratio of model 1 relative to model 2 is

$$B_{1,2} := \frac{h(y|m = 1)}{h(y|m = 2)}$$

  which is called the **Bayes factor**.

Model Selection

- Suppose we believe model $m \in \mathcal{M}$ is true with prior probability $\pi(m)$. Then, the posterior probability of model $m = 1$ is given as

$$\pi(m=1|y) = \frac{\pi(m=1)h(y|m=1)}{\pi(m=1)h(y|m=1) + \pi(m=2)h(y|m=2)}$$

- The posterior odd ratio of model 1 to model 2 is

$$\frac{\pi(m=1|y)}{\pi(m=2|y)} = \frac{\pi(m=1)h(y|m=1)}{\pi(m=2)h(y|m=2)}$$

- That is, the posterior odd ratio = prior odd ratio $\times$ Bayes factor
- If the posterior odd ratio > 1, we believe model 1 is more reasonable.
- BIC and AIC are rough approximations of this formal Bayesian model selection.

Bayesian Asymptotics

- Under mild conditions, posterior is consistent, i.e., the posterior asymptotically degenerates at the true parameter; Doob (1941), Schwartz (1965)
- The posterior and the sampling distribution of MLE are asymptotically equivalent under some regularity conditions on the model and the prior. (Bernstein - von Mises theorem)
- For large sample, therefore, the Bayesian analysis is robust to the choice of prior. Moreover, since MLE is efficient, Bayes is also efficient
- Asymptotic robustness can be used as a reaction to the criticism that Bayesian analysis is essentially subjective (depends on the prior).