CS 660: Mathematical Foundations of Analytics

Dr. Francis Parisi

Pace University

Spring 2018

IMS Chapter 11 – Bayesian Statistics

Bayesian Statistics

- Conceptually Bayesian analysis may seem backwards
- Use some known outcome to predict what lead to that outcome
- This is the basis of Bayes Theorem, and more broadly Bayesian analysis
- Start with some beliefs, then based on evidence, adjust those beliefs

Bayesian Statistics

Example

- It's a presidential election with only two candidates, Democrat and Republican, but you don't pay attention to the outcome
- Months later you learn that corporate taxes have been eliminated
- From this evidence, is it more likely that the Democrat was elected, or the Republican?

Bayesian Statistics

Example (continued...)

- Polls show the two candidates are equally likely to win so

$$P[R] = P[D] = 0.5$$

- Experts say the probability of a tax-cut if the Democrat wins is 0.25, and the probability of a tax-cut if the Republican wins is 0.85 or

$$P[T|D] = 0.25 \text{ and } P[T|R] = 0.85$$

- The probability of a tax-cut is then

$$P[T] = P[T|D] \times P[D] + P[T|R] \times P[R] = 0.55$$

Bayesian Statistics

Using Bayes Theorem we have ...

$$P[R|T] = \frac{P[T|R] \times P[R]}{P[T]} = 0.772$$

and

$$P[D|T] = \frac{P[T|D] \times P[D]}{P[T]} = 0.227$$

Using the outcome “corporate tax-cut” to infer that it is likely the Republican won

Bayesian Statistics

Prior belief

$$P[R] = P[D] = 0.5$$

Evidence: Corporate Tax cut

Posterior belief

$$P[R] = 0.772, \text{ and } P[D] = 0.227$$

Definition (Convergence in Probability)

A sequence of random variables $\{X_n\}$ getting "closer" to another random variable X, as $n \to \infty$

$$\lim_{n\to\infty} P\left[|X_n - X| \geq \epsilon\right] = 0$$

or

$$\lim_{n\to\infty} P\left[|X_n - X| < \epsilon\right] = 1$$

If so we write

$$X_n \xrightarrow{P} X$$

Definition (Weak Law of Large Numbers)

Let $\{X_n\}$ be a sequence of iid random variables having common mean μ and variance $\sigma^2 < \infty$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then

$$\overline{X}_n \xrightarrow{P} \mu$$

And...

Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$

Suppose $X_n \xrightarrow{P} X$ and a is a constant, then $aX_n \xrightarrow{P} aX$

Suppose $X_n \xrightarrow{P} a$ and the real function g is continuous at a, then $g(X_n) \xrightarrow{P} g(a)$

Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$

Bayesian Statistics

- A **statistic** is a function of a random sample, e.g. the sample mean
- $\overline{X}$ is a ststistic and a point estimate of the mean μ
- $\overline{X}$ is **unbiased** if $\mathbb{E}\,\overline{X} = \mu$

Definition (Consistent Statistic)

Let X be a random variable with cdf $F(x, \theta)$, $\theta \in \Omega$. Let $X_1, \ldots, X_n$ be a sample from the distribution of X and let T_n denote a statistic. We say T_n is a **consistent** estimator of θ if

$$T_n \xrightarrow{P} \theta$$

For Converge in Probability we said a sequence gets "close" to a random variable, but how close?

Definition (Convergence in Distribution)

Let $\{Xn\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of X_n and X. Let $C(F_X)$ denote the set of all points where F_X is continuous. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \text{ for all } x \in C(F_X).$$

We denote this by

$$X_n \xrightarrow{D} X.$$

Bayesian Statistics

- In our discussion of the normal distribution and statistical inference we learned that

$$\frac{\overline{X} - \mu}{\sigma / sqrtn} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1)$$

- The **Central Limit Theorem** (CLT) asserts that is $X_1, \ldots, X_n$ are observations from any distribution with finite variance $\sigma^2 > 0$ then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} N(0, 1)$$

Bayesian Statistics

Sufficient Statistics

- A statistic is a data reduction technique
- $X_1, \ldots, X_n$ are n observations that can be summarized as the sample mean, $\overline{X}$
- We often use statistics to make some inference about some parameter θ
- A statistic, say $Y = u(X_1, \ldots, X_n)$ which contains all the information about θ in the sample is a **sufficient statistic** and its conditional distribution does not depend on θ

Bayesian Statistics

Bayesian Procedures

- We want to learn something about a parameter θ
- We have some prior beliefs or assumptions about its distribution (prior distribution)
- We get some data and the data revise our beliefs (posterior distribution)
- Bayesian statistics accounts for any prior knowledge the statistician has about the experiment

Bayesian Statistics

- Consider a random variable X whose distribution depends on some parameter θ
- θ is an element of a well-defined set Ω
- Now Θ is a random variable with a probability distribution over Ω
- So x is a value that the random variable X can take, and θ is a value that the random variable Θ can take

Bayesian Statistics

- Denote the pdf of Θ as $h(\theta)$ and $h(\theta) = 0$ when $\theta \notin \Omega$
- $h(\theta)$ is the **prior** pdf of Θ
- Denote the conditional pdf of $X|\theta$ as $f(x|\theta)$

Bayesian Statistics

- If $X_1, \ldots, X_n$ is a random sample from the conditional distribution of $X|\Theta = \theta$ and $\mathbf{X}^T = (X_1, \ldots, X_n)$ and $\mathbf{x}^T = (x_1, \ldots, x_n)$
- The joint conditional pdf of $\mathbf{X}$ given $\Theta = \theta$ is

$$L(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

- The joint pdf of $\mathbf{X}$ and Θ is

$$g(\mathbf{x}, \theta) = L(\mathbf{x}|\theta)h(\theta)$$

Bayesian Statistics

- The joint marginal pdf of $\mathbf{X}$ is

$$g_1(\mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{x}, \theta) d\theta$$

  when Θ is continuous – replace the integral with summation when Θ is discrete
- Finally, the conditional pdf of Θ given the sample $\mathbf{X}$ is

$$k(\theta|\mathbf{x}) = \frac{g(\mathbf{x}, \theta)}{g_1(\mathbf{x})}$$

- $k(\theta|\mathbf{x})$ is the **posterior** pdf and defines the posterior distribution of Θ

Bayesian Statistics

- The **prior distribution** reflects the belief **before** the sample is drawn
- The **posterior distribution** reflects the belief **after** the sample is drawn
- The posterior distribution contains all the information about the parameter

Bayesian Statistics

Point Estimation

- Finding a point estimator for θ amounts to selecting a *decision* function δ so $\delta(\mathbf{x})$ is a predicted value of θ
- A Bayes estimator is a decision function δ that minimizes the expected value of the loss function $\mathcal{L}[\theta, \delta(\mathbf{x})]$

$$\mathbb{E}[\mathcal{L}[\theta, \delta(\mathbf{x})]|\mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} \mathcal{L}[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

- So

$$\delta(\mathbf{x}) = \text{Argmin} \int_{-\infty}^{\infty} \mathcal{L}[\theta, \delta(\mathbf{x})]k(\theta|\mathbf{x})d\theta$$

Bayesian Statistics

- The random variable $\delta(\mathbf{X})$ is the Bayes estimator θ
- If the loss function $\mathcal{L}[\theta, \delta(\mathbf{x})] = [\theta - \delta(\mathbf{x})]^2$ the Bayes estimator is the mean, $\mathbb{E}[\Theta|\delta(\mathbf{x})]$
- If the loss function $\mathcal{L}[\theta, \delta(\mathbf{x})] = |\theta - \delta(\mathbf{x})|$ the Bayes estimator is the median

Bayesian Statistics

- Suppose there is a sufficient statistic $Y = u(\mathbf{X})$ for the parameter so that

$$L(\mathbf{x}|\theta) = g[u(\mathbf{x})|\theta]H(\mathbf{x})$$

  where $g(y|\theta)$ is the pdf of $Y|\Theta = \theta$
- Then $k(\theta|\mathbf{x}) \propto g[u(\mathbf{x})|\theta]h(\theta)$ or

$$k(\theta|y) \propto g(y|\theta)h(\theta)$$

- The marginal pdf of Y is

$$g_1(y) = \int_{-\infty}^{\infty} g(y|\theta)h(\theta)d\theta$$

Bayesian Statistics

Example

Consider the model

$$\begin{aligned} X_i|\theta &\sim \ iid \ \text{Bin}(1,\theta) \\ \Theta &\sim \ \text{Beta}(\alpha,\beta), \qquad \alpha,\beta \text{ known} \end{aligned}$$

so the prior pdf is

$$h(\theta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0<\theta<1 \\ 0, & \text{elsewhere,} \end{cases}$$

where α, β positive and constant
We seek a decision function δ that's a Bayes solution (estimator)

Example (Continued...)

Our sufficient statistic is $Y = \sum_{i=1}^{n} X_i$ which has a $\text{Bin}(n, \theta)$ distribution

The conditional pdf is

$$g(y\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

The posterior distribution is

$$k(\theta|y) \propto \theta^y (1-\theta)^{n-y} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \quad 0 < \theta < 1$$

Bayesian Statistics

The posterior is proportional to a Beta distribution so we have

$$k(\theta|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+y)\Gamma(n+\beta-y)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}, \quad 0 < \theta < 1$$

If we use the squared-error loss, $\mathcal{L}[\theta, \delta(\mathbf{x})] = [\theta - \delta(\mathbf{x})]^2$ then the Bayesian point estimate of θ that minimizes the expected value of $\mathcal{L}$ is mean of this beta pdf

$$\delta(y) = \frac{\alpha + y}{\alpha + \beta + n}$$

Example (Continued...)

Our prior distribution assumption was

$$\Theta \sim \text{Beta}(\alpha, \beta)$$

and our posterior distribution given our sample $X_1, \ldots, X_n$ and $Y = \sum_1^n X_i$ is

$$\Theta \sim \text{Beta}(\alpha + y, \beta + n - y)$$

Bayesian Statistics

Example

Suppose we are interested in the true mortality rate θ in a hospital H which is about to try a new operation. On average in the country around 10% of people die, but mortality rates in different hospitals vary from around 3% up to 20%. Hospital H has no deaths in their first 10 operations. What is the posterior distribution for the mortality rate? What is the expected mortality rate for hospital H?

$X = 1$ if the patient dies, zero otherwise. $X_1, \ldots, X_n$ is a random sample from a Bernoulli distribution, and suppose the prior distribution is $\text{Beta}(\alpha, \beta)$.

$$X_i|\theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(\alpha, \beta) = h(\theta)$$

Bayesian Statistics

Example (Continued...)

Now $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ and

$$L(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

so

$$L(\mathbf{x}|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

and

$$k(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)h(\theta)$$

or

$$\begin{aligned} k(\theta|\mathbf{x}) &\propto \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta-1} \end{aligned}$$

Bayesian Statistics

Example (Continued ...)

This is the $\text{Beta}\left(\sum x_i + \alpha, n - \sum x_i + \beta\right)$ distribution

$$k(\theta|\mathbf{x}) = \frac{\theta^{\sum x_i + \alpha - 1}(1-\theta)^{n - \sum x_i + \beta - 1}}{B(\sum x_i + \alpha, n - \sum x_i + \beta)}$$

This answers the first question. Now what is an estimate of the mortality rate for this procedure at Hospital H?

Bayesian Statistics

We know that on average about 10% die from the procedure so if we choose, $\alpha = 3, \beta = 27$ we get $\alpha/(\alpha + \beta) = 0.1$

Additionally these parameters give $P\left[0.03 < \theta < 0.20\right] = 0.9$ so these values for α and β match our data well

From the data, $\sum x_i/n = 0$ so the mle is $\hat{\theta} = 0$, although unlikely no one will ever die from the procedure

But our Bayesian estimate from $\text{Beta}(\sum x_i + \alpha, n - \sum x_i + \beta)$ is $3/40 = 0.075$, is more plausible

So Hospital H can expect about a 7.5% mortality rate for this procedure

Bayesian Statistics

Interval Estimation

- We can find an interval estimate for θ
- Find two functions $u(\mathbf{x})$ and $v(\mathbf{x})$ so

$$P\left[u(\mathbf{x}) < \Theta < v(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right] = \int_{u(\mathbf{x})}^{v(\mathbf{x})} k(\theta|\mathbf{x}) d\theta$$

  is large, for example 0.95
- This is a 95% **credible** interval, not to confuse this with a confidence interval

Bayesian Statistics

The simplified comparison between a **confidence** interval in classical statistics and a **credible** interval in Bayesian statistics is as follows...

- In classical statistics, we take the true parameter to be a specific but unknown value and we say there is a 95% chance the true value lies within the confidence interval. Thus, the probability statement is about the interval rather than about the location of the true parameter value.
- In Bayesian statistics we take the true but unknown parameter to be a random variable and we estimate the uncertainty in the parameter location. Credible intervals capture our current uncertainty in the location of the parameter values and thus can be interpreted as probabilistic statement about the parameter.

Bayesian Statistics

- We can use Bayesian methods to test hypotheses
- $H_0 : \theta \in \omega_0$ versus $H_1 : \theta \in \omega_1$
- We use a simple rule to inform our decision, we accept H_0 if

$$P\left[\Theta \in \omega_0 | \mathbf{x}\right] \geq P\left[\Theta \in \omega_1 | \mathbf{x}\right]$$

Bayesian Statistics

Example

$\mathbf{X}^T = (X_1, X_2, \ldots, Xn)$ is a random sample from a Poisson distribution with mean θ, suppose we are interested in testing

$$H_0 : \theta \leq 10 \text{ versus } H_1 : \theta > 10.$$

Suppose we think θ is about 12, but we're not quite sure. So we choose the $\Gamma(10, 1.2)$ pdf as our prior. The mean of the prior is 12 ($\alpha\beta$) and the variance of the prior distribution is 14.4 ($\alpha\beta^2$).

We observe a sample of size $n = 20$. The value of the sufficient statistic is $y = \sum_{i=1}^{20} x_i = 177$. Hence, the posterior distribution is a $\Gamma(177 + 10, 1.2/[20(1.2) + 1]) = \Gamma(187, 0.048)$ distribution.

Bayesian Statistics

Example (Continued...)

The data result in a mean of 187(0.048) = 8.976, which is the Bayes estimate (under squared-error loss) of θ.

We compute the posterior probability (use the **pgamma** command in R), of H_0 as

$$P\left[\leq 10|y=177\right] = P\left[\Gamma(187, 0.048) \leq 10\right] = 0.9368$$

and

$$P\left[> 10|y=177\right] = 10.9368 = 0.0632$$

Under the rule we would accept H_0. The 95% credible interval is (7.77, 10.31), which also contains 10.

Bayesian Statistics

What happens if we get new information after we conduct our analysis?

The last posterior distribution becomes our new prior distribution and we repeat the analysis

Bayesian analysis is an excellent way to deal with sequential analysis

Each time we get new data we revise our posterior distribution

Bayesian Statistics

From the examples we've looked at we note that we have some choice as to the prior distribution we use

Our choice of prior is based on some assumption or possibly on information about the distribution of the parameter

For example, if we were analyzing hurricane arrivals and we needed a prior for storm counts, a Poisson distribution would make sense

When we don't have any information about the distribution of the parameter we use a **noninformative prior**

A noninformative prior is a prior distribution that treats all values of θ the same

Bayesian Statistics

Gibbs Sampler

- From the preceding sections, it is clear that integration techniques play a significant role in Bayesian inference
- We now touch on some of the Monte Carlo techniques used for integration in Bayesian inference

Bayesian Statistics

Example

Suppose we have the Bayes model where $X \sim N(\theta, \sigma^2)$, then $Y = \overline{X}$ is a sufficient statistic

$$Y|\theta \sim N(\theta, \sigma^2/n)$$
$$\Theta \sim h(\theta) \propto b^1 \exp\{-(\theta - a)/b\}/(1 + \exp\{-(\theta - a)/b\})^2$$

the prior distribution is the logistic pdf The posterior distribution for this is very messy and involves two integrals and has no closed form solution

Monte Carlo simulation allows use to estimate the mean of the logistic and helps us solve this

Bayesian Statistics

Theorem

Suppose we generate random variables by the following algorithm:

1. *Generate* $Y \sim f_Y(y)$
2. *Generate* $X \sim f_{X|Y}(x|Y)$

Then $X \sim f_X(x)$

Bayesian Statistics

- The idea behind this is if we know the distribution of a random variable we and can find the inverse of the cdf, then we can easily simulate a sample
- We use the The Inverse Transformation Method
  If U is a Unif(0,1) random variable and F is a continuous distribution and if

$$Y = F^{-1}(U)$$

  then the random variable Y has distribution function F

Simulating an exponential random variable

Example

If $F(X) = 1 - e^{-x}$ then $F^{-1}(u)$ is that value of x such that

$$1 - e^{-x} = u$$

or

$$x = -\log(1 - u).$$

If U is a Unif(0,1) then

$$F^{-1}(U) = -\log(1 - U)$$

is exponentially distributed with mean 1.

Bayesian Statistics

- The main purpose of presenting this algorithm is to motivate another algorithm, called the **Gibbs Sampler**, which is useful in Bayes methodology
- Suppose (X, Y) has pdf $f(x, y)$
- Our goal is to generate two streams of iid random variables, one on X and the other on Y

Bayesian Statistics

The Gibbs Sampler Algorithm: Let m be a positive integer, and let X_0, an initial value, be given. Then for $i = 1, 2, 3, \ldots, m$,

1. Generate $Y_i | X_{i-1} \sim f(y|x)$
2. Generate $X_i | Y_i \sim f(x|y)$

- Note that before entering the i^{th} step of the algorithm, we have generated X_{i-1}
- Let x_{i-1} denote the observed value of X_{i-1}
- Then, using this value, generate sequentially the new Y_i from the pdf $f(y|x_{i-1})$ and then draw (the new) X_i from the pdf $f(x|y_i)$, where y_i is the observed value of Y_i

It can be shown that $Y_i \xrightarrow{D} Y \sim f_Y(y)$ and $X_i \xrightarrow{D} X \sim f_X(x)$

Bayesian Statistics

- Consider the sequence of generated pairs

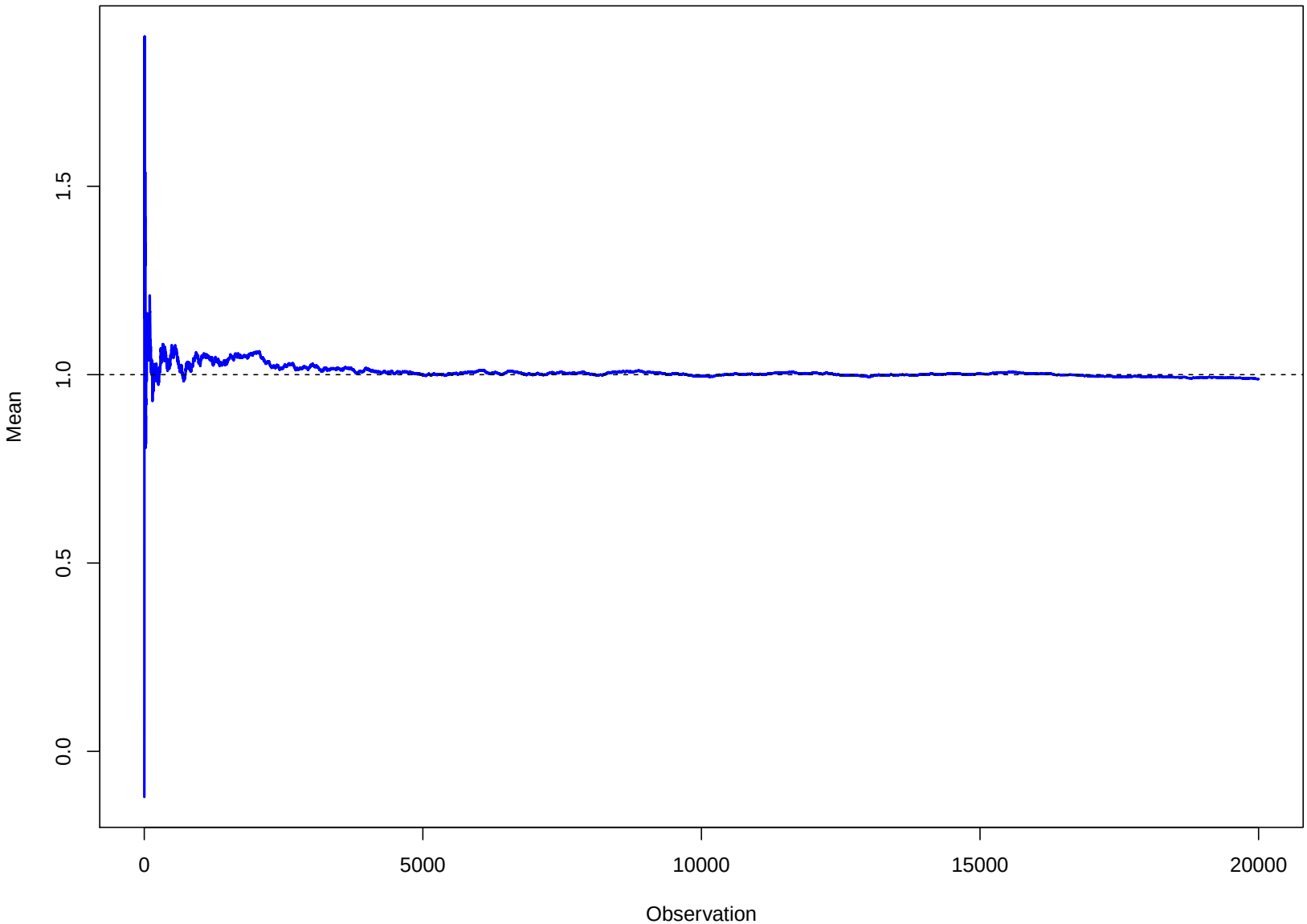
$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k), (X_{k+1}, Y_{k+1})$$

- Note that to compute (X_{k+1}, Y_{k+1}), we need only the pair (X_k, Y_k) and none of the previous pairs from 1 to $k - 1$
- Given the present state of the sequence, the future of the sequence is independent of the past
- In stochastic processes such a sequence is called a **Markov chain**
- The distribution of Markov chains stabilizes (reaches an equilibrium distribution) as the length of the chain increases

Bayesian Statistics

- For the Gibbs sampler, the equilibrium distributions are the limiting distributions in the expression (44) as $i \to \infty$
- How large should i be? In practice, usually the chain is allowed to run to some large value i before recording the observations
- This is known as the “burn-in”

Bayesian Statistics



Bayesian Statistics

We'll consider a couple of other approaches to Bayesian analysis

1. Hierarchical Bayes
2. Empirical Bayes

Bayesian Statistics

Hierarchical Bayes

- We've seen that the prior pdf plays an important role in Bayesian inference
- We find the the different Bayes estimators based on different priors and loss functions
- A way to have more control over the prior is to model the prior in terms of another random variable
- This approach is called the **hierarchical** Bayes model
- The form of the hierarchical Bayes model extends the regular Bayes model

$$X|\theta \sim f(x|\theta)$$
$$\Theta|\gamma \sim h(\theta|\gamma)$$
$$\Gamma \sim \psi(\gamma)$$

Bayesian Statistics

- Note that in model the prior distribution for Θ, $h(\theta|\gamma)$, is now a conditional distribution
- We have some control over the prior by modifying the pdf of the random variable Γ
- The parameter γ can be thought of as a nuisance parameter, and is often called a **hyperparameter**
- In hierarchical Bayes the inference focuses on the parameter θ so our primary interest is still the posterior conditional pdf $k(\theta|\mathbf{x})$

Bayesian Statistics

- We use the Gibbs sampler to find the Bayes estimator for the hierarchical Bayes model
- For the arithmetic average for Θ we have

$$\frac{1}{m}\sum_{i=1}^{m} W(\Theta_i) \xrightarrow{P} \mathbb{E}[W(\Theta_i)|\mathbf{x}] = \delta(\mathbf{x})$$

  as $m \to \infty$
- The Gibbs sampler produces a stream of values $(\theta_1, \gamma_1),\ (\theta_2, \gamma_2),\ \ldots$ via Monte Carlo simulation

Bayesian Statistics

- By choosing large values of m and $n^* > m$ we estimate an average for $W(\theta)$

$$\frac{1}{n^* - m} \sum_{i=m+1}^{n^*} W(\theta_i)$$

- Because we use Monte Carlo simulations to generate a Markov Chain, these procedures are often called **Markov Chain Monte Carlo** or **MCMC**

Bayesian Statistics

Example

Let's look at the following hierarchical Bayes model

$$
\begin{aligned}
X|\lambda &\sim \text{Poisson}(\lambda) \\
\Lambda|b &\sim \Gamma(1, b) \\
B &\sim g(b) = \tau - 1b - 2\exp -1/b\tau,\ b > \theta,\ \tau > 0
\end{aligned}
$$

Bayesian Statistics

We need two conditional pdfs for the Gibbs sampler

$$g(\lambda|x,b) \propto \lambda^{x+1-1}e^{-\lambda[1+(1/b)]}$$

which is $X|\lambda \times \Lambda|b$

This is the $\Gamma(x+1, b/[b+1])$distribution and

$$g(b|x,\lambda) \propto b^{-3}\exp\left\{ -\frac{1}{b}\left[\frac{1}{\tau}+\lambda\right]\right\}$$

which is $\Lambda|b \times B$

Applying the change of variable $y = 1/b$ to the last expression we have the $\Gamma(2, \tau/[\lambda\tau+1])$ distribution
The joint pdf is

$$g(x,\lambda,b) = f(\mathbf{x}|\lambda)h(\lambda|b)\psi(b)$$

Bayesian Statistics

The Gibbs sampler for this model is then

$$\Lambda_i | x, b_{i-1} \sim \Gamma(x+1, b_{i-1}/[b_{i-1}+1])$$

$$B_i = Y_i^{-1}, \quad \text{where} \quad Y_i | x_i, \lambda_i \sim \Gamma(2, \tau/[\lambda_i \tau + 1])$$

The authors provide a function **`hierach1.s`** as part of the R package **`hmcpkg`** that can run this sampler

An example of the output is in the text

Bayesian Statistics

Empirical Bayes

- In the hierarchical Bayes models we model γ as a random variable
- Instead **Empirical Bayes**, obtains an estimate of γ and plugs it into the posterior pdf
- Thus the first two parts of the model are the same as hierarchical Bayes

$$\begin{aligned} X|\theta &\sim f(x|\theta) \\ \Theta|\gamma &\sim h(\theta|\gamma) \end{aligned} \tag{11.1}$$

- We estimate γ on the data

Bayesian Statistics

Recall

$$g(\mathbf{x}, \theta|\gamma) = f(\mathbf{x}|\theta)h(\theta|\gamma)$$

then the likelihood function is

$$m(\mathbf{x}|\gamma) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)d\theta$$

Using maximum likelihood we get $\hat{\gamma} = \hat{\gamma}(\mathbf{x})$

Finally, for inferences about θ, our parameter of interest, we use the posterior pdf

$$k(\theta|\mathbf{x}, \hat{\gamma})$$

Bayesian Statistics

Key Points

- In Bayesian analysis we start with some prior belief about tthe distribution of a parameter of interest θ
- We make observations and use the data to inform our beliefs
- We end up with a revised (posterior) distribution of θ

Bayesian Statistics

Key Points

Bayes theorem in plain English:

$$Posterior\ distribution = \frac{prior \times likelihood}{\sum prior \times likelihood}$$

or if the distribution is continuous the denominator is an integral

Denominator is summed (integrated) over all possible priors and is often difficult to evaluate

The Gibbs sampler and Markov Chain Monte Carlo offer a solution

References

Hogg, R. V., McKean, J. W. and Craig, A. T. (2018).
Introduction to Mathematical Statistics.
Pearson, Boston, eighth edition.

Ross, S. (2010).
A First Course in Probability.
Prentice Hall, Upper Saddle River, eigth edition.

Wu, J. and Coggeshall, S. (2012).
Foundations of Predictive Analytics.
CRC Press, Boca Raton, FL.