Econ 413R: Computational Economics

Spring Term 2012

Bayesian Inference and Computation

1 Statistical Preliminaries

Understanding what marginal pdf's are and how they are derived will allow us to go a bit deeper into the theory behind Bayesian methods. It will be helpful if you familiarize/re-familiarize yourself with a few things:

i. Conditional probability: $p(x|y)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

ii. Joint density functions: $f(x,y)$

$$f(x,y) = P(x = X, y = Y)$$

iii. Marginal density functions: $f_x(x), f_y(y)$

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy, \quad f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

iv. Conditional pdf's:

$$f(x|y) = \frac{f(x,y)}{f_y(y)}, \quad f(y|x) = \frac{f(x,y)}{f_x(x)}$$

1.1 Example:

Let X and Y be continuous random variables with a joint pdf of the form

$$f(x,y) = 2(x+y), \qquad 0 \leq x \leq y \leq 1$$

Find the conditional pdf's $f(x|y)$ and $f(y|x)$.

Solution:

The marginal distributions, $f_1(x)$ and $f_2(y)$ with $x, y \in (0,1)$ are given by

$$f_1(x) = \int_x^1 2(x+y)dy = (1+3x)(1-x)$$

$$f_2(y) = \int_0^y 2(x+y)dx = 3y^2$$

Hence the conditional pdf's, $f(x|y)$ and $f(y|x)$ are

$$f(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{2(x+y)}{3y^2}, \qquad 0 \leq x \leq y$$

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{2(x+y)}{(1+3x)(1-x)}, \quad x \leq y \leq 1$$

2 Introduction to Bayes

2.1 What is the Bayesian approach?

Bayesian statistics is distinguished by its ability to use outside information or prior knowledge in making statistical estimates, rather than simply extrapolating estimates from the raw data. This is particularly useful when our sample size n is small, but can also be useful in much broader applications.

2.2 Example 1

You are told the probability of successfully navigating an asteroid field is 3,720 to 1. You successfully navigate the asteroid field on your first try. What is your new assessment of the probability of successfully navigating an asteroid field on your second try?

Without considering prior information, the direct probability estimate would be 100% (one success out of one attempt). If you consider prior information and guess that the odds are less than 100%, then you are an honorary Bayesian.

2.3 Example 2

Consider the following data array. Suppose the values are accurate estimates of healthcare centers (i.e. outpatient clinics, hospitals, family practice), cross tabulated by zip codes. Note we have one missing value.

82	88	75	97	98
100	102	x	105	103
89	98	90	101	100

What would be an appropriate estimate for the missing value x? Would a value such as 300 be suitable? Why not? Now suppose that the estimate is 300 (3 people in a region with 100 people becomes 300 per 10,000 people). You may still want to consider a weighted average between the direct estimate and the prior estimate, considering how small the population is in the given region. What if the population is not so small (i.e. 3,000 people out of 300,000)?[1]

2.4 Bayesians vs. Frequentists

If you are not a Bayesian, then statisticians would refer to you as a “frequentist.” To be clear in how these two parties differ, we describe them below.

[1] Carlin and Louis (2009)

Frequentist approach: Uses repeated sampling from a particular model (pdf) usually referred to as the likelihood function. The model defines the probability distribution of the data based on unknown, but fixed, parameter values. The likelihood function is often written as $L(y;\theta)$ with θ unknown and fixed.

Bayesian approach: Uses a sampling model like the frequentist approach, but θ is not fixed. Rather, it is considered to be a random variable over a *prior* distribution. The prior and the likelihood function are used to compute the the conditional distribution of the unknown values, based on the observed data. This conditional distribution is called the *posterior*.

2.5 Advantages of Bayesian inference

i. Bayesian methods allow the user to formally include prior information

ii. Inferences are based on a pdf rather than a finite sample

iii. The reason for stopping the experimentation does not affect Bayesian estimates (see hw problem 1)

iv. All analysis is based on the posterior distribution. No separate tests are needed.

v. Posterior adjusts to changes in data stream

3 Bayes Theorem

Bayes' Theorem, given in equation 3.1, gives the posterior as a function of the prior and the likelihood function.

Prior distribution: $\pi(\theta)$

Posterior distribution: $p(\theta|y)$

Data/likelihood: $f(y|\theta)$

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y,\theta)}{\int p(y,\theta)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \tag{3.1}$$

or discretely,

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A \cap B_j)}{\sum_{j=1}^{J} P(A \cap B_j)} = \frac{P(A|B_j)P(B_j)}{\sum_{j=1}^{J} P(A|B_j)P(B_j)} \tag{3.2}$$

Note that the posterior is simply the product of the prior and the likelihood, renormalized so that the distribution integrates to one. This gives us the Bayesian catch phrase, which you should now memorize for now and forever:

$$\boxed{\text{The posterior is proportional to the likelihood times the prior}}$$

Mathematically, this can be written with the following:

$$p(\theta|y) \propto f(y|\theta)\pi(\theta) \tag{3.3}$$

Because the prior and posterior are proportional to a multiplicative constant, the likelihood can be multiplied by any constant without affecting the posterior.

In the event that the prior depends on unknown parameters, $\pi(\theta|\eta)$ where η is unknown. We find a prior for $\pi(\cdot)$, called the hyperprior. The same principles in using a prior to find the posterior are applied in this situation, so we will not examine the use of hyperpriors in detail.

4 Prior Distributions

A necessary step in building any sort of statistical model using Bayes' method is determining an appropriate prior. There is unfortunately no step-by-step procedure for finding the best prior. A prior might be based on:

- industry experience, existing literature
- theory (i.e. structural model)
- intuition
- histogram of data

Statisticians often attempt to match important descriptive statistics (i.e. mean, median, 25% or 75% percentiles). Attempting to match percentiles at the extreme ends of the distribution, however, can prove to be both frustrating and can distort the posterior in bad ways. It is even possible to use a prior that is essentially empty, so that the information from the data quickly dominates the posterior. This part of determining a prior is important, but also somewhat artsy so we ignore it.

4.1 Conjugate Priors

Nice properties follow from the use of certain priors that are *conjugate* with the likelihood function $f(y|\theta)$, meaning that they belong to the same distributional family. These properties can simplify computation considerably and give us an easy-to-write, analytic expression for our posterior.

4.1.1 Example: conjugate priors

Let X be the number of Payday lenders located within a given radius around banks in the US. Because X is discrete and is probably clustered around lower numbers, we assume a Poisson likelihood function.

$$f(x|\theta) = \frac{e^{-\theta}\theta^{x}}{x!}, \quad x \in \{0, 1, 2, \ldots\}, \quad \theta > 0$$

A sensible prior, therefore, would have support on the positive real line. For flexibility, we use the gamma distribution

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \quad \theta > 0, \quad \alpha > 0, \quad \beta > 0$$

where β is the *scale* parameter and α determines whether the distribution is one-tailed ($\alpha \leq 1$) or two-tailed ($\alpha > 1$). Using equation **??** and dropping any multiplicative constants gives

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)\pi(\theta) \\ &\propto (e^{-\theta}\theta^{x})(\theta^{\alpha-1}e^{-\theta/\beta}) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)} \qquad (4.1) \end{aligned}$$

This is proportional to gamma distribution with parameters $\hat{\beta} = (1+1/\beta)^{-1}$ and $\hat{\alpha} = x+\alpha$. Density functions uniquely characterize distributions, so $G(\hat{\alpha}, \hat{\beta})$ is the only distribution proportional to 4.1. This information allows us to find means, medians, take random samples, or conduct statistical experiments without having to count every single Payday lender in the country, or doing any kind of numerical integration.

Figure 1 shows the relationship between several other conjugate priors and their resulting posterior distributions.[2]

4.1.2 Can I use more than one prior?

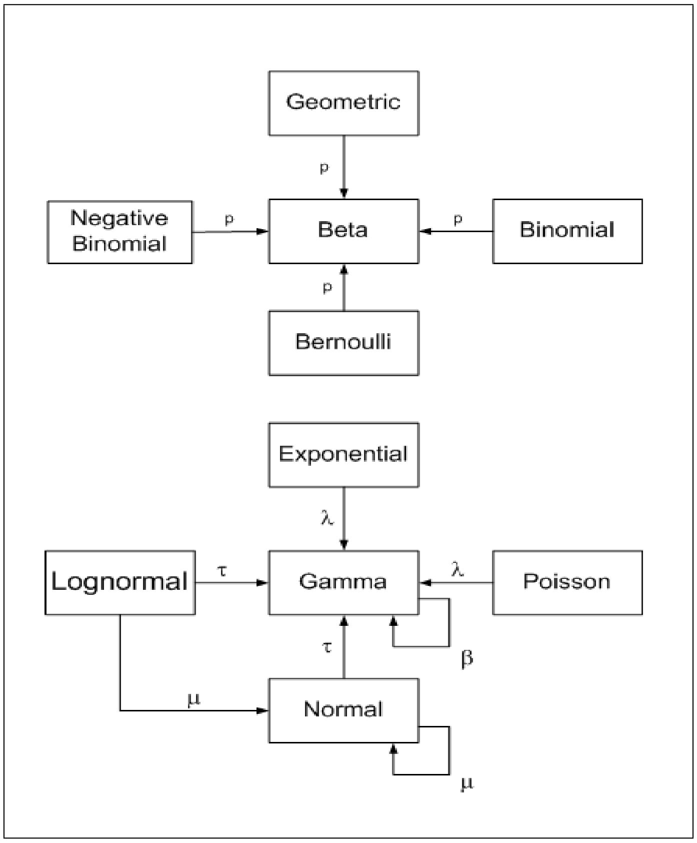
In some cases, it may be useful to use a mixture of conjugate priors. Consider the likelihood function $f(x|\theta)$ and the two component prior $\pi(\theta) = \alpha\pi_1(\theta) + (1-\alpha)\pi_2(\theta)$ with $0 \leq \alpha \leq 1$. Then

$$\begin{aligned} p(\theta|x) &= \frac{f(x|\theta)\pi_1(\theta)\alpha + f(x|\theta)\pi_2(\theta)(1-\alpha)}{\int[f(x|\theta)\pi_1(\theta)\alpha + f(x|\theta)\pi_2(\theta)(1-\alpha)]d\theta} \\ &= \frac{\frac{f(x|\theta)\pi_1(\theta)}{m_1(x)}m_1(x)\alpha + \frac{f(x|\theta)\pi_2(\theta)}{m_2(x)}m_2(x)(1-\alpha)}{m_1(x)\alpha + m_2(x)(1-\alpha)} \\ &= \frac{f(x|\theta)\pi_1(\theta)}{m_1(x)}w_1 + \frac{f(x|\theta)\pi_2(\theta)}{m_2(x)}w_2 \\ &= p_1(\theta|x)w_1 + p_2(\theta|x)w_2 \end{aligned}$$

with $w_1 = m_1(x)\alpha/[m_1(x)\alpha + m_2(x)(1-\alpha)]$ and $w_2 = 1 - w_1$. Thus a combination of priors leads to a similar combination of posteriors.

[2] Wikipedia also has a comprehensive table listing conjugate priors.

Figure 1: Conjugate priors



Source: Cook, J. http://www.johndcook.com

4.2 I don't want to "skew" my data with a prior?

For those who wish to be as objective as possible, it is possible to use "noninformative" priors, such as uniform distributions. This is most easily done over a finite-discrete space or a compact portion of the real line.

Naive empirical Bayesian analysis would directly estimate the prior from the data (using a kernal density function or a close histogram approximation). This is generally frowned upon since it is "using the data twice" and can result in overconfident inferences.

4.3 Using sufficient statistics

With a sample of n independent observations, our likelihood function $f(y|\theta)$ becomes $\prod_{t=1}^{n} f(y_t|\theta)$ which still allows us to use equation **??**. If n is large, we may overcome the curse of dimensionality by using statistic S(y) which is *sufficient* for θ (that is, $f(y|\theta) = h(y)g(S(y)|\theta)$). Here let S(y) = s.

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|u)\pi(u)du} = \frac{h(y)g(S(y)|\theta)\pi(\theta)}{\int h(y)g(S(y)|u)\pi(u)du} = \frac{g(s|\theta)\pi(\theta)}{m(s)} = p(\theta|s)$$

where $m(s) > 0$. Note that h(y) cancels in the middle expression. Thus we can bypass using the entire dataset y if this sufficient statistic is available (see homework problem 4).

4.4 Sequential estimation

Suppose we now have two independent samples of data, y_1 and y_2, drawn at two different periods in time (perhaps a clinical trial or election results that come in from different states at different times). Baye's Theorem may be used sequentially as follows:

$$\begin{aligned} p(\theta|y_1, y_2) &\propto f(y_1, y_2|\theta)\pi(\theta) \\ &= f_2(y_2|\theta)f_1(y_1|\theta)\pi(\theta) \\ &\propto f_2(y_2|\theta)p(\theta|y_1) \end{aligned}$$

In this way, the prior for the second dataset y_2, is simply the posterior from the first dataset, y_1.

5 Confidence Intervals/Credible Sets

The Bayesian term for a confidence interval is a "credible set." A credible set is the probability that θ lies in set C, given the data y is at least $(1 - \alpha)\%$ in C. Mathematically, a $100 * (1 - \alpha)\%$ credible set is expressed as

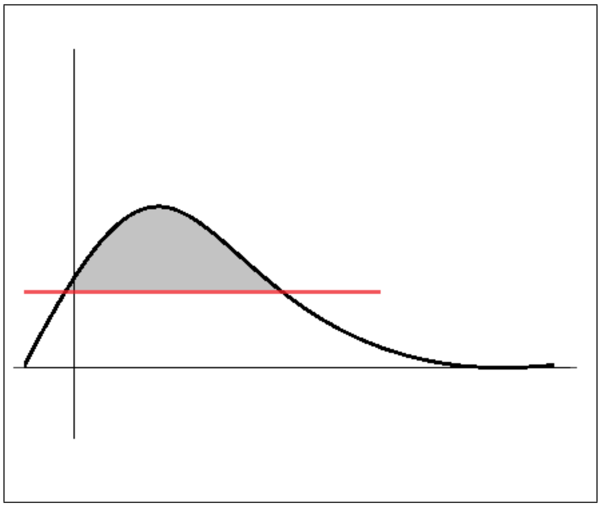
$$1 - \alpha \leq P(C|y) = \int_C p(\theta|y)d\theta. \tag{5.1}$$

Compare this to the definition for a confidence interval: if we could recompute C for a large number of datasets collected in the same way as ours, C would contain θ $(1-\alpha)\%$ of the time.

The difference between the two is that C often cannot be recollected (i.e. voter turnout in the 1960). The true value of θ is either in C or it is not. Hence a confidence interval is not a probability estimate, but rather just a signal as to the quality of the estimate. A credible set, however, is an actual probability statement.

A credible set is determined via a "horizontal line test." Intuitively, this is accomplished by lowering a horizontal line over a pdf (see figure 2) until 95% of the distribution is under the line. This differs from the frequentist approach of chopping off 2.5% of the distribution from each tail for a two-tailed test, or 5% from either tail for a one-tailed test. Observe that these approaches are only the same for symmetric distributions.

Figure 2: Horizontile line test



5.1 Hypothesis testing

TODO

Given two pdf's, $f(y|\theta)$ and $g(y|\theta)$, it is important to be able to determine whether or not the the distributions are statistically different.

Po-Boy's hypothesis test

Draw random samples from both distributions

Sort both draws, call them A_{sort} and B_{sort}

Difference draws, $C = A_{sort} - B_{sort}$

Calculate percentage of entries in C_{sort} above(below) H_0

6 Bayesian Inference

TODO

7 Bayesian Computation

7.1 Multiple parameters

Previously, we have only considered a single parameter θ. What if our likelihood function has several variables? We begin with an example with $n = 2$, and then proceed to the general case, known as *Gibbs sampling*.

7.1.1 Example

Suppose we have a normal likelihood function $f(\cdot)$ with mean μ and variance σ^2. Then $f(y|\mu, \sigma^2) \sim N(\mu, \sigma^2)$. How do we deal with two variables, μ and σ^2, especially when a particular prior might be appropriate for μ but not for σ^2 or vice versa? We proceed with two steps, one for each variable:

- Hold σ is known and estimate μ
- Hold μ is known and estimate σ^2

For the first of these parameters, μ, we use a normally distributed prior $\pi(\mu', \tau)$since $-\infty \leq \mu \leq \infty$. By exercise 3 we have the following result $p(\mu|y, \sigma^2) = N(\sigma|\frac{\sigma^2\mu'+\tau}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$.

For the second of these parameters, σ^2, a normal prior would not be optimal, since the normal distribution has a support of the entire real line, whereas $0 \leq \sigma^2$. Instead we use an inverse gamma distribution.

To simplify our approach, we begin with a gamma distribution, but use the measure for "precision" defined to be the inverse of the variance $\tau = \frac{1}{\sigma^2}$.

Using this transformation, our gamma prior becomes

$$\begin{aligned}
\pi(y|a,c) &= \frac{1}{\Gamma(a)b^a}\left(\frac{1}{y}\right)^{a-1} e^{\frac{1}{yb}} \\
&= \frac{1}{\Gamma(a)b^a}\left(\frac{1}{y}\right)^{a-1} e^{-1/by}|\left(-\frac{1}{y^2}\right)| \\
&= \frac{c^a}{\Gamma(a)}\,(y)^{-a-1}\,e^{-c/y} \qquad (7.1)
\end{aligned}$$

with $y = \sigma^2$ and $c = 1/b$. This gives us the following posterior:

$$\begin{aligned}
p(\sigma^2|x,\mu,a,c) &\propto \left((2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{\frac{s}{2\sigma^2}}\right)\left(\frac{c^a}{\Gamma(a)}\left(\sigma^2\right)^{-a-1}e^{-c/\sigma^2}\right) \\
&\propto (\sigma^2)^{-n/2}e^{-s/(2\sigma^2)}e^{-c/\sigma^2}(\sigma^2)^{-a-1} \\
&\propto (\sigma^2)^{-n/2-a-1}e^{-s/(2\sigma^2)-c/(\sigma^2)} \qquad (7.2)
\end{aligned}$$

where $s = \sum(x_i - \mu)^2$. Let $c' = (s/2 + c)$ and $a' = (n/2 + a)$. Then 7.2 is $\Gamma(a', c')$, which is equivalent to

$$p(\sigma^2|\mu,x,a,b) = \frac{(s/2+b)^{n/2+a}}{\Gamma(n/2+a)}\left(\sigma^2\right)^{n/2+a-1}e^{\frac{-(s/2+b)}{\sigma^2}} \qquad (7.3)$$

7.2 Gibbs Sampling

Suppose we have a distribution based on n parameters. Suppose also, that we do now know the closed form for the likelihood function, but that we are able to get the value for the pdf at any individual point. When direct sampling is difficult, we can use a Gibbs sampler to get a sequence of random samples that can help us find out the shape of the pdf.

Gibbs sampling works by constructing a Markov chain of draws from the likelihood function. We can then *thin* out the sample to get "random" draws by taking every 100^{th} or 1000^{th} observation. With additional iterations, the estimation improves. Hence the first portion of the estimation, called the "burn-in" phase, is often thrown out before we look at the distribution of the sample.

```
**Gibb's Sampling**
$\theta = (\theta_1, \theta_2, \ldots, \theta_k)$
Assign $\theta_1^0, \theta_2^0, \ldots, \theta_k^0$ arbitrary values
**for m loops:**
    Update $\theta_1$ from $[\theta_1 | \theta_2^{i-1}, \theta_3^{i-1}, \ldots, \theta_k^{i-1}]$
    Update $\theta_2$ from $[\theta_2 | \theta_1^{i}, \theta_3^{i-1}, \ldots, \theta_k^{i-1}]$
    Update $\theta_3$ from $[\theta_3 | \theta_1^{i}, \theta_2^{i}, \ldots, \theta_k^{i-1}]$
    $\vdots$
    Update $\theta_k$ from $[\theta_k | \theta_1^{i}, \theta_2^{i}, \ldots, \theta_{k-1}^{i}]$
```

7.3 Metropolis Algorithm

TODO

```
**Metropolis Algorithm**
Determine a starting value for $\theta$
**for m loops:**
    Sample $\theta^*$ from a symmetric proposal distribution: $J(\theta^* | \theta_{t-1})$
    Compute $r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$, where y is the data
    Update $\theta_t = \begin{cases} \theta^* & \text{with probability } r \\ \theta_{t-1} & \text{otherwise} \end{cases}$
```

Exercises

Exercise 1: Bayes' Theorem 1

A certain kind of tumor is benign about 85% of the time. The biopsy tests are not perfect, and err toward over-reporting the tumor as cancerous. A study just reported the following conditional probabilities based on the state of the tumor, cancerous or benign:

$$P(\text{test } +|cancerous) = .99 \text{ and } P(\text{test } +|benign) = .23$$

If you have a biopsy that tests positive for cancer, what is the probability that he actually does not have cancer?

Exercise 2: Bayes' Theorem 2

Consider the normal likelihood

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

In this case, σ is a known constant and θ is unknown. The prior distribution for θ is given as $\pi(\theta) \sim N(\mu, \tau^2)$, where μ and τ are known parameters. Verify that by plugging these expressions into equation (2.1) (see notes) we get the following posterior distribution:

$$p(\theta|y) = N\left(\theta|\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$$

Exercise 3: updating priors

Consider the situation in problem 3, but now suppose that we have a sample of n independent observations from $f(y|\theta)$. Since $S(y) = \bar{y}$ is sufficient for θ, we can say

$p(\theta|y) = p(\theta|\bar{y})$. Hence the posterior expression in the homework can be written as

$$\begin{aligned} p(\theta|\bar{y}) &= N\left(\theta|\frac{(\sigma^2/n)\mu + \tau^2\bar{y}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}\right) \\ &= N\left(\theta|\frac{\sigma^2\mu + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right) \end{aligned}$$

(a) Suppose $\mu = 2$, $\tau = 1$, $y = 6$, $\sigma = 1$, and $n = 1$ (the case in HW 3). Use Python to plot the prior, likelihood, and posterior distributions. Which distribution (prior or likelihood) dominates the posterior?

(b) What if n = 10?

Exercise 4: conjugate priors

You've been hired by Yellowstone to provide them with info relative to the number of eagles in the park. It is assumed that eagle sightings follow a poisson likelihood. Your daily eagle count for the first week is $(0, 1, 2, 0, 1, 0, 3)$.

(a)Use a gamma prior with shape of 0.3 and scale of 4. Report the following:

posterior shape and scale parameters

expected number of eagle sightings per day

95% "confidence interval" for number of eagle sightings per day

(b) The following week your daily eagle count is $(1, 3, 2, 2, 0, 1, 1)$. Use the posterior from part as as your prior and report the same numbers. Is this any different than using the prior from part (a) and using all the data at once?

Exercise 5: conjugate priors

You've been hired by a motorcycle shop to determine if their storewide sale days are effective. You have sales data for merchandise, bikes, and repairs (motorcycle.txt).

(a) Plot a histogram for each sales category (merch, bikes, repairs, total) for both sale and non-sale days.

(b) Choose a conjugate prior/likelihood pair for each category (do not distinguish between sale/non-sale days). Explain your choice.

(c) Update your prior based on the data and report the following:

-are revenues significantly higher during the store-wide sale days?

-probability that revenue for a given sale day will exceed \$50, \$100, \$1000

-plot the priors and posteriors for total revenue for sale and non-sale days

Exercise 6: Hyperpriors and Conditional Distributions

In some settings we may wish to model our uncertainty about the parameters of our prior distribution. We may do so by simply specifying a prior distribution for those parameters. We sometimes call this a hyperprior. This approach is called hierarchical modeling. Consider the Poisson/gamma model.

$$Y_i|\theta_i \sim Poisson(\theta_i t_i), \theta_i \sim G(\alpha, \beta), i = 1, ..., k$$

where t_i are known constants. Also assume α is known, but β is not. Use an inverse gamma hyperprior on β with known parameters c and d. Thus the density functions corresponding to the three stages of the model are

$$\begin{aligned} f(y_i|\theta_i) &= \frac{d^{-(\theta_i t_i)}(\theta_i t_i)^{y_i}}{y_i!}, y_i \geq 0, \theta_i > 0 \\ g(\theta_i|\beta) &= \frac{\theta_i^{\alpha-1} e^{-\theta_i/\beta}}{\Gamma(\alpha)\beta^\alpha}, \alpha > 0, \beta > 0 \\ h(\beta) &= \frac{e^{-1/(\beta d)}}{\Gamma(c) d^c \beta^{c+1}}, c > 0, d > 0 \end{aligned}$$

The main interest lies in finding the posterior distributions of the θ_i, $p(\theta_i|\mathbf{y})$. While the gamma prior is conjugate with the Poisson likelihood and the inverse gamma hyperprior is conjugate with the gamma prior, no closed form for $p(\theta_i|\mathbf{y})$ is available. We can, however, find the full conditional distributions of β and θ_i. Find these distributions $p(\theta_i|\theta_{j\neq i}, \beta, \mathbf{y})$ and $p(\beta|\{\theta_i\}, \mathbf{y})$.

Hint: Begin by verifying that each conditional distribution is proportional to $\left[\prod_{i=1}^{k} f(y_i|\theta_i)g(\theta_i|\beta)\right] h(\beta)$.

Exercise 7: Gibbs Sampling

Note that using the model and results of Exercise 6 we have functional forms that we can sample directly from for our conditional distributions. Write a Gibbs sampler for this model to find posterior information about β and the θ_i using the data from the table below and letting $c = 0.1$,$d = 1.0$, and $\alpha = 0.7$. . The data is on gas pump failures form Garver and O'Muircheartaigh (1987). Each observation is the number of pump failures observed (Y) in a given amount of time (t). The failure rate (r) is also reported for convenience. Because the length of time each pump was observed varied, we need to scale our results accordingly (thus the parameter in the Poisson distribution is $\theta_i t_i$ and not just θ_i). This ensures that our θ_i can be compared as rates.

i	Y_i	t_i	r_i
1	5	94.320	.053
2	1	15.720	.064
3	5	62.880	.080
4	14	125.760	.111
5	3	5.240	.573
6	19	31.440	.604
7	1	1.048	.954
8	1	1.048	.954
9	4	2.096	1.910
10	22	10.480	2.099

Report the means, standard deviation, 95 percent credible interval, and median for each parameter using a single chain run of 10,000 samples after a 1,000 sample burn-in. Comment on your results for the θ_i (the parameters of interest). What do you notice about the means and standard deviations for θ_5 and θ_6? Can you explain this?

Exercise 8: Metropolis-Hastings Algorithm

In this problem we will consider data from Bliss (1935) reported below. These data record the number of adult beetles killed after five hours of exposure to various levels of gaseous carbon disulphide (CS_2). We will use a generalized logit model suggested by Prentice (1976).

$$P(\text{death}|w) \equiv g(w) = \{\exp(x)/(1+\exp(x))\}^{m_1}$$

Here w is the predictor variable (dose) and $x = (w-\mu)/\sigma$ where $\mu \in \mathbb{R}$ and $\sigma^2, m_1 > 0$. Suppose there are y_i flour beetles dying out of n_i exposed at level w_i, $i = 1, ..., N$.

Dosage	# Killed	# Exposed
w_i	y_i	n_i
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

For the prior distributions, assume $m_1 \sim G(a_0, b_0)$, $\mu \sim N(c_0, d_0^2)$, and $\sigma^2 \sim IG(e_0, f_0)$, where a_0, b_0, c_0, d_0, e_0, and f_0 are known and m_1, μ, and σ^2 are independent. You may be tempted to think that these families of distributions have been chosen to preserve some sort of conjugate structure, but this is not the case for the joint posterior distribution or any of the conditional distributions. Thus we must resort to the Metropolis-Hastings algorithm.

Verify that

$$p(\mu, \sigma^2, m_1|y) \propto \left\{\prod_{i=1}^{N} [g(w_i)]^{y_i} [1-g(w_i)]^{n_i-y_i}\right\} \frac{m_1^{a_0-1}}{\sigma^{2(e_0+1)}}$$
$$\times \exp\left[-\frac{1}{2}\left(\frac{\mu-c_0}{d_0}\right)^2 - \frac{m_1}{b_0} - \frac{1}{f_0\sigma^2}\right]$$

Next make a convenient change of variables to $\theta = (\theta_1, \theta_2, \theta_3) = (\mu, \frac{1}{2} \log sigma^2, \log m_1)$ so that our parameter space is $\mathbb{R}^3$ (don't forget the Jacobian). Why is this necessary and helpful?

Verify that

$$p(\theta|y) \propto h(\theta) = \left\{ \prod_{i=1}^{N} [g(w_i)]^{y_i} [1 - g(w_i)]^{n_i - y_i} \right\} \exp(a_0\theta_3 - 2e_0\theta_2)$$

$$\times \exp\left[-\frac{1}{2} \left(\frac{\theta_1 - c_0}{d_0} \right)^2 - \frac{\exp(\theta_3)}{b_0} - \frac{\exp(-2\theta_2)}{f_0} \right]$$

Suppose we let $a_0 = .25$ and $b_0 = 4$ so that m_1 has a prior mean of 1 and prior standard deviation 2. Also, let $c_0 = 2$, $d_0 = 10$, $e_0 = 2.000004$, and $f_0 = 1000$ so that our priors for μ and σ^2 are rather vague. The latter two choices imply a prior mean of .001 and prior standard deviation of .5 for σ^2.

Write a Metropolis-Hastings sampler using a jumping distribution of $N_3(\theta^{(t-1)}, \tilde{\Sigma})$ where $\tilde{\Sigma} = D = Diag(.00012, 0.033, .10)$. (Tip: Calculating r is more numerically stable if you compute it with $r = \exp[\log h(\theta^*) - \log h(\theta^{(t-1)})]$. Use three parallel chains of 10000 draws with initial conditions of $(1.8, -3.9, -1)$, $(1.5, -5.5, -2.5)$, and $(2.1, -2.3, 1.5)$.

Generate monitoring plots for μ, $\log \sigma$, and $\log m_1$. (Monitoring plots are plots of the draws for each parameter with the parameter values on the vertical axis and the draw number on the horizontal axis.) Use the monitoring plots to choose a reasonable burn-in length. Use the remaining draws from one chain to estimate the lag 1 sample autocorrelations. Use the post burn-in draws from all three chains to generate histograms for each parameter and report the .025, .5, and .975 quantiles. Also report the overall acceptance rate.

What do you notice about the acceptance rate? The autocorrelations? What might be the source of these results? (Hint: You may want to compute the sample correlations between parameters to help inform your reasoning.)

Exercise 9: Metropolis-Hastings Algorithm

Consider one possible improvement we could make in generating our variance-covariance matrix in Exercise 8. Use the output of the first algorithm to estimate the posterior covariance matrix ($\hat{\Sigma} = \frac{1}{G}\sum_{g=1}^{G}(\theta_g - \bar{\theta})(\theta_g - \bar{\theta})'$ where $g = 1, ..., G$ indexes the post-convergence Monte Carlo samples). Then let $\tilde{\Sigma} = 2\hat{\Sigma}$. This reflects "folklore" that says two times the posterior covariance should perform well.

Report the same output as in Exercise 8 (monitoring plots, burn-in length, lag 1 autocorrelations, acceptance rate, histograms, and quantiles) for the new covariance matrix in your jumping distribution. Comment on differences between the two algorithms.

Exercise 10: Independence Chains

Reanalyze the flour beetle mortality data and model from Exercise 8, replacing the multivariate Metropolis algorithm with a Hastings algorithm employing independence chains drawn from a $N(\tilde{\theta}, \tilde{\Sigma})$ candidate density where $\tilde{\theta} = (1.8, -4.0, -1.0)'$ (roughly the true posterior mode) and $\tilde{\Sigma}$ as estimated in 9.

Exercise 11: Metropolis within Gibbs

As in Exercise 10, reanalyze the flour beetle mortality data and model, this time using a univariate Metropolis (Metropolis within Gibbs) algorithm using proposal densities $N(\theta_i^{(t-1)}, D_{ii}), i = 1, 2, 3$ with D as given in Exercise 8. The Metropolis within Gibbs algorithm is just the Gibbs sampler but employing the Metropolis-Hastings algorithm to sample from the conditional distributions.

Exercise 12: Metropolis-Hastings Algorithm

Notice that the decision to use 2 in Exercise 9 was rather arbitrary. The optimal amount of variance inflation might well depend on the dimension and precise nature of the target distribution, the type of sampler used (multivariate versus univariate, Metropolis versus Hastings (independent chains), etc.) and any number of other factors.

Explore these issues in the context of the flour beetle mortality data by resetting $\tilde{\Sigma} = c\hat{\Sigma}$ for $c = 1$ (candidate variance matched to the target) and $c = 4$ using the multivariate Metropolis algorithm (Exercise 9), the multivariate Hastings algorithm (Exercise 10), and the univariate Metropolis algorithm (Exercise 11).

References

Carlin, B., and T. Louis (2009): *Bayesian Methods for Data Analysis.* Chapman and Hall/CRC Press.