Session 1

Bayesian Methods

Iraj Kazemi

i.kazemi@lancaster.ac.uk

Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, England.

March 10-11, 2005

- This course introduces students to the use of Bayesian methods for data analysis in the social sciences and related disciplines.
- This also provides the basic concepts of the Bayesian approach to statistics such as:

☐ the subjective interpretation of probability,

☐ types of prior distributions,

☐ the use of Bayes theorem in updating information and inference procedures such as Bayesian estimates.

These will include incorporating classical likelihood within the Bayesian framework, and

- fitting linear regression models
- generalized linear models including
- the binary response model, and
- the Poisson model for counts.

Finally more advanced topics such as

- Hierarchical models,
- Markov chain Monte Carlo, and
- Gibbs sampling

★ The main focus of the course will be the application of Bayesian models by using a various type of real examples from social sciences and the other disciplines.

♦ All the models fitted in this course use WinBUGS, a Bayesian MCMC package,

- This is distributed freely from the web site of the Medical research Council Biostatistics Research Unit in Cambridge (http://www.mrc-bsu.cam.ac.uk/bugs/).

Prerequisites

It is assumed that each participant has

- a basic knowledge about probability theory,
- continuous and discrete probability distributions,
- MLEs, linear regression, and
- generalized linear models.

The level of statistics required is not extremely high, but

- a basic background in deriving likelihood functions is necessary for an understanding of the prior and posterior distributions, which are most common in Bayesian methods.

★ We will spend time in the first session to cover the necessary background for the properties of distribution functions, though

★ it is assumed that everyone is familiar with these topics.

- There will be many practical sessions with the main focus on applying Bayesian models to real data analysis.

References

- Carlin, B. and Louis, T.A. (1996). Bayes and Empirical Bayes Methods for Data Analysis, Chapman and Hall.
- Congdon, P. (2001). Bayesian Statistical Modelling, John Wiley & Sons, New York.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). Bayesian Data Analysis, 2nd ed., Chapman & Hall.
- Gill, J. (2002). Bayesian Methods: A Social and Behavioral Sciences Approach, Chapman & Hall/CRC.
- Lancaster, T. (2004). An Introduction to Modern Bayesian Econometrics, Blackwell Publishing.
- Lee, P.M. (2004). Bayesian Statistics: An Introduction, 3rd edition, Arnold, London.
- Spiegelhalter, D. J. S., Abrams, K. R., and Myles, J. P. (2004). Bayesian Approaches to Clinical Trials and Health-Care Evaluation, John Wiley & Sons, Ltd.

Introduction

- Classical statistics provides methods to analyze data, from simple descriptive measures to complex models.
- The available data are processed and then conclusions about a hypothetical population are drawn.
- However, data are not the only available source of information about the population.
- Bayesian methods provide a principled way to incorporate the external information into the data analysis process.

- In a Bayesian approach, the data analysis process starts already with a given probability distribution.
- As this distribution is given before any data is considered, it is called *prior* distribution.
- The Bayesian data analysis process consists of using the sample data to **update** this prior distribution into a *posterior* distribution.
- The basic tool for this updating is a theorem, proved by Thomas Bayes.

Conditional Probability

- The key concept for thinking about conditional probabilities is that the occurrence of B reshapes the sample space for subsequent events.
- When A and B are two events defined on a sample space S, the conditional probability of A given that B looks just at the subset of the sample space for B.
- For two events A and B,

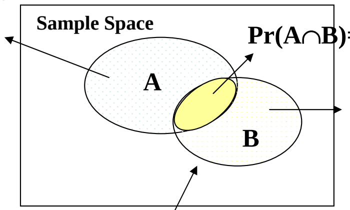
$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

- $Pr(A \cap B)$: the probability that both A and B occur and
- $Pr(A|B)$: the probability that A occurs given the knowledge that B has occurred.

The Venn Diagram

Pr(A-B)=The probability that A occur and B does not.

Pr(A∩B)=Pr(Both A and B occur)



Pr(B-A)=The probability that B occur and A does not.

- A-B and B-A are two independent events; i.e., the intersection of them is an empty set. So, for example,

P(B)=P(B-A)+P(**B** ∩A)
= P(B ∩A')+P(**B** ∩A)

- **Multiplication rule:** we also have $Pr(A \cap B) = P(A)P(B|A)$, then

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

- When written in this form the definition is called Bayes' Theorem.
- $P(A)$ is the probability assigned to the truth of A before the data have been seen and
- $P(A|B)$ is its probability after the evidence is in
- When thought of in this way we call $P(A)$ the *prior* probability of A and
- $P(A|B)$ the *posterior* probability of A.
- Bayes theorem can then be interpreted as showing how $P(A)$ is changed by the evidence into $P(A|B)$.

Who's Bayes?

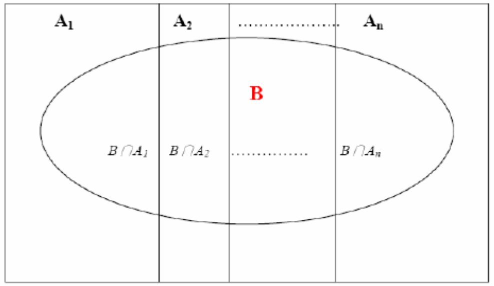
Born: 1702 in London, England

Died: 1761, Turnbridge, Kent, England

Elected a fellow of the Royal Society in 1742 (at that time he had no published work!)

Conditional Probability and Partitions of a Sample Space

- The set of events $A_1, \cdots, A_k$ form a partition of a sample space S if $\bigcup_{i=1}^{k} A_i = S$.
- If the events $A_1, \cdots, A_k$ partition S and if B is any other event in S, then
- the events $A_1 \cap B, A_2 \cap B, \cdots, A_k \cap B$ will form a partition of B.



- Thus,

$$B = \bigcup_{i=1}^{k} (A_i \cap B)$$

and

$$Pr(B) = \sum_{i=1}^{k} Pr(A_i \cap B)$$

- Finally, if $Pr(A_i) > 0$ for all i, then

$$Pr(B) = \sum_{i=1}^{k} Pr(A_i)Pr(B|A_i)$$

Total probability rule (average rule)

Bayes' Theorem: for $i = 1, .., k$,

$$Pr(A_i|B) = \frac{Pr(A_i)Pr(B|A_i)}{\sum_{i=1}^{k} Pr(A_i)Pr(B|A_i)}$$

- This shows how to update the prior probability $Pr(A_i)$ to the posterior probability $Pr(A_i|B)$, after the event (i.e., data) A has been observed.

Interpretation of Bayes' Theorem

$\Pr(A_i)$ = Prior distribution for the A_i. It summarizes your beliefs about the probability of event A_i before A_i or B are observed.

$\Pr(B \mid A_i)$ = The conditional probability of B given A_i. It summarizes the *likelihood* of event B given A_i.

$$\Pr(A_i \mid B) = \frac{\Pr(A_i)\ \Pr(B \mid A_i)}{\sum_k \Pr(A_k)\Pr(B|A_k)}$$

$\Pr(A_i \mid B)$ = The posterior distribution of A_i given B. It represents the probability of event A_i after B has been observed.

$\Sigma_k \Pr(A_k)\Pr(B \mid A_k)$ = The normalizing constant. This is equal to the sum of the quantities in the numerator for all events A_k. Thus, $P(A_i \mid B)$ represents the likelihood of event A_i relative to all other elements of the partition of the sample space.

Example

In the United Kingdom in 1975, a referendum was to be held as to whether the UK should stay part of the EC.

- The proportion supporting Labour (L) was 52%; $Pr(L) = 0.52$
- the proportion supporting the Conservatives (C) was 48%; $Pr(C) = 0.48$
- Many polls indicated that 55% of L supporters and 85% of C supporters intended to vote "Yes" (Y) in the EC referendum

$$Pr(Y|L) = 0.55 \text{ and } Pr(Y|C) = 0.85$$

- the remainder intended to vote "No" (N).

- Suppose we meet someone and she says that she intends to vote "Yes" in the referendum. What should we conclude about her partisan support?

$$\begin{aligned} Pr(L|Y) &= \\ &= \frac{Pr(Y|L)P(L)}{Pr(Y|L)Pr(L) + Pr(Y|C)Pr(C)} \\ &= \frac{0.55 \times 0.52}{(0.55 \times 0.52) + (0.85 \times 0.48)} \\ &= 41.2\% \end{aligned}$$

Example

- A patient is given a blood test to determine whether she has a certain disease.
- Suppose the test returns a positive result 95% of the time when a patient has the disease and Pr(B|A)=0.95

Pr(B|A')=0.01

- 1% of the time when a patient does not have the disease.
- Moreover, suppose 0.1% of the population is known to have the disease. Pr(A)=0.001 =>Pr(A)=0.999
- What is the probability that a person who tests positive for the disease actually has the disease? Pr(A|B) = ?

- Let B be the event the patient tests positive and A be the event that the patient has the disease. Then

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)} \\ &= \frac{(0.001)(0.95)}{(0.001)(0.95) + (0.999)(0.01)} \\ &= 0.08684. \end{aligned}$$

- Note that although the test is 95% effective in correctly identifying a person who has the disease if they have the disease, nevertheless, given a positive result, there is only a 9% chance that the person actually has the disease.

Bayes' Theorem and Statistics

- The foundation of Bayesian statistics is Bayes' theorem.
- Bayes' theorem can be used to revise probability distributions for a parameter of a statistical model.
- If the prior distribution of a parameter θ, with n possible outcomes $(\theta_1, \cdots, \theta_k)$, is discrete and
- the new information x comes from a discrete model,
- then

$$Pr(\theta_i|x) = \frac{Pr(\theta_i)Pr(x|\theta_i)}{\sum_{i=1}^{k} Pr(\theta_i)Pr(x|\theta_i)},$$

- $Pr(\theta)$: the prior distribution of the possible θ values
- $Pr(\theta_i|x)$: the posterior distribution of θ given the observed data x.

Example

- Suppose **one** in every **1000** families has a genetic disorder (sex-bias) in which they produce only female offspring.
- Define the random variable

$$\theta = \begin{cases} 0 & \text{normal family} \\ 1 & \text{sex-bias family} \end{cases}$$

- Suppose we observe a family with 5 girls and *no* boys. What is the probability that this family is a sex-bias family?
- From prior information, there is a $1/1000$ chance that any randomly-chosen family is a sex-bias family, so $Pr(\theta = 1) = 0.001$.
- If x = five girls, then Pr(five girls | sex-bias family) $= 1$. This is $Pr(x|\theta = 1)$.

- We need to compute $Pr(x)$; the probability that a family with five children has all girls.

$$\begin{aligned} Pr(x) &= Pr(x|normal) \cdot Pr(normal) + \\ &\quad Pr(x|sex\text{-}bias) \cdot Pr(sex\text{-}bias), \end{aligned}$$

giving

$$Pr(x) = (\frac{1}{2})^5(\frac{999}{1000}) + 1 \cdot (\frac{1}{1000}) = 0.0322$$

- Hence,

$$\begin{aligned} Pr(\theta = 1|x) &= \frac{Pr(x|\theta = 1)Pr(\theta = 1)}{Pr(x)} \\ &= \frac{1 \cdot 0.001}{0.0322} = 0.032 \end{aligned}$$

- Thus, a family with five girls is 32 times more likely than a random family to have the sex-bias disorder.

- We also have

$$Pr(\theta = 0|x) = 1 - 0.032 = 0.968.$$

θ	$Pr(\theta)$	$Pr(x\|\theta)$	$Pr(\theta\|x)$
0	0.999	0	0.968
1	0.001	1	0.032

- It is, however, highly unrealistic to assume that there are only a few possible values of the parameter of interest, θ.
- The possibility of treating θ as a continuous random variable should be allowed.
- Specifically, the more usual form of the theorem is in terms of random variables.

Bayes' Theorem and the Likelihood Function

- Let the observation vector $\mathbf{x} = (x_1, \cdots, x_n)'$ denote the numerical realization of a random vector $\mathbf{X} = (X_1, \cdots, X_n)'$,
- Denote the density or probability mass function with $f(\mathbf{x}|\theta)$, for $\mathbf{x} \in S$ and $\theta \in \Theta$.
- Before observing $\mathbf{x}$, most scientists will possess some prior information about θ, such as previous data sets, knowledge, or subjective scientific advice, which might suggest plausible values for θ.

- If the *prior distribution* of θ is represented by a density function $\pi(\theta)$,
- then the *posterior distribution* is also represented by a density function $\pi(\theta|\mathbf{x})$.
- The Bayes's theorem

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x})}$$

- We note that the joint probability distribution of the data and the parameter is given by $f(\mathbf{x}, \theta)$. The probability

$$f(\mathbf{x}|\theta) = L(\theta) = \prod_i f(x_i|\theta)$$

  is the likelihood.
- It gives your predictions as to what the data should look like if the parameter takes the particular value given by θ.
- The prior distribution, $\pi(\theta)$, gives your beliefs about the possible values of θ.
- Both likelihood and prior distribution are required in order to reach probabilistic conclusions about the consistency of the model with the evidence.

- The function $f(\mathbf{x})$ is called the *marginal distribution* of data.
- For most inference problems, $f(\mathbf{x})$ does no have a closed form.
- Since $f(\mathbf{x})$ depends only on the $\mathbf{x}$ (and not on θ) and our concern is the distribution over θ, we may write

$$\begin{aligned} \pi(\theta|x) &= \frac{1}{f(\mathbf{x})}\pi(\theta)L(\theta) \\ &= (\text{normalizing constant})\, \pi(\theta)L(\theta) \\ &= \mathit{constant} \cdot \mathit{likelihood} \cdot \mathit{prior} \end{aligned}$$

  Because of this, the posterior distribution is often written as

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

  where the symbol $\propto$ means "is proportional to".

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

$$Posteriordistribution \propto Priordistribution \times Likelihood$$

- Taking logs on $\pi(\theta|\mathbf{x})$ (and ignoring the normalizing constant) gives

$$log(posterior) = log(likelihood) + log(prior).$$

$\Longrightarrow$ the posterior density summarizes the total information, after viewing the data, and provides a basis for posterior inference regarding θ.

Definition: A Kernel

- A probability density function of a r.v. X typically has the form $cg(x)$.
- The purpose of c is to make the density function integrate to one. .
- The remaining portion, $g(x)$, which does involve x, is called the kernel of the function.

Example: The gamma distribution

- Suppose that X is Gamma distributed:

$$f(x|\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \qquad x > 0; \alpha, \lambda > 0$$

the kernel is

$$x^{\alpha-1}e^{-\lambda x}$$

where c is a function of α and λ.

The Gamma Distribution

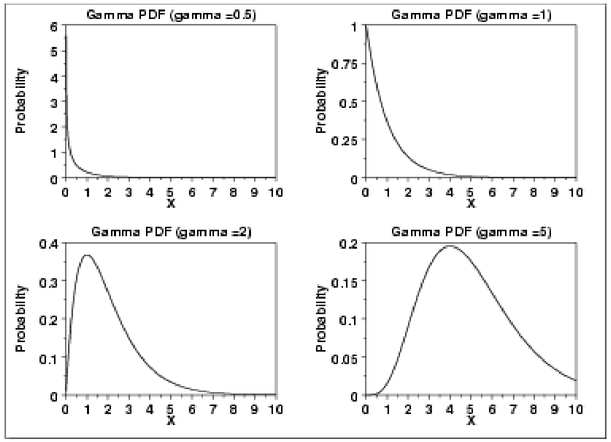
- The general formula for the probability density function of the gamma distribution is

$$f(x|\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, \qquad x > 0;\ \alpha, \lambda > 0$$

$$E(X) = \frac{\alpha}{\lambda}$$

$$var(X) = \frac{\alpha}{\lambda^2}$$

- λ is the shape parameter
- for λ=*n*/2, α=1/2 we have the Chi-square distribution.



- **Example:** $X \sim N(\mu, \sigma^2)$. If μ is of interest, the kernel is

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

where the constant is $1/\sqrt{2\pi\sigma^2}$.

- If μ and σ^2 are of interest then the kernel would be

$$\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

Remark

- In deriving posterior densities it is usually convenient to omit constants
- It makes for algebra that is much easier to follow.
- This avoids a direct computation of $f(\mathbf{x})$.
- If the kernel cannot be recognized, then $f(\mathbf{x})$ must be computed directly.

Example: Exponential Distribution

- Suppose that X has the exponential density

$$f(x|\theta) = \theta e^{-\theta x}, \quad x > 0; \theta > 0.$$

- Assume that the prior distribution of θ is given by

$$\pi(\theta) = 1, \quad 0 < \theta < 1$$

- The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta s}, \qquad \text{where } s = \sum x_i.$$

The posterior distribution of θ is

$$\pi(\theta|\mathbf{x}) \propto \theta^n e^{-\theta s}$$

$\theta|\mathbf{x} \sim Gamma\,(n+1, s)$.

Example: Exponential Distribution (cont.)

- If the prior distribution of θ is given by

$$\pi(\theta) = e^{-\theta}, \quad \theta > 0$$

The posterior distribution of θ is

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto e^{-\theta} \cdot \theta^n e^{-\theta \sum x_i} \\ &= \theta^n e^{-\theta(\sum x_i + 1)} \end{aligned}$$

$\theta|\mathbf{x} \sim Gamma\,(n+1, s+1)$.

- For example, if we observe s=15 for a sample size of 10, then

θ | **x** ~ *Gamma*(11 , 16)

Example: Poisson Distribution

- Suppose that X is Poisson distributed with mean θ:

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

- The likelihood function is

$$L(\theta) \propto \theta^s e^{-n\theta}, \qquad \text{where } s = \sum_i x_i.$$

- Assume that the prior distribution of θ is

$$\pi(\theta) = e^{-\theta}, \ \theta > 0$$

The posterior density is

$$\pi(\theta|x) \propto \theta^s e^{-(n+1)\theta}, \quad \theta > 0$$

Therefore, the posterior distribution is $Gamma(s+1, n+1)$.

Example: Poisson Distribution (cont.)

- Assume that the prior distribution of θ is Gamma with known hyper-parameters a and b:

$$\pi(\theta) \propto \theta^{a-1} e^{-b\theta}, \ \theta > 0$$

- The term hyper-parameter is used to distinguish a and b from the parameter of the sampling model.

- Then the posterior density is

$$\pi(\theta|x) \propto \theta^{x+a-1} e^{-(1+b)\theta}, \quad \theta > 0$$

$\Longrightarrow$ the posterior distribution is $Gamma(x + a, 1 + b)$

Example: Poisson Distribution (cont.)

- The marginal distribution of X follows a negative binomial distribution: $X \sim Negbin(a, b)$

$$P(X = x) = \binom{a + x + 1}{x} \left(\frac{b}{b+1}\right)^a \left(\frac{1}{b+1}\right)^x, \quad x = 0, 1, 2, \cdots$$

$\Longrightarrow$ $Negbin(\alpha, \beta)$ = mixture of Poisson distributions with rates θ, that follow a $Gamma(\alpha, \beta)$ distribution.

Bernoulli Trials.

- Suppose the variable of interest, X, is binary and takes the values zero or one.

- The probability mass function of X

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1; 0 < \theta < 1.$$

- The likelihood function is

$$L(\theta) = \theta^s (1-\theta)^{n-s}.$$

where $s = \sum_{i=1}^n x_i$ is the total number of successes (ones) in the n trials.

- The parameter θ is constrained to be in the interval $(0,1)$. A class of possible priors is of the form

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1$$

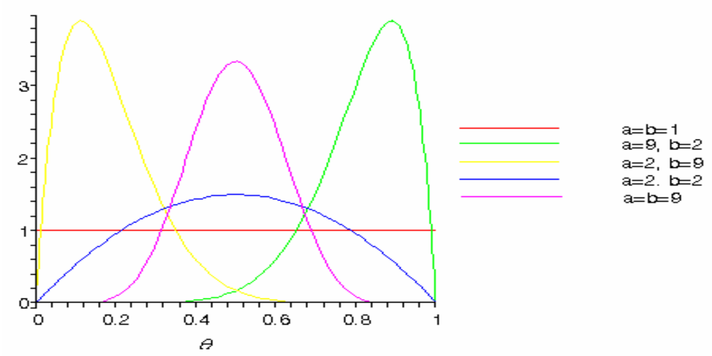
- The posterior density is readily found to be

$$\begin{aligned}\pi(\theta|x) &\propto \theta^{a-1}(1-\theta)^{b-1} \times \theta^{s}(1-\theta)^{n-s} \\ &= \theta^{a+s-1}(1-\theta)^{n+b-s-1}, \quad 0 < \theta < 1\end{aligned}$$

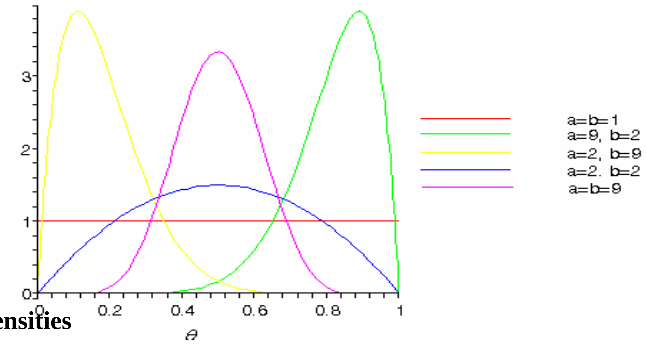
$\Longrightarrow$ a Beta distribution with hyper-parameters $a+s$ and $n+b-s$.

Priors for Various Choices of *a* and *b*

- A choice *a*=*b*=1 yields a prior distribution which is uniform in (0,1), so that all values are equally likely.
- For a choice of *a*<*b*, large values of θ are more likely.
- Choosing *a*=*b* implies that the prior distribution is symmetrical about the prior mean and mode that are both equal to 0.5.
- Several examples are given in Figure below with *n*=10, and *s*=6.

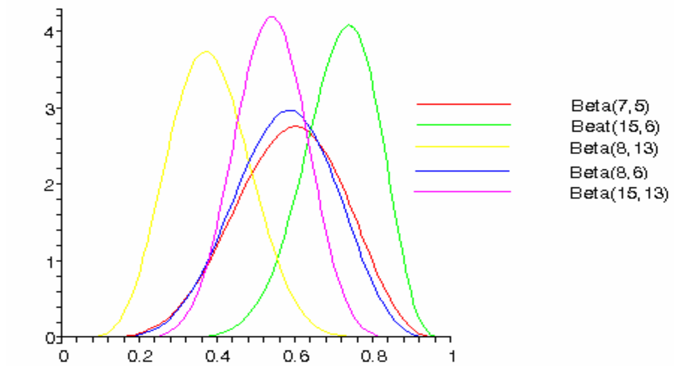


- Prior densities
- For instance, when the prior parameters are *a* = 2 and *b* = 2 (or a=b=9), the prior distribution assigns smaller probability to values of larger and smaller than 0.5.



- **Corresponding posterior densities**

posterior densities are very different from prior densities but have a similar shape.



o

End of Session