Session 3

Bayesian Methods

Iraj Kazemi

i.kazemi@lancaster.ac.uk

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

March 10-11, 2005

The Choice of a Prior

- A critical feature of any Bayesian analysis is the choice of a prior.
- We can check the impact of the prior by seeing how stable to posterior distribution is to different choices of priors.
- If the posterior is highly dependent on the prior, then the data (the likelihood function) may not contain sufficient information.
- However, if the posterior is relatively stable over a choice of priors, then the data indeed contain significant information.

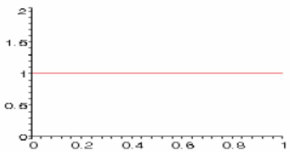
Prior Distributions

- There are various types of prior distributions that we need to discuss.

Noninformative Priors

- A Prior distribution is noninformative if the prior is flat relative to the likelihood function. That is, the prior is simply a constant,

$$\pi(\theta) = c = \frac{1}{b-a} \quad \text{for } a < \theta < b$$



- Thus a prior $\pi(\theta)$ is noninformative if it has minimal impact on the posterior distribution of θ.
- Other names for noninformative prior are *reference prior*, *vague prior*, or *flat prior*.

- With a noninformative prior, the posterior just a constant times the likelihood,

$$\pi(\theta|x) \propto cons \cdot L(\theta|x)$$

and we typically write that $\pi(\theta|x) \propto L(\theta|x)$.

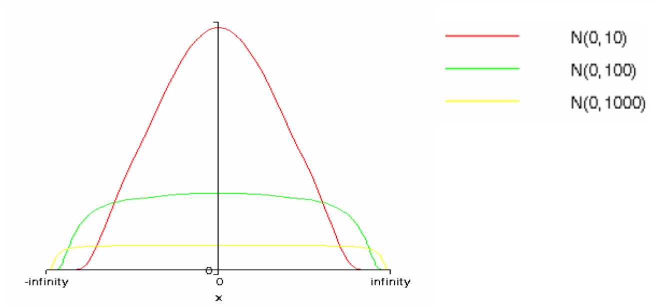
- In many cases, classical expressions from frequentist statistics are obtained by Bayesian analysis assuming a vague prior.
- Thus, for a flat prior the **posterior mode = MLE**.

Examples of noninformative priors

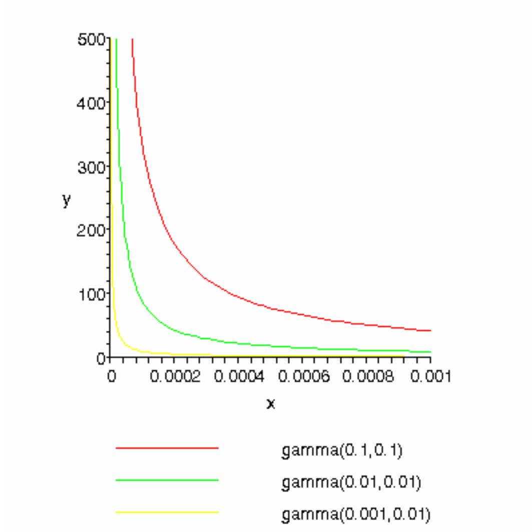
(1) If $0 < \theta < 1$, then $\theta \sim U(0,1)$ is a noninformative prior for θ; i.e., $\pi(\theta) = 1,\ 0 < \theta < 1$.

(2) If $-\infty < \theta < \infty$, then if $\theta \sim N(\mu_0, \sigma_0^2)$, and $\sigma_0^2 \to \infty$, then we get a noninformative prior.

- In WinBUGS, the flat prior can be approximated by a vague normal density prior, with mean 0 and variance 1000000; for example.



- The inverse prior $\pi(\sigma^2) = 1/\sigma^2$, can be approximated by a Gamma density (with a very small shape and rate parameters.)



Improper Priors

- If the variable of interest ranges over $(0, \infty)$ or $(-\infty, +\infty)$, then strictly speaking a flat prior does not exist,
- i.e, the integral does not exist.
- In such cases a flat prior (assuming $\pi(\theta|x) \propto L(\theta|x)$) is referred to as an improper prior.
- A prior $\pi(\theta)$ is said to be *improper* if

$$\int_{\Theta} \pi(\theta) d\theta = \infty.$$

- Thus a prior is improper if its normalizing constant is equal to ∞.
- Improper priors are often used in Bayesian inference since they usually yield noninformative priors.

Example

- Suppose that for $-\infty < \theta < \infty$, the prior $\pi(\theta) \propto 1$.
- That is θ has a uniform prior distribution on the real line. Clearly,

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta = \int_{-\infty}^{\infty} d\theta = \infty.$$

Remarks:

(1) An improper prior may result in an improper posterior. We cannot make inference with improper posterior distributions.

(2) An improper prior may still lead to a proper posterior distribution.

Example

- Consider a random sample $X_1, X_2, \cdots, X_n$ drawn from a $N(\mu, 1)$ distribution.
- Suppose $\pi(\mu) \propto 1$, then $\mu|x \sim N\left(\overline{x}, \frac{1}{n}\right)$.
- Therefore, the uniform improper prior on μ still leads to a **proper posterior**.
- i.e., a normal distribution with mean $\overline{x}$ and variance $\frac{1}{n}$.

Example

- Consider random sample $X_1, X_2, \cdots, X_n$ that conditional on θ is distributed as $Poisson(\theta)$.
- Suppose $\pi(\theta) \propto \theta^{-\frac{1}{2}}$. Here $\Theta = \{\theta : 0 < \theta < \infty\}$, and therefore

$$\int_0^\infty \pi(\theta) d\theta = \int_0^\infty \theta^{-\frac{1}{2}} d\theta = \infty.$$

- Thus $\pi(\theta)$ is improper for θ.

$$\begin{aligned} g(\theta|x) &\propto \left(\theta^{\sum x_i} e^{-n\theta}\right)\left(\theta^{-\frac{1}{2}}\right) \\ &= \theta^{\sum x_i + \frac{1}{2} - 1} e^{-n\theta} \end{aligned}$$

- Thus $\theta | x_1, \cdots, x_n \sim gamma\left(\sum x_i + \frac{1}{2}, n\right)$.
- The posterior of θ is **proper** and is a gamma distribution.

- In most cases, improper priors can be used in Bayesian analysis without major problems. However,

(1) In a few models, the use of improper priors can result in improper posteriors.

(2) Use the improper priors makes model selection and hypothesis testing difficult.

(3) WinBUGS does not allow the use of improper priors.

Informative Priors

- An informative prior is a prior not dominated by the likelihood, and has an impact on the posterior distribution.
- Informative priors must be specified with care in actual practice.
- They are useful to use if we have real prior information from a previous similar study, for example.

Example: Normal distribution

- Given θ, suppose $X_1, X_2, \cdots, X_{10}$ are i.i.d. $N(\theta, 10)$.
- Suppose $\theta \sim N(\theta_0, 1)$.
- Then this represents an informative prior for θ. We have

$$L(\theta|\mathbf{x}) \propto exp\left\{-\frac{1}{2}(\theta - \overline{x})^2\right\}$$
$$\pi(\theta) \propto exp\left\{-\frac{1}{2}(\theta - \theta_0)^2\right\}$$

Example: Binomial distribution

- Given θ, suppose $X_1, X_2, \cdots, X_{10}$ are i.i.d. $Bin(1, \theta)$.
- Suppose $\sum x_i = 5$, so that

$$\begin{aligned} f(x|\theta) &\propto \theta^{\sum x_i - 1}(1-\theta)^{n-\sum x_i} \\ &= \theta^5 (1-\theta)^5 \\ &= \theta^{6-1}(1-\theta)^{6-1} \end{aligned}$$

- If $\theta \sim beta(5, 5)$, then this would be an informative prior for θ.
- In this case, $\pi(\theta) \propto \theta^{5-1}(1-\theta)^{5-1}$.

Conjugate Priors

- A prior is said to be a *conjugate* prior for a family of distributions if the prior and posterior distributions are of the same family.

Example: Poisson Distribution

- Suppose that X is Poisson distributed with mean θ.
- Assume that the prior distribution of θ is $Gamma(a, b)$:

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

- 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

- Consider the density of X_i, $i = 1, \cdots, n$ conditional on θ is

$$f(x_i|\theta) = \theta(1-\theta)^{x_i} \quad i = 1, 2, \cdots$$

- Assume the prior is $Beta(a, b)$

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

- The posterior distribution is

$$\begin{aligned} g(\theta|x) &\propto \prod_i f(x_i|\theta) \times \pi(\theta) \\ &= \theta^{n+a-1}(1-\theta)^{\sum x_i + b - 1} \end{aligned}$$

- This is the density of a $Beta(n + a, \sum x_i + b)$.

Conjugate Priors for the Exponential Family of Distributions

- Many common distributions (normal, gamma, Poisson, binomial,, etc.) are members of the exponential family, whose general form is given by

$$\begin{aligned} f(x|\theta) &= a(\theta)\, b(x)\, e^{\mathbf{c}(\boldsymbol{\theta})'\mathbf{d}(x)} \\ &= a(\theta)\, b(x)\, e^{\sum_j c_j(\boldsymbol{\theta}) d_j(x)} \end{aligned}$$

for a suitable choice of functions a, b, c, and d.

Example: Bernoulli Distribution

Suppose a binary variable X has a probability mass function

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

then

$$f(x|\theta) = (1-\theta)\, e^{x \log\left(\frac{\theta}{1-\theta}\right)}$$

We can take $a(\theta) = 1-\theta, b(x) = 1, c(\boldsymbol{\theta}) = \log(\theta/(1-\theta))$, and $d(x) = x$; so $f(x|\theta)$ belongs to the exponential family.

Example: Exponential distribution

If X is a exponential variable with a probability density function

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

then $f(x|\theta)$ belongs to the exponential family for $a(\theta) = \theta, b(x) = 1, c(\boldsymbol{\theta}) = -\theta$, and $d(x) = x$.

Example: Normal distribution with unknown mean and variance

Let $f(x|\mu, \sigma^2)$ be the $N(\mu, \sigma^2)$ family of pdf's. Then $\theta = (\mu, \sigma^2)$ where $-\infty < \mu < \infty$ and $\sigma > 0$. This family is a 2-parameter exponential family

$$\begin{aligned} f(x|\mu, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \\ &= \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}\mu^2} \cdot \left(\frac{1}{2\pi}\right)^{1/2} \cdot e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x} \end{aligned}$$

We can take $a(\theta) = \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}\mu^2}, b(x) = (1/2\pi)^{1/2}, c_1(\boldsymbol{\theta}) = -1/(2\sigma^2), d_1(\boldsymbol{\theta}) = x^2, c_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}$, and $d_2(x) = x$; so $f(x|\theta)$ belongs to the exponential family.

When the density or the probability mass function is in the form of an exponential family, a conjugate prior can be found. Suppose we consider a prior on θ of the form

$$\pi(\boldsymbol{\theta}) \propto [a(\boldsymbol{\theta})]^b \, e^{\sum_j c_j(\boldsymbol{\theta}) d_j}$$

where b and d_j are specified hyper-parameters. We note that this prior is also a member of the exponential family of distributions. Using this prior and the likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod f(x_i|\boldsymbol{\theta}) \propto a(\boldsymbol{\theta})^n \, e^{\sum_j c_j(\boldsymbol{\theta}) t_j(\mathbf{x})}$$

where $t_j(\mathbf{x}) = \sum d_j(x_i)$,

the posterior density will be of the form

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}) &\propto L(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) \\ &\propto a(\boldsymbol{\theta})^{n+b} \, e^{\sum_j c_j(\boldsymbol{\theta}) s_j(\mathbf{x})} \end{aligned}$$

where

$$s_j(\mathbf{x}) = d_j + t_j(\mathbf{x})$$

Thus $\pi(\boldsymbol{\theta})$ is the conjugate prior density for the likelihood, with the posterior having the same form as the prior, with $n+b$ (in the posterior) replacing b and $s_j(\mathbf{x})$ replacing d_j.

Example: Bernoulli Distribution

For a binary variable X with

$$\begin{aligned} f(x|\theta) &= \theta^x (1-\theta)^{1-x}, \; x = 0, 1; \; 0 < \theta < 1, \\ &= (1-\theta)\, e^{x \log\left(\frac{\theta}{1-\theta}\right)} \end{aligned}$$

the prior is

$$\begin{aligned} \pi(\theta) &\propto (1-\theta)^b \, e^{dlog\left(\frac{\theta}{1-\theta}\right)} \\ &\propto \theta^d (1-\theta)^{b-d}, \; 0 < \theta < 1 \end{aligned}$$

Thus the conjugate prior for the binomial family is a beta prior.

Example: If the density of random variable X is exponential:

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

and

$$\pi(\theta) \propto \theta^b e^{-\theta d}, \; 0 < \theta < 1$$

then the prior is equivalent to a gamma prior for θ.

The use of a prior density that conjugates the likelihood allows for analytic expressions of the posterior density.

Conjugate priors for common likelihood function

Family	*Conjugate Priors*
$Binomial(n, \theta)$	$\theta \sim Beta\,(a, b)$
$Poisson\,(\theta)$	$\theta \sim Gamma\,(\alpha_0, \lambda_0)$
$N\left(\mu, \sigma^2\right),\ \sigma^2\ known$	$\mu \sim N\left(\mu_0, \sigma_0^2\right)$
$N\left(\mu, \sigma^2\right),\mu\ known$	$\dfrac{1}{\sigma^2} \sim Gamma\,(\alpha_0, \lambda_0)$
$Gamma\,(\alpha, \lambda),\ \alpha\ known$	$\lambda \sim Gamma\,(\alpha_0, \lambda_0)$
$Beta\,(a, b),\ b\ known$	$\lambda \sim Gamma\,(\alpha_0, \lambda_0)$

Jeffreys' Prior

- Jeffreys' rule is to choose the prior proportional to the square root of the information,

$$I\,(\theta) = -E\left(\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right)$$

where the expectation is taken with respect to $f(\mathbf{x}|\theta)$. That is,

$$\pi\,(\theta) \propto \sqrt{I\,(\theta)}$$

- If θ is a vector parameter, informations are replaced by information matrices and Jeffreys' takes the form $det\left\{\sqrt{\mathbf{I}\,(\theta)}\right\}$ – the square root of the determinant of the information matrix.

- If a distribution for θ is non-informative, and we make a parameter transformation $\gamma = h(\theta)$, then the distribution of γ must be non-informative.
- The Jeffreys' rule allows us to find prior distributions that are *invariant* under reparameterizations.
- If a prior density $\pi\,(\theta) \propto \sqrt{I\,(\theta)}$ is used, then $\pi(\gamma) \propto \sqrt{I(\gamma)}$.
- For example: if $\pi\left(\sigma^2\right) \propto \dfrac{1}{\sigma^2}$, then $\pi\,(\sigma) \propto \dfrac{1}{\sigma}$
- In most cases, Jeffreys' priors are improper priors.
- However, posterior distributions are proper.

Example: Jeffrey's Prior for Bernoulli Trials

With n Bernoulli trials the likelihood for θ is $L(\theta) \propto \theta^s(1-\theta)^{n-s}$. To calculate Jeffres' prior we need to differentiate the log likelihood twice and take expectations. The calculation is as follows.

$$\begin{aligned} logL(\theta) &\propto slog\theta + (n-s)log(1-\theta), \\ \frac{\partial logL}{\partial \theta} &= \frac{s}{\theta} - \frac{n-s}{1-\theta}, \\ \frac{\partial^2 logL}{\partial \theta^2} &= -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}, \end{aligned}$$

since $E(s|\theta) = n\theta$,

$$I\,(\theta) = \frac{n}{\theta(1-\theta)}.$$

It follows that Jeffreys' prior is

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}, \quad \theta > 0$$

This is a $beta(1/2, 1/2)$ density which is **proper**.

Example: Jeffrey's Prior for Poisson

Suppose that the sample $X_1, X_2, \cdots, X_n$ be i.i.d. $Poisson(\theta)$.
The likelihood function

$$
\begin{aligned}
L(\theta|\mathbf{x}) &= \prod_i \frac{\theta^{x_i} e^{-\theta}}{x_i!} \\
\frac{\partial}{\partial\theta} log\ L(\theta|\mathbf{x}) &= \frac{\sum x_i}{\theta} - n \\
\frac{\partial^2}{\partial\theta^2} log\ L(\theta|\mathbf{x}) &= -\frac{\sum x_i}{\theta^2}
\end{aligned}
$$

Therefore,

$$I(\theta) = \frac{\sum E(X_i)}{\theta^2} = \frac{n}{\theta},$$

and

$$\pi(\theta) \propto 1/\sqrt{\theta}.$$

which is improper, since $\int_0^\infty \pi(\theta)\, d\theta = \int_0^\infty \theta^{-1/2} d\theta = \infty$.

$\theta \mid x \sim Gamma\left(s + \frac{1}{2}, n\right)$ is proper

Example: Exponential Distribution

If $X_1, \cdots, X_n$ is a random sample form exponential distribution with mean $1/\theta$:

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

then

$$
\begin{aligned}
logL(\theta|\mathbf{x}) &= nlog(\theta) - \theta \sum x_i \\
\frac{\partial^2}{\partial\theta^2} log\ L(\theta|\mathbf{x}) &= -\frac{n}{\theta^2}
\end{aligned}
$$

and

$$I(\theta) = \frac{n}{\theta^2},$$

Thus,

$$\pi(\theta) \propto \frac{1}{\theta},\ \theta > 0$$

Therefore, the Jeffreys prior for θ is **improper**.

Example: Exponential Distribution (cont.)

The posterior distribution is given by

$$
\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \theta^n e^{-\theta \sum x_i} \times \frac{1}{\theta} \\
&= \theta^{n-1} e^{-\theta \sum x_i},\ \theta > 0
\end{aligned}
$$

This is $gamma\,(n, \sum x_i)$, which is proper.

Example: Jeffreys' Prior for the mean of a Normal distribution:

Let $X_1, X_2, \cdots, X_n$ be i.i.d. with mean μ and variance 1. Then,

$$L(\mu|\mathbf{x}) = (2\pi)^{-n/2} exp\left[-\frac{1}{2}\sum (x_i - \mu)^2\right]$$

We can show that

$$I(\mu) = n,$$

Thus,

$$\pi(\mu) \propto c,\ -\infty < \mu < \infty$$

where c is an arbitrary constant

- Therefore, the Jeffreys prior for μ is the (**improper**) uniform distribution over the real numbers.
- The posterior distribution (we will see it later) is **proper.**

Example: Jeffreys' Prior for the variance of a Normal distribution

Let $X_1, X_2, \cdots, X_n$ be independent, normally distributed variates with known mean μ and unknown variance σ^2. Then,

$$L(\sigma^2|\mathbf{x}) = \left(2\pi\sigma^2\right)^{-n/2} exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)$$

then

$$I(\sigma^2) = \frac{n}{2\sigma^4}$$

Thus,

$$\pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2},\ \sigma^2 > 0$$

which is an improper prior.

Example: Jeffreys' Prior for the mean and the variance of a Normal distribution

Suppose $X_1, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, where (μ, σ^2) are both unknown. We can easily show that Jeffreys' prior for (μ, σ^2) is

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2},\ -\infty < \mu < \infty,\ \sigma^2 > 0$$

which is an improper prior.

Summary Table

Distribution	$\pi(\theta) \propto$	Parameter Space
$Bin(\theta)$	$\theta^{-1/2}(1-\theta)^{-1/2}$	$0 < \theta < 1$
$Pois(\theta)$	$\theta^{-1/2}$	$\theta > 0$
$Exp(\theta)$	θ^{-1}	$\theta > 0$
$N(\mu, 1)$	$cons.$	$-\infty < \mu < \infty$
$N\left(\mu, \sigma^2\right), known\ \mu$	$1/\sigma^2$	$\sigma^2 > 0$
$N\left(\mu, \sigma^2\right)$	$1/\sigma^2$	$-\infty < \mu < \infty, \sigma^2 > 0$

Prior Selection

- There are several approaches to select a sensible prior.
- A simple one is the maximum likelihood-type II (ML-II) approach.
- Let C be a class of priors under consideration. ML-II approach is to find $\hat{\pi} \in C$ satisfying

$$m_{\hat{\pi}}(x) = \underset{\pi\in C}{Max}\, m_{\pi}(x)$$

where $m_{\pi}(x)$ is called the *predictive distribution* for X.

Example

- Let $X_1, X_2, \cdots, X_n$ be i.i.d. variables with mean μ and variance 1.
- Suppose $\mu \sim N\left(\mu_0, \sigma_0^2\right)$.
- The predictive distribution for X_i is $N\left(\mu_0, 1+\sigma_0^2\right)$.
- The ML-II method is to find μ_0 and σ_0^2 by maximizing the predictive distribution.
- Taking the logarithm of $m_\pi(\mathbf{x})$, or $N\left(\mu_0, 1+\sigma_0^2\right)$, and putting the first derivatives of $m_\pi(\mathbf{x})$ with respect to μ_0 and σ_0^2 to zero, we have

$$\hat{\mu}_0 = \overline{x};\ \hat{\sigma}_0^2 = \frac{1}{n}\sum (x_i - \overline{x})^2 - 1$$

- Suppose we know $\overline{x} = 1$ and $\frac{1}{10}\sum_{i=1}^{10} (x_i - \overline{x})^2 = 3$, then $\pi(\mu) = N(1, 2)$ is an appropriate prior.

End of Session