STAT40380/STAT40390/STAT40850
Bayesian Analysis

Dr Niamh Russell

School of Mathematics and Statistics
University College Dublin

niamh.russell@ucd.ie

February 2016

Some topics to tidy up

Before we move on to deeper Bayesian ideas in hypothesis testing, computing and modelling, we need to tidy up a few things we have missed along the way.

The topics we will cover today include:

- Summarising posterior distributions
- Predictive distributions
- Jeffreys' priors

Summarising posterior distributions

- Recall the Bayesian definition of probability: *the degree of belief* in an event occuring
- This definition allows us to consider any parameter or data value as a random variable
- Contrast this with the traditional frequentist approach in which parameters are *fixed but unknown*
- In the frequentist case, all inference occurs on the sample estimators, eg $\hat{\theta}$, which have sampling distributions and thus allow for confidence intervals to be calculated.
- If we are Bayesian, we can make probability statements about our parameters without having to worry about sampling properties

Summarising posterior distributions 2

Under the Bayesian definition of probability we can:

- Report the full posterior distribution $p(\theta|\mathbf{x})$ as our degree of belief in the possible values of parameter θ given the data we have observed
- Report any summary statistics that are important to us. eg $p(\theta < 0)$ or $p(\theta = 6)$ (if θ is discrete), or the mean/mode/variance etc of θ
- Report a posterior interval of any size or proportion we require, eg $p(0.2 < \theta < 0.6) = 0.95$

In each case, the probability relates to θ (rather than to $\mathbf{x}$ in the frequentist case) so $p(a < \theta < b) = 0.95$ means *the probability that* θ *lies in the range* (a, b) *given the data* $\mathbf{x}$ *(and the proposed model) is 0.95*

Highest Density Regions (HDRs) and Credible Intervals (CIs)

There are different ways of constructing confidence intervals under the Bayesian Method

- A 95% *credible interval* (CI) represents the central 95% of the posterior probability distribution, i.e the 2.5th percentile to the 97.5th percentile.
- Example: for the normal distribution, we know that the 95% interval lies between 1.96 standard deviations of the mean.
- A *Highest (posterior) Density Region* (HDR) represents the region of values that contains the highest 95% of the probability distribution.
- An algorithm for computing HDRs involves computing a histogram of the posterior, and then including the bins with the highest frequency until 95% of the distribution is covered. A HDR can contain more than one interval.

CIs and HDRs - some pictures

Predictive distributions

- It is sometimes helpful to calculate the distribution:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$$

  This is the probability distribution of the data $\mathbf{x}$ taking into account our likelihood and prior assumptions. It is often known as the *prior predictive distribution* as it describes the probability distribution of the data before it is observed

- The prior predictive distribution is sometimes written as $p(\mathbf{x}|\mathcal{M})$ where $\mathcal{M}$ represents our modelling assumptions (for example a Binomial likelihood and a Beta prior)
- The predictive distribution is also known as the *normalising constant* as it appears in the denominator of Bayes' Theorem:

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

Predictive distributions 2

- After the data have been observed, we may wish to predict the next observation $\tilde{x}$
- This new distribution is known as the *posterior predictive distribution* and can be written as:

$$
\begin{aligned}
p(\tilde{x}|\mathbf{x}) &= \int p(\tilde{x}|\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int p(\tilde{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}
\end{aligned}
$$

- In many cases, $p(\mathbf{x})$ and $p(\tilde{x}|\mathbf{x})$ cannot be calculated analytically and so must be numerically simulated

Jeffrey's prior distributions

- We often want a *vague prior distribution* when we have little information about our parameters
- However, we know that some prior distributions give inconsistent results under transformation
- Jeffreys suggests a technique to produce vague prior distributions which are invariant under transformation
- Recall that the likelihood is written as $p(\mathbf{x}|\theta)$, ie the probability of observing the data given the parameters
- Define the Fisher information to be:

$$I(\theta|\mathbf{x}) \hat{=} -\mathbb{E}\left[\frac{\partial^2 \log p}{\partial \theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial \log p}{\partial \theta}\right)^2\right]$$

- It is important to note that the information depends on the distribution of the data rather than any particular value of it, so that $x = 5$ and $x = -2$ carry the same amount of information

Jeffrey's prior distributions 2

- Let $\psi = \psi(\theta)$, ie a transformation of the parameter θ
- Note that:

$$\frac{\partial \log p(\mathbf{x}|\psi)}{\partial \psi} = \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta} \frac{\partial \theta}{\partial \psi}$$

- Squaring and taking expectations wrt $\mathbf{x}$ we get:

$$I(\psi|\mathbf{x}) = I(\theta|\mathbf{x}) \left[\frac{\partial \theta}{\partial \psi}\right]^2$$

So if we use a prior $p(\theta) \propto \sqrt{I(\theta|\mathbf{x})}$ then, by the change-of-variable rule $p(\psi) \propto \sqrt{I(\psi|\mathbf{x})}$. Thus the prior distribution is invariant to the change of scale.

- The prior distribution $p(\theta) \propto \sqrt{I(\theta|\mathbf{x})}$ is often known as the *Jeffrey's prior* for θ.

Example 1: Jeffreys' priors

Example

Let $x \sim N(\theta, \phi)$ with θ known. Find the Jeffreys' prior for ϕ.

Jeffreys' prior for multi-parameter models

- With several unknown parameters, the Fisher information becomes

$$I(\boldsymbol{\theta}|\mathbf{x})_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p}{\partial\theta_i\theta_j}\right]$$

- Now define a vector of transformations $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_k\}$ so that $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$. Let $\mathbf{J}$ be the matrix such that

$$J_{ij} = \frac{\partial\theta_i}{\partial\psi_j}$$

- It is now the case that

$$\mathbf{I}(\boldsymbol{\psi}|\mathbf{x}) = \mathbf{J}\mathbf{I}(\boldsymbol{\theta}|\mathbf{x})\mathbf{J}^T \text{ and } \det\mathbf{I}(\boldsymbol{\psi}|\mathbf{x}) = \{\det\mathbf{I}(\boldsymbol{\theta}|\mathbf{x})\}(\det\mathbf{J})^2$$

- Our Jeffreys' prior is:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det\mathbf{I}(\boldsymbol{\theta}|\mathbf{x})}$$

Example 2: Jeffreys' priors

Example

Let $x \sim N(\theta, \phi)$ with both θ and ϕ unknown. Find the joint Jeffreys' prior for (θ, ϕ).