STAT40380/STAT40390/STAT40850
Bayesian Analysis

Dr Niamh Russell

School of Mathematics and Statistics
University College Dublin

niamh.russell@ucd.ie

March 2016

Bayesian statistics: revision

- For parameters θ and data x:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

- Posterior: $p(\theta|x)$, likelihood: $p(x|\theta)$, prior: $p(\theta)$
- For Bayesian statisticians, probability is a measure of *the degree of belief* in an event. Thus, both parameters and data can be treated as stochastic
- For frequentist statisticians, parameters are *fixed* so statements can only be made about the data
- The Bayesian method allows us to avoid the *error of the transconditional*
- Bayesian inference follows the *likelihood* and *stopping rule principles*

Bayesian statistics: revision 2

- The proportional symbol in Bayes' theorem is required because the *constant of proportionality*:

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

  (also known as the *normalising constant*) is often very tricky to calculate
- We often do not need to bother calculating it as in many cases we can recognise the form of the posterior distribution as another distribution we already know
- In such cases, and where the posterior distribution is in a similar form to the prior, we have a *conjugate prior distribution*
- It is very easy to find conjugate prior distributions for member of the *exponential family*

Bayesian statistics: revision 3

- When we have little prior information, we may wish to use an *improper* or *vague* prior distribution
- Many vague prior distributions, however, have the property that they are not invariant to transformations of the parameter space. In such cases we may like to use a *Jeffreys' prior distribution*, defined as:

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

  where $I(\theta|x)$ is the Fisher Information
- To test hypotheses (eg competing models $\mathcal{M}_0$ and $\mathcal{M}_1$) via Bayes' theorem, we may calculate a *Bayes Factor*:

  $$BF = \frac{p(x|\mathcal{M}_1)}{p(x|\mathcal{M}_0)}$$

- The Bayes Factor will give the odds in favour of $\mathcal{M}_1$ over $\mathcal{M}_0$

Computatioinal Bayesian statistics

- So far, we have been restricted to dealing with likelihoods and prior distributions which are conjugate
- What if somebody specifies a prior distribution which does not yield a neat posterior distribution?
- Example: $x_i \sim N(\theta, 1), \;\; \theta \sim exp(4)$.
- What if there are so many parameters that we could not feasibly find marginal distributions for our parameters of interest?
- The problem usually exists because we cannot calculate the *normalising constant*
- The solution lies in Monte Carlo techniques such as the *EM algorithm, Gibbs' sampling* and *the Metropolis-Hastings algorithm*.

The EM algorithm

- A useful tenique for finding the posterior mode, but not necessarily the full posterior distribution.
- (Sometimes the posterior mode is the main quantity of interest)
- The EM algorithm works by *augmenting* the dataset with fictitious extra data, and then iterating through estimates of the parameters until convergence
- EM stands for *Expectation-Maximisation* which are the two steps taken at every iteration
- We will write the original data as **x**, the augmented data as **y** (ie made up of **x** and extra hypothetical observations), and the parameter interations as $\theta^{(0)}, \theta^{(1)}, \ldots$. Let $\theta^{(t)}$ be our guess at the value of θ at iteration *t*.

EM steps

1. Take a guess for $\theta^{(0)}$
2. E step: compute

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{y|x,\theta}[\log p(\theta|\mathbf{y}]_{\theta=\theta^{(t)}}$$

3. M step: find

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)})$$

   (Usually this is done by differentiating Q by θ and setting to zero)
4. We repeat the above steps until convergence

Example:the EM algorithm

Example

Suppose that balls are thrown into 4 pots, such that each ball lies in each pot with probabilities:

$$\left(\frac{1}{2}+\frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right)$$

we observe data $\mathbf{x} = \{125, 18, 20, 34\}$. Use the EM algorithm to estimate θ when the reference prior $Be(0, 0)$ is used.

Why does the EM algorithm work?

- Main result: $p(\theta^{(t)}|\mathbf{x})$ increases with t, so with each iteration we will *head towards the posterior mode*
- Outline proof. Start with:

$$\log p(\theta|\mathbf{x}) = \log p(\theta|\mathbf{y}) - \log p(\mathbf{y}|\boldsymbol{\theta},\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x})$$

- Multiplying both sides by $p(\mathbf{y}|\theta^{(t)},\mathbf{x})$ and integrating over $\mathbf{y}$ gives:

$$\log p(\theta|\mathbf{x}) = Q(\theta,\theta^{(t)}) - H(\theta,\theta^{(t)}) + K(\theta^{(t)})$$

- Proof proceeds by calculating $\log p(\theta^{t+1}|\mathbf{x}) - \log p(\theta^{t}|\mathbf{x})$ which depends on $-[H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)})]$. This is shown to be greater than or equal to 0
- More complete proof in Lee textbook