Logistics and events

- Fire exits
- I/O: Tea/coffee/lunch (Blackett 311), toilets
- Breakfast 8.15-8.45 a.m.
- Events:
  - Talk by Tom Babbedge (Winton) today ~5 p.m.
  - Barbecue tonight 6 p.m. 58 Princes Gate
  - Drinks reception 5:30 p.m. tomorrow
  - Public engagement lunch, Wednesday
Outline of course

• Basic principles
• Sampling
• Numerical methods (Parameter inference)
• MCMC
• Hybrid/Hamiltonian Monte Carlo
• Bayesian Hierarchical Models
• Bayesian Evidence (Model selection)

Outline

• Inverse problems: from data to theory
• Probability review, and Bayes’ theorem
• Parameter inference
• Priors
• Marginalisation
• Posteriors

ICIC Data Analysis Workshop: the Bayesics

Alan Heavens
Imperial College London
ICIC Data Analysis Workshop
8 September 2014

LCDM fits the WMAP data well.
Inverse problems

- Most cosmological problems are inverse problems, where you have a set of data, and you want to infer something.
- Generally harder than predicting the outcomes when you know the model and its parameters.
- Examples
  - Hypothesis testing
  - Parameter inference
  - Model selection

What is probability?

- Frequentist view: $p$ describes the relative frequency of outcomes in infinitely long trials.
- Bayesian view: $p$ expresses our degree of belief.
- Bayesian view is what we seem to want from experiments: e.g., given the Planck data, what is the probability that the density parameter of the Universe is between 0.9 and 1.1?

Bayes’ Theorem

- Rules of probability:
  - $p(x) + p(\text{not } x) = 1$ sum rule
  - $p(x, y) = p(x|y) p(y)$ product rule
  - $p(x) = \sum_k p(x, y_k)$ marginalisation
  - Sum → integral continuum limit (p=pdf)
    
    $p(x) = \int dy \, p(x, y)$
  - $p(x,y) = p(y|x)$ gives Bayes’ theorem

Examples

- Hypothesis testing
  - Is the CMB radiation consistent with (initially) gaussian fluctuations?
- Parameter inference
  - In the Big Bang model, what is the value of the matter density parameter?
- Model selection
  - Do cosmological data favour the Big Bang theory or the Steady State theory?
  - Is the gravity law General Relativity or a different theory?
p(x | y) is not the same as p(y | x)

- x = female, y = pregnant
- p(y | x) = 0.03
- p(x | y) = 1

Bayes’ Theorem and Inference

- If we accept p as a degree of belief, then what we often want to determine is
  \[ p(\theta | x) \]
- \( \theta \): model parameter(s), \( x \): the data

To compute it, use Bayes’ theorem

\[ p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)} \]

Note that these probabilities are all conditional on a) prior information \( I \), b) a model \( M \)

\[ p(\theta | x) = p(\theta | x, I, M) \text{ or } p(\theta | x, I) \]

*This is RULE 1: start by writing down what it is you want to know
RULE 2: There is no RULE n, n>1

The Monty Hall problem:
An exercise in using Bayes’ theorem

Do you change your choice?

This is the Monty Hall problem

Posterior, likelihoods, priors and evidence

\[ p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)} \]

Remember that we interpret these in the context of a model \( M \), so all probabilities are conditional on \( M \) (and on any prior info \( I \)). E.g. \( p(\theta) = p(\theta | M) \)

The evidence looks rather odd – what is the probability of the data? For parameter estimation, we can ignore it – it simply normalises the posterior. If you need it,\n
\[ p(x) = \sum_k p(x | \theta_k) p(\theta_k) \text{ or } p(x) = \int d\theta \ p(x | \theta) p(\theta) \]

Noting that \( p(x) = p(x | M) \) makes its role clearer.

In model selection (from \( M \) and \( M' \)), \( p(x | M) \neq p(x | M') \)
Forward modelling $p(x|\theta)$

With noise properties we can predict the Sampling Distribution (the probability of obtaining a general set of data). The Likelihood refers to the specific data we have - it isn't a probability, strictly.

Set up the problem

- What is the model for the data, $M$?
- $M$: $x = \mu + n$
- Data: a set of values $\{x_i\}, i=1...N$
- Prior info $\theta$: noise $<n>=0 <n^2>=\sigma^2$ (known); gaussian distributed
- $\theta$: the mean, $\mu$
- Rule 1: what do we want?
- $p(\mu|\{x_i\})$
- See Jonathan’s lectures for the solution

Case study: the mean

- Given a set of $N$ independent samples $\{x_i\}$ from the same distribution, with gaussian dispersion $\sigma$, what is the mean of the distribution $\mu = \langle x \rangle$?
- Bayes: compute the posterior probability $p(\mu|\{x_i\})$
- Frequentist: devise an estimator $\hat{\mu}$ for $\mu$. Ideally it should be unbiased, so $\langle \hat{\mu} \rangle = \mu$ and have as small an error as possible (minimum variance).
- These lead to superficially identical results (although they aren’t), but the interpretation is very different
- Bayesian: no estimators - just posteriors

State your priors

- In easy cases, the effect of the prior is simple
- As experiment gathers more data, the likelihood tends to get narrower, and the influence of the prior diminishes
- Rule of thumb: if changing your prior† to another reasonable one changes the answers a lot, you could do with more data
- Reasonable priors? Noninformative* – constant prior
- scale parameters in $[0, \infty)$; uniform in log of parameter (Jeffreys’ prior*)
- Beware: in more complicated, multidimensional cases, your prior may have subtle effects…

† I mean the raw theoretical one, not modified by an experiment
* Actually, it’s better not to use these terms – other people use them to mean different things – just say what your prior is!
From Sivia & Skilling’s *Data Analysis* book. **IS THE COIN FAIR?**

Model: independent throws of coin. Parameter $\theta =$ probability of $H$

**The effect of priors**

Priors = “It’s likely to be nearly fair”, “It’s likely to be very unfair”

VSA CMB experiment

(Slosar et al 2003)

Priors: $\Omega_\Lambda \geq 0$

$10 \leq \text{age} \leq 20 \text{ Gyr}$

$h \approx 0.7 \pm 0.1$

There are no data in these plots – it is all coming from the prior!

$p(\theta) = \int d\theta_{\neq 1} p(x|\theta) p(\theta)$
Inferring the parameter(s)

- What to report, when you have the posterior?
  - Commonly the **mode** is used (the peak of the posterior)
  - **Mode = Maximum Likelihood Estimator, if the priors are uniform**
  - The **posterior mean** may also be quoted, but beware
  - Ranges containing x% of the posterior probability of the parameter are called **credibility intervals** (or **Bayesian confidence intervals**)

Errors

- If we assume uniform priors, then the posterior is proportional to the likelihood.
  If further, we assume that the likelihood is single-moded (one peak at \( \theta_0 \)), we can make a Taylor expansion of \( \ln L \):  
  \[
  \ln L(x; \theta) = \ln L(x; \theta_0) + \frac{1}{2} (\theta_\alpha - \theta_0\alpha) H_{\beta\alpha} (\theta_\beta - \theta_0\beta) + \ldots
  \]
  where the Hessian matrix is defined by these equations. Comparing this with a gaussian, the **conditional error** (keeping all other parameters fixed) is  
  \[
  \sigma_\alpha = \frac{1}{\sqrt{H_{\alpha\alpha}}}
  \]
  Marginalising over all other parameters gives the **marginal error**  
  \[
  \sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}
  \]

Non-gaussian likelihoods: number counts

- A radio source is observed with a telescope which can detect sources with fluxes above \( S_0 \). The radio source has a flux \( S_1 = 2S_0 \) (assume it is precisely measured).
  What is the slope of the number counts?
  (Assume \( N(S) dS \propto S^{-\alpha} dS \))

Can you tell anything?

Multimodal posteriors etc

- Peak may not be gaussian
  - Multimodal? Characterising it by a mode and an error is probably inadequate. May have to present the full posterior.
  - Mean posterior may not be useful in this case – it could be very unlikely, if it is a valley between 2 peaks.
Summary

- Write down what you want to know. For *parameter inference* it is typically:

\[ p(\theta | xIM) \]

- What is \( M \)?
- What is/are \( \theta \)?
- What is \( I \)?
- You might want \( p(M|xI) \)...this is *Model Selection* - see later