BCCS 2008/09: Graphical models and complex stochastic systems: Lecture 10: Bayesian model choice

As we said early on, 'all models are wrong, some models are useful'. Very rarely is the choice of a model absolutely fixed, much more commonly we will want to use data to help formulate or choose a model, or to criticise a model currently under consideration. There are many different tasks of this kind, and various approaches, Bayesian or non-Bayesian. It is an area of particularly active current research, so it is impossible to give a brief overview. Here we will just discuss one problem – choosing between several specified models, and fitting the unknown parameters in each – that is, simultaneous inference about the model and its parameters.

10.1 Extending the hierarchical set-up

Let us label the models in question by k (perhaps $k = 1, 2, \ldots$, but we can label the models to suit the application). The prior probability that model k generated the data is $p(k)$. Model k has a parameter vector θ_k (NB, this does not mean the kth component of a vector θ). Note that there is no reason why the dimension of θ_k should be the same for all k. For each k we will have a prior distribution for the parameter θ_k, denoted $p(\theta_k|k)$, and a likelihood for the data Y, denoted $p(Y|k, \theta_k)$.

By the ordinary rules for probability, the joint distribution of everything is

$$p(k, \theta_k, Y) = p(k)p(\theta_k|k)p(Y|k, \theta_k)$$

This really just corresponds to adding one more 'top' level to bring together a collection of hierarchical models for the same data Y. Our task is to make inference about both k and θ_k; but as usual by Bayes' theorem,

$$p(k, \theta_k, Y) = \frac{p(k, \theta_k, Y)}{p(Y)} \propto p(k, \theta_k|Y)$$

In other words, the model indicator k becomes just like one more unknown parameter. We can compute this posterior distribution by any MCMC method that can cope with the possibility that the unknowns, the state variable of the Markov chain, (k, θ_k) is not of fixed dimension, e.g. reversible jump. (This kind of application was the original motivation for creating these methods).

Once we have a (simulation-based) approximation to $p(k, \theta_k|Y)$, we can derive the *posterior model probabilities* $p(k|Y)$ by simply ignoring the simulated values of the θ_k, so that $p(k|Y)$ is estimated by the sample relative frequency of 'visiting' model k. Similarly, we can approximate the *within-model parameter posterior* $p(\theta_k|k, Y)$ by considering only the θ values in the sample obtained when the MCMC was visiting the correct model.

Note that in this basic form, the Bayesian approach does not *choose* a specific model, it finds the probability for each of a range of models; if we need to choose just one, we might choose the model with the highest posterior probability.

10.2 Bayesian model averaging

It is often a good idea not to choose just one model and then behave as if it is known to be true. In reality there is some doubt, so it is better to recognise that doubt by averaging over models, weighted by their posterior probabilities.

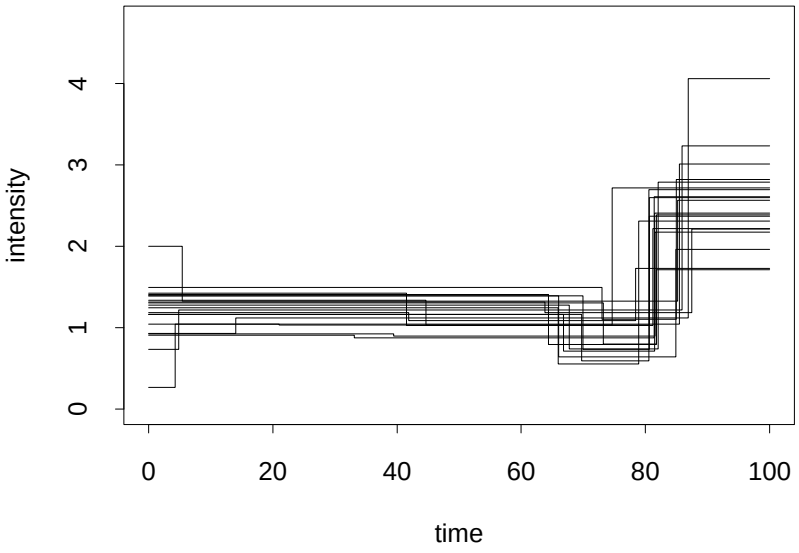

Figure 1: A small MCMC sample of $x(t)$ for the 3–changepoint model

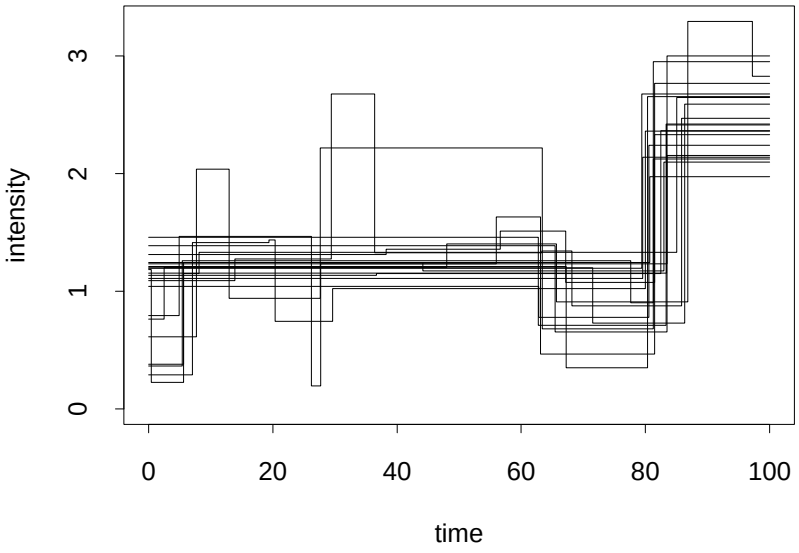

Figure 2: A small MCMC sample of $x(t)$ for the variable number of changepoints model

For example, suppose we want to do prediction – that is state the distribution of the *next* observation Y^+. Then

$$p(Y^+|Y) = \frac{p(Y, Y^+)}{p(Y)} = \frac{\sum_k p(k) \int p(\theta_k|k) p(Y|k, \theta_k) p(Y^+|k, \theta_k) d\theta_k}{\sum_k p(k) \int p(\theta_k|k) p(Y|k, \theta_k) d\theta_k}$$

which can be re-written as

$$\sum_k \int p(k, \theta_k|Y) p(Y^+|k, \theta_k) d\theta_k,$$

that is, as posterior model averaging.

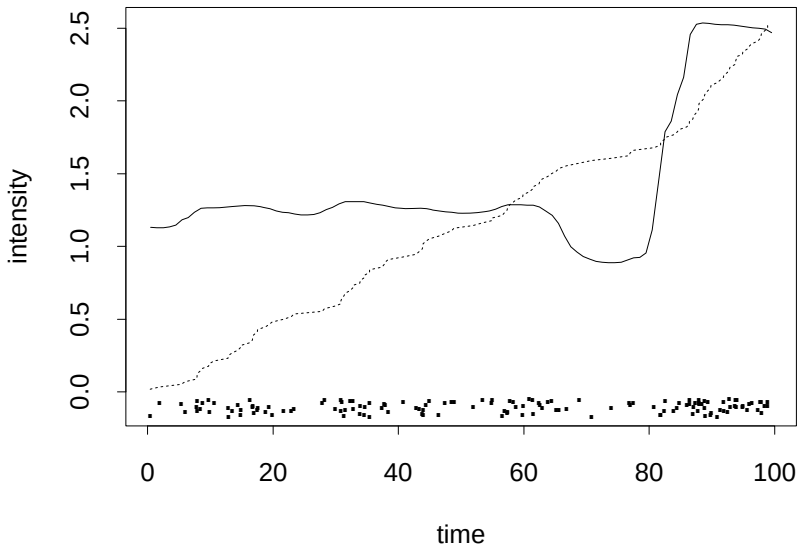

Figure 3: Posterior average of $x(t)$ for the variable number of changepoints model

10.3 Bayes factors

Some people find the idea of specifying prior model probabilities $p(k)$ logically difficult. We can get round that by only using the data to *compare* models. Note that comparing models k_1 and k_0,

$$\frac{p(k_1|Y)}{p(k_0|Y)} = \frac{p(k_1, Y)}{p(k_0, Y)} = \frac{p(k_1)}{p(k_0)} \times \frac{p(Y|k_1)}{p(Y|k_0)}.$$

We call the last term the *Bayes factor* for model k_1 vs. model k_0. So in words, posterior odds equal prior odds times Bayes factor. The probability $p(Y|k)$ is called the *marginal likelihood* for model k – marginal since the parameters θ_k have been 'integrated out'.

As a rough rule of thumb, the pioneering statistician Jeffreys proposed that Bayes factors greater that 3, 10 or 100 should be interpreted as saying that the data provided 'substantial', 'strong' or 'decisive' preference for model k_1 rather than k_0, respectively.

10.4 Variable selection

A very important special case of model choice is when you have a multiple linear regression model:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_d x_{di} + \text{error}$$

and you don't believe that all of the terms on the right are actually needed – we want to select the variables $\{x_j\}$ that have a significant effect on Y.

This is a model selection problem with 2^d models, corresponding to all the possible choices of covariates, and has been a particular focus of research interest.

10.5 Reading

Some articles in the book of Gilks, Richardson and Spiegelhalter (1996) are relevant, but most of the literature is still found only in journals.