How to compute posterior model probabilities ... and why that doesn’t mean that we have solved the problem of model choice

Peter Green

School of Mathematics
University of Bristol

15 March 2011, Groningen: All models are wrong...
Outline

1. Ideal Bayes
2. Across-model methodology
3. Examples
4. Alternatives to across-model sampling
5. Methodological relatives
6. Better Bayes factors
7. Real Bayes
Hierarchical model

Suppose given

- a prior $p(k)$ over models $k$ in a countable set $\mathcal{K}$, and
- for each $k$
  - a prior distribution $p(\theta_k|k)$, and
  - a likelihood $p(Y|k, \theta_k)$ for the data $Y$.

For definiteness and simplicity, suppose that $p(\theta_k|k)$ is a density in $n_k$ dimensions, and that there are no other parameters, so that where there are parameters common to all models these are subsumed into each $\theta_k \in \mathcal{R}^{n_k}$.

Additional parameters, perhaps in additional layers of a hierarchy, are easily dealt with. All probability distributions are proper.
Hierarchical model

Suppose given

- a prior $p(k)$ over models $k$ in a countable set $\mathcal{K}$, and
- for each $k$
  - a prior distribution $p(\theta_k|k)$, and
  - a likelihood $p(Y|k, \theta_k)$ for the data $Y$.

For definiteness and simplicity, suppose that $p(\theta_k|k)$ is a density in $n_k$ dimensions, and that there are no other parameters, so that where there are parameters common to all models these are subsumed into each $\theta_k \in \mathcal{R}^{n_k}$.

Additional parameters, perhaps in additional layers of a hierarchy, are easily dealt with. All probability distributions are proper.
Hierarchical model as DAG

\[ p(k) \quad \theta \quad p(Y | \kappa, \theta_k) \]

- \( p(k) \): model indicator
- \( \theta \): model-specific parameter
- \( p(Y | \kappa, \theta_k) \): data

\[ k \]

\[ \theta \]

\[ Y \]
The joint posterior

\[ p(k, \theta_k | Y) = \frac{p(k)p(\theta_k | k)p(Y | k, \theta_k)}{\sum_{k' \in K} \int p(k')p(\theta'_{k'} | k')p(Y | k', \theta'_{k'})d\theta'_{k'}} \]

can always be factorised as

\[ p(k, \theta_k | Y) = p(k | Y)p(\theta_k | k, Y) \]

– the product of posterior model probabilities and model-specific parameter posteriors.
– very often the basis for reporting the inference, and in some of the methods mentioned below is also the basis for computation.
Bayes factors, Evidence

Perhaps you didn’t think you wanted to report $p(k|Y)$? Of course, with these posterior probabilities, we can report Bayes Factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)} = \frac{p(k|Y)}{p(l|Y)} \div \frac{p(k)}{p(l)}$$

for pairwise comparison of models.

For some, the Evidence (= marginal likelihood) $p(Y|k)$ has an intrinsic meaning and interpretation.
Bayesian model averaging, prediction

We can do model averaging

\[ E(F|Y) = \sum_k \int F(k, \theta_k)p(k, \theta_k|Y)d\theta_k \]

for any function \( F \) with the same interpretation in each model; the expectation can be estimated simply by averaging \( F \) along the entire run, essentially ignoring the model indicator \( k \).

As an example, we can do prediction:

\[ p(Y^+|Y) = \sum_k p(Y^+|k, Y)p(k|Y) \]

a posterior-weighted mixture of the within-model-\( k \) predictions

\[ p(Y^+|k, Y) = \int p(Y^+|k, \theta_k)p(\theta_k|k, Y)d\theta_k \]
Bayesian model averaging, prediction

We can do model averaging

\[
E(F|Y) = \sum_k \int F(k, \theta_k) p(k, \theta_k|Y) d\theta_k
\]

for any function \( F \) with the same interpretation in each model; the expectation can be estimated simply by averaging \( F \) along the entire run, essentially ignoring the model indicator \( k \).

As an example, we can do prediction:

\[
p(Y^+|Y) = \sum_k p(Y^+|k, Y)p(k|Y)
\]

a posterior-weighted mixture of the within-model-\( k \) predictions

\[
p(Y^+|k, Y) = \int p(Y^+|k, \theta_k)p(\theta_k|k, Y) d\theta_k
\]
Note the generality of this basic formulation: it embraces both

- genuine model-choice situations, where the variable $k$ indexes the collection of discrete models under consideration, but also

- settings where there is really a single model, but one with a variable dimension parameter, for example a functional representation such as a series whose number of terms is not fixed (in which case, $k$ is unlikely to be of direct inferential interest).
Splines
Polynomials
Compatibility across models

Some would argue that responsible adoption of this Bayesian hierarchical model presupposes that, e.g., $p(\theta_k|k)$ should be compatible in that inference about functions of parameters that are meaningful in several models should be approximately invariant to $k$.

Such compatibility could in principle be exploited in the construction of MCMC methods (how?).

But it is philosophically tenable that no such compatibility is present, and we can’t assume it.
Across- and within-model simulation

How to compute $p(k, \theta_k | Y)$?

Two main approaches using MCMC:

- **across**: one MCMC simulation with states of the form $(k, \theta_k) \approx p(k, \theta_k | Y)$
- **within**: separate simulations of $\theta_k \approx p(\theta_k | k, Y)$ for each $k$.

and beyond straight MCMC:

- **particle filter**: SMC in place of MCMC
- **ABC**: ‘likelihood-free’ methods where $Y | k, \theta_k$ can be simulated but $p(Y | k, \theta_k)$ cannot be evaluated.
- **nested sampling**
- **variational methods**
Across- and within-model simulation

How to compute $p(k, \theta_k | Y)$?

Two main approaches using MCMC:

- **across**: one MCMC simulation with states of the form $(k, \theta_k) \approx p(k, \theta_k | Y)$
- **within**: separate simulations of $\theta_k \approx p(\theta_k | k, Y)$ for each $k$.

And beyond straight MCMC:

- **particle filter**: SMC in place of MCMC
- **ABC**: ‘likelihood-free’ methods where $Y | k, \theta_k$ can be simulated but $p(Y | k, \theta_k)$ cannot be evaluated.
- **nested sampling**
- **variational methods**
Across-model simulation: Reversible jump MCMC

The across-model MCMC simulator follows the ideal Bayesian, and treats \((k, \theta_k)\) as a single unknown. The state space for an across-model simulation is \(\{(k, \theta_k)\} = \bigcup_{k \in K} (\{k\} \times \mathbb{R}^{n_k})\).

Mathematically, this is not a particularly awkward object. But at least a little non-standard, when \(n_k\) varies with \(k\).

We use Metropolis-Hastings to build a suitable reversible chain.

On the face of it, this requires measure-theoretic notation, which may be unwelcome! The point of the ‘reversible jump’ framework is to render the measure theory invisible, by means of a construction using only ordinary densities. Even the fact that we are jumping dimensions becomes essentially invisible.
Across-model simulation: Reversible jump MCMC

The across-model MCMC simulator follows the ideal Bayesian, and treats \((k, \theta_k)\) as a single unknown. The state space for an across-model simulation is \(\{(k, \theta_k)\} = \bigcup_{k \in K} (\{k\} \times \mathcal{R}^{n_k})\).

Mathematically, this is not a particularly awkward object. But at least a little non-standard, when \(n_k\) varies with \(k\).

We use Metropolis-Hastings to build a suitable reversible chain.

On the face of it, this requires measure-theoretic notation, which may be unwelcome! The point of the ‘reversible jump’ framework is to render the measure theory invisible, by means of a construction using only ordinary densities. Even the fact that we are jumping dimensions becomes essentially invisible.
Let’s avoid all abstraction!

Consider how the transition will be implemented.

Assume $\mathcal{X} \subset \mathcal{R}^d$, and that $\pi$ has a density (also denoted $\pi$).

At the current state $x$, we generate, say, $r$ random numbers $u$ from a known joint density $g$, and then form the proposed new state as a deterministic function of the current state and the random numbers: $x' = h(x, u)$, say.

The reverse transition from $x'$ to $x$ would be made with the aid of random numbers $u' \sim g'$ giving $x = h'(x', u')$. 
Modelling the code

Reversible jump

\[ x = h'(x', u') \]

\[ x' = h(x, u) \]

\[ u' \]

\[ h' \]

\[ h \]

\[ x' = h(x, u) \]
The equilibrium probability of jumping from \( A \) to \( B \) is then an integral with respect to \((x, u)\):

\[
\int_{(x, x' = h(x, u)) \in A \times B} \pi(x) g(u) \alpha(x, x') \, dx \, du.
\]

The equilibrium probability of jumping from \( B \) to \( A \) is an integral with respect to \((x', u')\):

\[
\int_{(x = h'(x', u'), x') \in A \times B} \pi(x') g'(u') \alpha(x', x) \, dx' \, du'.
\]

If the transformation from \((x, u)\) to \((x', u')\) is a diffeomorphism (the transformation and its inverse are differentiable), then we can apply the standard change-of-variable formula, to write this as an integral with respect to \((x, u)\).
Detailed balance says the two integrals must be equal: it holds if

\[ \alpha(x, x') = \min \left\{ 1, \frac{\pi(x')g'(u')}{\pi(x)g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \right\}, \]

involving only ordinary joint densities.
What’s the point?

Perhaps a little indirect! – but it’s a flexible framework for constructing quite complex moves using only elementary calculus.

The possibility that \( r < d \) covers the typical case that given \( x \in \mathcal{X} \), only a lower-dimensional subset of \( \mathcal{X} \) is reachable in one step.

(The Gibbs sampler is the best-known example of this, since in that case only some of the components of the state vector are changed at a time, although the formulation here is more general as it allows the subset not to be parallel to the coordinate axes.)
What’s the point?

Perhaps a little indirect! – but it’s a flexible framework for constructing quite complex moves using only elementary calculus.

The possibility that $r < d$ covers the typical case that given $x \in \mathcal{X}$, only a lower-dimensional subset of $\mathcal{X}$ is reachable in one step.

(The Gibbs sampler is the best-known example of this, since in that case only some of the components of the state vector are changed at a time, although the formulation here is more general as it allows the subset not to be parallel to the coordinate axes.)
The trans-dimensional case

But the main benefit of this formalism is that

\[ \alpha(x, x') = \min \left\{ 1, \frac{\pi(x')g'(u')}{\pi(x)g(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right| \right\}, \]

applies, without change, in a variable dimension context.

(Use the same symbol \( \pi(x) \) for the target density whatever the dimension of \( x \) in different parts of \( \mathcal{X} \).)

Provided that the transformation from \((x, u)\) to \((x', u')\) remains a diffeomorphism, the individual dimensions of \( x \) and \( x' \) can be different. The dimension-jumping is ‘invisible’.
Dimension matching

Suppose the dimensions of $x, x', u$ and $u'$ are $d, d', r$ and $r'$ respectively, then we have functions $h : \mathbb{R}^d \times \mathbb{R}^r \rightarrow \mathbb{R}^{d'}$ and $h' : \mathbb{R}^{d'} \times \mathbb{R}^{r'} \rightarrow \mathbb{R}^{d}$, used respectively in $x' = h(x, u)$ and $x = h'(x', u')$.

For the transformation from $(x, u)$ to $(x', u')$ to be a diffeomorphism requires that $d + r = d' + r'$, so-called ‘dimension-matching’; if this equality failed, the mapping and its inverse could not both be differentiable.

Dimension matching is necessary but not sufficient.
Dimension matching

\[ d + r = d' + r' \]

\[ 1 + 2 = 2 + 1 \]
Dimension matching

\[ d + r = d' + r' \]

\[ 1 + 1 = 2 + 0 \]
We wish to use these reversible jump moves to sample the space \( \mathcal{X} = \bigcup_{k \in \mathcal{K}} \{k\} \times \mathcal{R}^{n_k} \) with invariant distribution \( \pi \), which here is \( p(k, \theta_k | Y) \).

Just as in ordinary MCMC, we typically need multiple types of moves to traverse the whole space \( \mathcal{X} \). Each move is a transition kernel reversible with respect to \( \pi \), but only in combination do we obtain an ergodic chain.

The moves will be indexed by \( m \): consider a particular move \( m \) that proposes to take \( x = (k, \theta_k) \) to \( x' = (k', \theta_{k'}') \) or vice-versa for a specific pair \( (k, k') \).
Across-model methodology

The acceptance probability is derived as before, and yields

\[ \alpha_m(x, x') = \min \left\{ 1, \frac{\pi(x')}{\pi(x)} \frac{j_m(x') g_m(u')}{j_m(x) g_m(u)} \left| \frac{\partial (x', u')}{\partial (x, u)} \right| \right\}. \]

Here \( j_m(x) \) is the probability of choosing move type \( m \) when at \( x \), the variables \( x, x', u, u' \) are of dimensions \( d_m, d'_m, r_m, r'_m \) respectively, with \( d_m + r_m = d'_m + r'_m \), we have \( x' = h_m(x, u) \) and \( x = h'_m(x', u') \), and the Jacobian has a form correspondingly depending on \( m \).
Toy example

..... of no statistical use at all!
Suppose $x$ lies in $\mathcal{R} \cup \mathcal{R}^2$: $\pi(x)$ is a mixture:
with probability $p_1$, $x \sim \text{Beta}(2, 3)$,
with probability $p_2$, $(x_1, x_2, 1 - x_1 - x_2) \sim \text{Dirichlet}(2, 3, 4)$.
I will use three moves:

1. within $\mathcal{R}$: $x \rightarrow U(x - \epsilon, x + \epsilon)$.
2. within $\mathcal{R}^2$: $(x_1, x_2) \rightarrow (x_2, x_1)$.
3. between $\mathcal{R}$ and $\mathcal{R}^2$

In $\mathcal{R}$, choose (1) or (3) with probabilities $1 - r_1, r_1$.
In $\mathcal{R}^2$, choose (2) or (3) with probabilities $1 - r_2, r_2$.
Thus $j_3(x) = r_1$ for all $x \in \mathcal{R}$ and $j_3(x') = r_2$ for all $x' \in \mathcal{R}^2$. 
Dimension-changing with move (3)

Proposal:
To go from \( x \in \mathcal{R} \) to \((x_1, x_2) \in \mathcal{R}^2\), draw \( u \) from \( U(0, 1) \) [so \( g_3(u) = 1 \) if \( 0 < u < 1 \)] and propose \((x_1, x_2) = (x, u)\). For reverse move, no \( u' \) required [write \( g_3'(u') \equiv 1 \)] and set \( x = x_1 \). This certainly gives a bijection: \((x, u) \leftrightarrow (x_1, x_2)\), with Jacobian = 1.

Acceptance decision:

\[
\alpha = \min \left\{ 1, \frac{\pi(x') j_3(x') g_3'(u')} {\pi(x) j_3(x) g_3(u)} \left| \frac{\partial(x', u')} {\partial(x, u)} \right| \right\}
\]

\[
= \min \left\{ 1, \frac{p_2 f_D(x, u) r_2} {p_1 f_B(x)} \frac{1} {r_1} |1| \right\}
\]

\[
= \min \left\{ 1, \frac{p_2 r_2 f_D(x, u)} {p_1 r_1 f_B(x)} \right\}
\]

For reverse move, \( \alpha = \min\{1, (p_1 r_1 f_B(x))/(p_2 r_2 f_D(x, u))\} \).
Toy example

after 1 sweeps
Toy example

after 2 sweeps
Toy example

after 3 sweeps
Toy example

after 4 sweeps
Toy example

after 5 sweeps
Examples

Toy example

after 6 sweeps
Toy example

after 7 sweeps

Peter Green (Bristol) Computing posterior model probabilities Groningen, March 2011 33 / 69
Examples

Toy example

after 8 sweeps
Toy example

after 9 sweeps

Peter Green (Bristol)
Computing posterior model probabilities
Groningen, March 2011 35 / 69
Toy example

after 10 sweeps
Toy example

after 20 sweeps
Toy example

after 30 sweeps
Toy example

after 40 sweeps
Toy example

after 50 sweeps
Toy example

after 100 sweeps
Toy example

after 200 sweeps
Toy example

after 300 sweeps
Toy example

after 400 sweeps
Toy example
Toy example

after 1000 sweeps
Examples

\[ x_1 \sim \text{Beta}(2,3) \]

\[ x_2[1] \sim \text{Beta}(2,7) \]

\[ x_2[2] \sim \text{Beta}(3,6) \]
Cyclones in Bay of Bengal

A point process – of cyclone events – in time: modelled here as an inhomogeneous Poisson process, with piecewise constant intensity.

Examples

Cyclones in Bay of Bengal

Multiple change-point analysis
A Bayesian approach to change-point analysis for point processes

\[
\text{Poisson-process likelihood: } p(y_j | x) = \exp \left( -\mu \sum_{i=1}^{n} x(t_i) \right) \prod_{i=1}^{n} x(t_i)^{y_i} \frac{1}{y_i!}
\]

Prior model for step function \(x(t)\), representing intensity

\[
\text{Prior model: represent step function by } x(t) = \sum_{j=1}^{k} h_j I_{[s_j, s_{j+1})}
\]

\(k\): number of steps

\(h_j\): step heights

\(s_j\): step positions

\(\mu = \lambda_1 + \lambda_2\)

\(\lambda_1, \lambda_2\) are gamma distributed

18

MCMC for step functions

I will use four moves:

(a) Metropolis change to a randomly chosen step height \(h_j\).

(b) Metropolis change to a randomly chosen step position \(s_j\).

(c) Jump move: birth/death of steps – birth: choose new step position \(s_j\) at random, split current step height \(h_j\) into two: \(h_j \rightarrow h_j^+ + h_j^-\) – death: choose step at random to kill, combine current step heights \(h_j \rightarrow h_j^+ + h_j^-\) into one: \(h_j \rightarrow h_j^+ + h_j^-\).

(d) Update hyperparameters \(\lambda_1, \lambda_2\).

Birth and death of steps

\[
h_j^+ + h_j^- = h_j \]

Example:
cyclones hitting the Bay of Bengal

141 cyclones over a period of 100 years (a cyclone is a storm with winds \(>88\) km h\(^{-1}\)).
Cyclones in Bay of Bengal

Illustrating the model-jumping move: splitting and merging steps.
Cyclones in Bay of Bengal

A small sample from the posterior
Examples

Cyclones in Bay of Bengal

Posterior of the number of steps

<table>
<thead>
<tr>
<th>k</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
</tr>
<tr>
<td>6</td>
<td>0.10</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
</tr>
<tr>
<td>12</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Prior on $k$: Poisson($\lambda$), with $\lambda = 3$.

Prior on $h_j$: Gamma($\alpha$, $\beta$), with $\alpha = \beta = 1$.

Zero change points is ruled out; $k = 1$ or 2 more probable than under the prior.

Posterior density estimates for change-point positions

Model-averaged estimate: $E(x | y)$ (the expectation of a random step function is not a step function).
Cyclones in Bay of Bengal

An example of inference conditional on $k$: posterior densities of step positions
An example of Bayesian model averaging: posterior expectation of the intensity
Alternatives to joint model-parameter sampling

- When marginal likelihoods are tractable, it’s usually a good idea to compute them (thus integrating out $\theta_k$) then conduct search/sampling only over the model indicator $k$.

- This direct approach has been taken a little further by Godsill (2001), who considers cases of ‘partial analytic structure’, where some of the parameters in $\theta_k$ may be integrated out, and the others left unchanged in the move that updates the model, to give an across-model sampler with probable superior performance.

- Marginal likelihoods via within-model sampling: Nial’s talk!
Some issues in choosing a sampler

- How many models are there?
- Do we want results across $k$, within each $k$, or for one $k$ of interest?
- Do we need the Evidence (marginal likelihood) values $p(Y|k)$ absolutely, or only relatively?
- Jumping between models as an aid to mixing (c.f. simulated tempering: mixing may be better in the ‘other’ model)
- Are samplers for individual models already written and tested?
- Are standard strategies like split/merge likely to work?
- Trade-off between remembering and forgetting $\theta_k$ when leaving model $k$
The different ways that models can interconnect

- completely unrelated
- nested in some irregular way
- mixture models (nesting and exchangeability)
- variable selection (factorial structure)
- graphical models (possibly with constraints such as decomposability)
Methodological connections and extensions

- Point process representations and samplers (Geyer & Møller, 1994, Stephens, 2000)
- Delayed rejection (Green & Mira, 2001)
- Connections between reversible jump and continuous time birth-and-death samplers (Cappé, Robert & Rydén, 2001)
- Composite model space framework (Godsill, 2001)
- Efficient construction of proposals (Brooks, Giudici & Roberts, 2003)
- Automatic RJ sampler (Hastie, 2005)
- Population RJMCMC and Interacting SMC (Jasra et al, 2007/8)
The obvious estimator of $p(k|Y)$ is the empirical frequency

$$\hat{p}(k|Y) = \frac{\#\{ t : k(t) = k \}}{N}$$

the proportion of the MCMC run spent in model $k$, but Bartolucci, Scaccia and Mira (Biometrika, 2006) showed that we can do better using one of several ‘Rao-Blackwellised’ estimators, exploiting the fact that the sampler moves into each model with a probability that is known at the time.

We can also use this in BMA, etc.
The basic idea of the simplest version.

Let $\alpha_{kl}^{(t)}$ be the Metropolis-Hastings acceptance ratio for moving from model $k$ to model $l$ on sweep $t$, or 0 if we are not in $k$ at time $t - 1$ or do not propose such a move. Define

$$\hat{B}_{kl} = \frac{\sum_{t=1}^{N} \alpha_{lk}^{(t)} / n_l}{\sum_{t=1}^{N} \alpha_{kl}^{(t)} / n_k}$$

where $n_k = \#\{ t : k(t) = k \}$ is the number of visits to model $k$. This derives from Meng–Wong type bridge-sampling ideas, somewhat analogous to Chib–Jeliazkov.
Better Bayes factors

Bartolucci, Scaccia and Mira

Computing posterior model probabilities

Groningen, March 2011
In reality, it’s not that easy...

- **Prior model probabilities may be fictional**

  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

  Arbitrariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!
In reality, it’s not that easy...

- Prior model probabilities may be fictional
  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

  Arbitrariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!
In reality, it’s not that easy...

- Prior model probabilities may be fictional
  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

Arbirariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!
In reality, it’s not that easy...

- No chance of passing the test of a sensitivity analysis

In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it’s clear that a test of sensitivity to model probabilities will always fail:

\[
\frac{p^*(k \mid Y)}{p^*(l \mid Y)} = \frac{p(k \mid Y)}{p(l \mid Y)} \times \left( \frac{p^*(k)}{p^*(l)} \div \frac{p(k)}{p(l)} \right)
\]
In reality, it’s not that easy...

- **No chance of passing the test of a sensitivity analysis**

  In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it’s clear that a test of sensitivity to model probabilities will always fail:

  \[
  \frac{p^*(k|Y)}{p^*(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left( \frac{p^*(k)}{p^*(l)} \div \frac{p(k)}{p(l)} \right)
  \]
In reality, it’s not that easy...

- No chance of passing the test of a sensitivity analysis
  In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it’s clear that a test of sensitivity to model probabilities will always fail:

\[
\frac{p^*(k|Y)}{p^*(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left( \frac{p^*(k)}{p^*(l)} \div \frac{p(k)}{p(l)} \right)
\]
In reality, it’s not that easy...

**Improper parameter priors problems**

In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.
In reality, it’s not that easy...

- **Improper parameter priors problems**
  
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.
In reality, it’s not that easy...

- **Improper parameter priors problems**
  
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.
In reality, it’s not that easy...

- **Improper parameter priors problems**
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.
In reality, it’s not that easy...

**Improper parameter priors problems, continued**

In certain circumstances, ideas such as Intrinsic or Fractional Bayes factors, or Expected Posterior priors, can be applied, essentially based on tying together improper priors across different models. These ideas lose much of the appeal of ideal Bayes arguments, have arbitrary aspects, and are not widely accepted.
In reality, it’s not that easy...

- Improper parameter priors problems, continued

In certain circumstances, ideas such as Intrinsic or Fractional Bayes factors, or Expected Posterior priors, can be applied, essentially based on tying together improper priors across different models. These ideas lose much of the appeal of ideal Bayes arguments, have arbitrary aspects, and are not widely accepted.
Model uncertainty? Yes, but do we have to choose?

Model uncertainty is a fact of life.
- When can we quantify it?
- When can we eliminate it?
- When can we accommodate it?

Why choose?
- Prediction
- Scientific understanding
- Presentation
- Policy
- Defence
An illusion of unity?

Is ‘model’ too much of a catch-all?

- different scientific mechanisms
- selection of predictors in regression
- number of components in mixture
- order of AR model
- complexity of polynomial or spline
All the other criteria

- Bayesian hypothesis testing, and ‘alternative-prior carpentry’
- AIC, BIC, DIC, DIC+, MDL, $C_p$
- Decision theory
- Bayesian p-values
- Posterior predictive checks
Acknowledgements

Thanks to all my collaborators on trans-dimensional MCMC problems: Carmen Fernández, Paolo Giudici, Miles Harkness, David Hastie, Matthew Hodgson, Antonietta Mira, Agostino Nobile, Marco Pievatolo, Sylvia Richardson, Luisa Scaccia and Claudia Tarantola.

References and preprints available from

http://www.stats.bris.ac.uk/~peter/Research.html
P.J.Green@bristol.ac.uk

©University of Bristol, 2011