Unit 7: Model Assessment and Comparison
Course Units

- Introduction to Bayesian Statistics
- Prior Distributions
- Simple Models
- Bayesian Asymptotics
- Hierarchical Modeling
- Bayesian Computation
- Model Assessment and Comparison
- Regression Modeling
- Bayesian Nonparametrics
- Survival Analysis and Missing Data
- Clinical Trials and Bayesian Design
Outline of the Unit

Model diagnostics

Robustness and sensitivity

Model comparison
  Bayes factors
  Information criteria
  Other approaches
  Reversible Jump MCMC

Bayesian testing

The Bayesian Method
Overview

- Model diagnostics are strategies for assessing how well the model fits the data (a question of statistical adequacy).
- This is NOT the same as, nor related to, MCMC diagnostics, which aim to assess whether one’s MCMC is producing samples from the model being implemented (a question of simulation adequacy).
- However, note that poor MCMC performance or bugs in your code might cause the model to fit poorly.
Residual diagnostics

- We can often define quantities analogous to classical diagnostics. For example:
  - Bayesian residual: \( r_i = y_i - E(Y_i|X, Y) \)
  - Standardized Bayesian residual: \( d_i = \frac{y_i - E(Y_i|X, Y)}{\text{Var}(Y_i|X, Y)} \)
  - Cross-validated Bayesian residual: \( r_i = y_i - E(Y_i|X, Y_{-i}) \)

- Basically, one can define quantities in the Bayesian framework, often based on posterior expectations and predictive distributions, that can be used in a sensible manner to check the model.

- One can always compare \( \{y_i\} \) to \( \{E(Y_i|Y)\} \) or \( \{E(Y_i|Y_{-i})\} \) even if subtraction does not make sense.
Using the posterior predictive distribution

The basic idea is that the observations should be consistent with the posterior predictive distribution. There are several ways to consider this question.

1. Draw samples of data, $y^{\text{rep}}$ (which are replicates like $y$) from $P(y^{\text{rep}}|y)$ under the posterior predictive distribution and compare to the observations in aggregate graphically (see GCSR for some examples)

2. Conditional Predictive Ordinate (CPO): the density of the posterior predictive distribution evaluated at an observation:

$$P(y_i | y_{-i}) = \int P(y_i | \theta, y_{-i})P(\theta | y_{-i})d\theta$$

- Estimate as $\frac{1}{T} \sum_t P(y_i | \theta^t)$ where $\theta^t \sim P(\theta | y_{-i})$ (if observations are conditionally independent)
- Small values of CPO indicate poor fit (or chance!); one can also compute tail areas
- Plot CPO versus $i$ to look for outliers
- Basing this on C-V is ideal, but could be done with $P(\theta | y)$

3. Posterior predictive p-values: idea is to see if the value of an observed test statistic is extreme relative to the posterior distribution of the statistic.
Posterior predictive $p$-value

Basic idea is to define a test quantity (or discrepancy measure) which is a scalar summary of parameters and data, and then examine the tail-area probability of the test quantity. Let $T(y, \theta)$ represent the discrepancy measure.

Then:

**Classical $p$-value** for the test statistic, $T(y)$, is

$$Pr(T(y^{\text{rep}}) \geq T(y)|\theta)$$

where the probability is taken over the distribution of $y^{\text{rep}}$ with $\theta$ fixed.

**Posterior Predictive $p$-value** for a test statistic, $T(y, \theta)$, is

$$Pr(T(y^{\text{rep}}, \theta) \geq T(y, \theta)|y)$$

which is calculated as:

$$\int \int I_{T(y^{\text{rep}}, \theta) \geq T(y, \theta)} P(y^{\text{rep}} | \theta) P(\theta | y) d\theta dy^{\text{rep}}$$

How can we use simulation to do the double integral?
The test statistic

Choosing the test statistic: Choose quantities that measure features of the data that are not directly addressed by the probability model and that are relevant to the scientific goal of the modeling.

- ranks of the sample (e.g., the minimum $y$).
- zero inflation with replicated Poisson or binomial data
- $\chi^2$ discrepancy measure: $\sum_i \frac{(y_i - E(y_i|\theta))^2}{\text{Var}(y_i|\theta)}$

Remarks:

- Large values (close to 1) or small values (close to 0) indicate the model is suspect.
- The Bayesian p-value should not be compared across different models.
- The Bayesian p-value is not the probability that the model is correct. It provides information about model adequacy.
- One could do this in a cross-validatory framework to guard further against overfitting.
- Bayarri and Berger (2000; JASA) and Robins et al. (2000; JASA) found that the distribution of the p-values under the true model is not uniform, raising doubts about this approach.
Overview

- Both prior and likelihood are subjective.
- One can examine sensitivity to each as sensitivity analysis.
- Or, the model comparison approach considers changes to either or both as competing models.
Robust models

- A simple way to make a model more robust to unusual observations is to make sure the likelihood is not overly influenced by such observations. Heavy-tailed/over-dispersed distributions are one way to do this, e.g., using:
  - \( t \) distribution rather than a normal distribution.
  - negative binomial in place of Poisson
  - beta-binomial in place of binomial
  - Dirichlet process (nonparametric) distributions have gotten a lot of attention recently

- Often this can take the form of expanding the model (e.g., the \( t \) is a normal with an additional scale parameter) and estimating the new overdispersion parameter(s).

- One may often want to consider robust models for random effects as well.
  - What could we consider for the Avandia analysis?
Prior sensitivity

The basic idea is to consider other priors and examine sensitivity of the conclusions from the modeling effort.

- change the variance
- change the distributional family
- change the bounds

Other more formal approaches include so-called epsilon-contamination priors.
Assessing sensitivity

- We can assess:
  - Sensitivity to the prior or likelihood.
  - Sensitivity to case deletion (influential observations).

- One can often re-use the posterior samples from the original model, provided the altered model is not substantially different (usually in terms of tail probabilities).
  - The possibilities are importance sampling and the weighted bootstrap (sampling-importance resampling).
  - CAUTION: if the prior is influential or there are very influential observations, the importance sampling weights can be very variable, so these should be checked as one proceeds.
Importance sampling revisited

Let $g(\theta)$, the importance sampling density be the original posterior, $P(\theta|y)$. Suppose we alter the model, and have $P^*(\theta|y)$ where the parameters are equivalent. We can then use $P(\theta|y)$ as our importance sampling density, $g(\theta)$.

The weight function is then

$$w(\theta) = \frac{P^*(y|\theta)P^*(\theta)}{P(y|\theta)P(\theta)}$$

This gives us

$$E^*(h(\theta)|y) \approx \frac{\sum_t h(\theta^t)w(\theta^t)}{\sum_t w(\theta^t)}$$

The modified expectation is just based on reweighting the samples based on their posterior probability under the new and old models.

For case deletion, we have $P(\theta|y_i)$ ∝ $\frac{P(\theta|y)}{P(y_i|\theta)}$ so we have weights

$$w(\theta) = \frac{1}{P(y_i|\theta)}$$

For case deletion or for changing the prior not to substantially, we should have $P^*(\theta|y) \approx P(\theta|y)$. 
Weighted bootstrap

Suppose instead we want a sample from the new posterior?

1. Suppose we have a sample from the posterior.
2. Let \( w_t = \frac{P^*(y|\theta^t)P^*(\theta^t)}{P(y|\theta^t)P(\theta^t)} \) and \( q_t = \frac{w_t}{\sum_{i=1}^{T} w_i} \)
3. Now draw \( \theta^* \) from the discrete distribution over \( \{\theta^1, \theta^2, \ldots, \theta^T\} \) which places mass \( q_t \) at \( \theta^t \). Then \( \theta^* \) is a sample from \( P^*(\theta|y) \) with the approximation improving as \( T \to \infty \).

This is a weighted bootstrap because rather than resampling from the set \( \{\theta^1, \theta^2, \ldots, \theta^T\} \) with equal probabilities of selection, we resample some points more than others because of the unequal weighting. Also known as sampling-importance resampling (SIR). Reference is Smith and Gelfand, 1992, American Statistician.
Why does it work?

\[
\hat{P}(\theta^* \leq a | y) = \sum_{t=1}^{T} q_t I(-\infty < \theta^t \leq a)
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} w_t I(-\infty < \theta^t \leq a)
\]

as \( T \to \infty \)

\[
E \left[ P^*(y|\theta)P^*(\theta) \frac{P^*(\theta|y)P(\theta)}{P(y|\theta)P(\theta)} I(-\infty < \theta \leq a) \right]
\]

\[
E \left[ \frac{P^*(y|\theta)P^*(\theta)}{P(y|\theta)P(\theta)} \right]
\]

\[
\int_{-\infty}^{a} \left[ \frac{P^*(y|\theta)P^*(\theta)}{P(y|\theta)P(\theta)} P(\theta|y) \right] d\theta
\]

\[
= \int_{-\infty}^{a} P^*(\theta|y) d\theta.
\]

- As with importance sampling, need \( g(\theta) = P(\theta|y) \approx P^*(\theta|y) \) or need very large \( T \).
Bayes factor overview

- The Bayes factor is a quantity used to compare models and test hypotheses in the Bayesian framework (Kass R, Raftery AE, 1995. Bayes Factors, *JASA*, 90, 773-795.)
- Suppose we have two candidate parametric models $M_1$ versus $M_2$ for data $y$, and the two models have respective parameter vectors $\theta_1$ and $\theta_2$. Suppose we specify priors as $P(\theta_1)$ and $P(\theta_2)$.
- The Bayes factor in favor of $M_1$ is defined as the odds ratio of the posterior to prior odds of $M_1$ and $M_2$:

$$BF = \frac{P(M_1|y)/P(M_1)}{P(M_2|y)/P(M_2)}$$
More details

\[ BF = \frac{P(M_1 | y)}{P(M_2 | y)} \]

By Bayes theorem:

\[ P(M_1 | y) = \frac{P(y | M_1)P(M_1)}{P(y | M_1)P(M_1) + P(y | M_2)P(M_2)} . \]

Substituting these into the definition, we get the Bayes factor in favor of \( M_1 \) to be

\[ BF = \frac{P(y | M_1)}{P(y | M_2)} \]

The marginal distribution (marginal likelihood, prior predictive) for \( y \) is:

\[ P(y | M_k) = \int P(y | \theta_k, M_k)P(\theta_k | M_k)d\theta_k, \ k = 1, 2. \]

I.e., how likely are the data based on each model, integrating over the uncertainty in the parameters, as represented in the prior.
Large values of $BF$ favor $M_1$.

<table>
<thead>
<tr>
<th>$\log_{10}(BF)$</th>
<th>$BF$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 0.5</td>
<td>$1 \leq BF \leq 3.2$</td>
<td>Weak</td>
</tr>
<tr>
<td>0.5 - 1.0</td>
<td>$3.2 &lt; BF \leq 10$</td>
<td>Substantial</td>
</tr>
<tr>
<td>1.0 - 2.0</td>
<td>$10 &lt; BF \leq 100$</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt; 2.0$</td>
<td>$BF &gt; 100$</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$2\log(BF)$</th>
<th>$BF$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td>$1 \leq BF \leq 3$</td>
<td>Weak</td>
</tr>
<tr>
<td>2 - 6</td>
<td>$3 &lt; BF \leq 20$</td>
<td>Positive</td>
</tr>
<tr>
<td>6 - 10</td>
<td>$20 &lt; BF \leq 150$</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt; 10$</td>
<td>$BF &gt; 150$</td>
<td>Very Strong</td>
</tr>
</tbody>
</table>
Computation

We want to compute:

\[ P(y \mid M) = \int P(y \mid \theta, M)P(\theta \mid M)d\theta \]

- **Sampling from the prior**: Sample \( \theta^l, l = 1, \ldots, L \) from \( P(\theta \mid M) \), the prior under \( M \):

\[ P(y \mid M) \approx \frac{1}{L} \sum_{l} P(y \mid \theta^l, M) \]

where \( \theta^l \) is a draw from \( P(\theta \mid M) \). Do this for both \( M_1 \) and \( M_2 \) and calculate BF. However, this is likely to be an inefficient estimator. **Why?**
Computation (2)

- **Sampling from the posterior**: Use samples from the posterior distribution rather than from the prior (Newton MA, Raftery AE, 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *JRSSB*, 56, 1-48).

\[
P(y|M) \approx \left[ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{P(y|\theta^t, M)} \right]^{-1}
\]

where \( \theta^t \) is a draw from \( P(\theta|M, y) \). This is the harmonic mean of the posterior sample likelihoods. This is efficient but unstable. Why?

- An alternative is to define a mixture of the posterior and the prior as an importance sampling density, which avoids the harmonic mean.
Computation (3)

- **Sampling from the posterior (2):** Gelfand and Dey (JRSSB 56:501) derive the following approach. Start with the identity

\[
\frac{1}{P(y|M)} = \int \frac{g(\theta)}{P(y|\theta, M)P(\theta|M)} \, P(\theta|y, M) \, d\theta
\]

which holds for any density \( g(\theta) \). This suggests estimating the following using the posterior sample:

\[
P(y|M) \approx \left[ \frac{1}{T} \sum_t \frac{g(\theta^t)}{P(y|\theta^t, M)P(\theta^t|M)} \right]^{-1}.
\]

- From the perspective of importance sampling, we should choose \( g(\theta) \) to approximately match the posterior, so one possibility is a \( t \) or normal estimated from the sample \( \{\theta^t\}_{t=1}^T \).

- Theory suggests the tails of \( g \) should be relatively thin, so a normal or \( t \) with large df might be a good choice.

- This approach seems to do pretty well except if the parameter space is very high-dimensional.
Remarks

1. We integrate over the parameter space instead of maximizing over it.
2. The $BF$ does **NOT** require nested models.
3. The $BF$ has a nice interpretation of telling you how the data change the odds in favor of model 1.
4. The $BF$ reduces to the likelihood ratio test in the case of a simple vs. simple hypothesis test.
5. $BF$ is defined **only** when proper prior distributions are defined for $\theta$. The $BF$ is **NOT** defined for improper prior distributions as it would involve calculating an expectation with respect to something that is not a distribution.
6. The $BF$ may be sensitive to the choice of prior distribution and/or the choice of prior hyperparameters. For example, if $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$, the $BF$ may change markedly as $\sigma_0^2 \to \infty$.
7. Computation of $BF$ can be a real pain, and a variety of approaches have been suggested (see Carlin & Louis).
Bayesian Information Criteria (BIC). (Schwarz, 1978; Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464) showed that for large sample size $n$, one can approximate $-2 \log BF$ by:

$$
\Delta BIC = W - (p_2 - p_1) \log n
$$

where $p_m$ is the number of parameters in model $M_m$; and $(p_1 - p_2) \log n$ acts as a penalty term which accounts for differences in size between models. Thus,

$$
BF \approx \exp \left( -\frac{1}{2} \Delta BIC \right).
$$

Therefore, to compare models, choose the model that minimizes

$$
BIC_k = -2 \log (\text{max likelihood}_k) + p_k \log n
$$

BIC is also justified by an asymptotic argument that it is consistent in choosing the correct model (or the model closest in K-L to the correct model).
Some issues with BIC

How to choose $p$ and $n$ in hierarchical models? Suppose we have a longitudinal model where we have $m_i$ observations on the $i$th individual and a total of $J$ individuals?

- Questions:
  1. Suppose we have $J$ random effects, one for each individual. What is $p$?
  2. What is $n$?
BIC, AIC and BF

- BIC is defined for improper priors (note that it depends on maximizing the likelihood and is justified asymptotically, hence with diminishing influence of the prior)

- Recall that AIC is

\[-2 \log (\text{max likelihood}_k) + 2p_k\]

- In general, BIC favors simpler models than AIC. This relates to the natural Bayesian penalty on complexity that we will discuss later. Basically Bayesian methods account for the fact that more complicated models have large parts of the parameter space with parameter values that do poorly in explaining the data. Bayesian models integrate over the parameter space rather than maximizing.

- Both BIC and AIC share the difficulty involved in estimating the number of parameters.
DIC

**Deviance Information Criteria (DIC).** Spiegelhalter et al. (2002) JRSSB, 64:583.

Define the *deviance statistic* as

\[ D(\theta) = -2 \log P(y|\theta) \]

where \( P(y|\theta) \) is the likelihood.

- Fit of the model is assessed by the posterior expectation of the deviance, \( \bar{D} = E(D|y) \).
  - See GCSR for some discussion of why the deviance is a good measure of model fit, based on its relationship to Kullback-Leibler information.
- Complexity of the model is summarized by the effective number of parameters, \( p_D \),

\[
p_D \equiv E(D|y) - D(E(\theta|y))) = \bar{D} - D(\bar{\theta})
\]

- Note that this involves the posterior means only of the parameters directly involved in the likelihood.
- In normal regression models, \( p_D \) is the trace of the hat matrix.
- Note that \( p_D \) depends on the data, which makes sense, as the amount of shrinkage in a model is estimated based on the data.
An information criterion

By analogy with other information criteria, e.g., AIC, the expected predictive deviance can be estimated by

$$ D(\bar{\theta}) + 2p_D = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta}) $$

where $p_D$ is a measure of the optimism involved in estimating the deviance for new data based on the deviance for the fitted data. Hence, the idea is to choose the model that minimizes

$$ DIC_k = \bar{D}_k + p_{D,k} $$

- DIC is very popular, in part because it is easily calculated from MCMC output.
- One drawback is that it varies with parameterization.
- What is a substantial difference in DIC between models?
  - We need an idea of the sampling variance of DIC. One approach is to use blocks of the posterior iterates.
  - Spiegelhalter et al. (2002) suggest a difference of 3 or more for substantive difference.
Motivating DIC

We can justify this asymptotically in the non-hierarchical setting: Let’s expand $D(\theta)$ around the posterior mode,

$$D(\theta) \approx D(\hat{\theta}) - 2L'(\hat{\theta})(\theta - \hat{\theta}) - (\theta - \hat{\theta})^T L''(\hat{\theta})(\theta - \hat{\theta})$$

where $L = -D(\theta)/2$. By the BCLT, we have

$$P(\theta|y) \sim \mathcal{N}(\hat{\theta}, (-L''(\theta))^{-1})$$

The last term is approximately $\chi^2_p$, so we have

$$D(\theta) \approx D(\hat{\theta}) + \chi^2_p$$

Taking expectations,

$$p \approx E(D) - D(\hat{\theta}) \approx E(D) - D(\bar{\theta})$$
Continuous families

- GCSR emphasize continuous families of models rather than discrete choices.
- In my personal experience, one often defines more complicated models as expansions of current models and can do model comparison based on posterior inference on variance components, regression coefficients and other parameters.
- However, there are lots of circumstances where one might have completely different models or parameterizations that one wants to compare.
Cross-validation

- Very useful for prediction problems
- Essentially, it involves assessing $P(\tilde{y}|y, M_k)$.
  - evaluate mean squared error, bias, coverage and length of credible intervals for quantities of interest
- One could also define a measure in terms of probability density (i.e., deviance) such as
  \[
  LS^CV_k = \log P(\tilde{y}|y, M_k) \approx \log \frac{1}{T} \sum_t P(\tilde{y}|y, \theta^t_k, M_k)
  \]
  - This looks a lot like a C-V version of a Bayes factor, with the crucial distinction that we don’t have the sensitivity to the prior.
  - See Draper and Krnjajic (UCSC tech report); who note connections with DIC
  - Note that unlike MSE, this assesses how well we have characterized uncertainty as well

- Assesses model adequacy in a concrete way
Posterior model probabilities

The coherent Bayesian approach to model selection and comparison is to use the posterior model probabilities, $P(M_k|y), \ k = 1, \ldots, K$.

$$P(M_k|y) = \frac{P(y|M_k)P(M_k)}{\sum_i P(y|M_i)P(M_i)} \propto P(y|M_k)P(M_k)$$

$$= \int P(y|\theta_k, M_k)P(\theta_k|M_k)P(M_k)d\theta_k$$

$P(y|M_k)$ is the marginal likelihood, i.e., the prior predictive distribution, or the normalizing constant in the single model case.
Model averaging

- As statisticians, we prefer not to fix parameters in advance; instead we want to account for their uncertainty in making inference.

- Similarly, choosing a single model is like fixing a parameter; we don’t account for uncertainty in the model.

- Model averaging is a way to account for this uncertainty, averaging over the possible models, just as we average over possible parameter values.

- This is primarily done in the context of prediction, where different models can generate predictions for the same system.
  - If the loss function is $\log P(\tilde{y}|y)$ then model averaging can be shown to be optimal for prediction.

- Often for estimation, the meaning of the parameters changes between models, so model averaging may not make sense. But in some cases (e.g., with regression parameters) it may still make sense.

- In some cases, we may wish to treat model uncertainty as a sensitivity analysis, rather than averaging across very different model structures.
Model averaging (2)

Suppose we want to estimate $\phi$, which is well-defined (and has the same meaning!) for all of our models.

$$P(\phi|y) = \sum_k P(\phi|M_k, y)P(M_k|y)$$

$$= \sum_k \left[ \int P(\phi|\theta_k, M_k, y)P(\theta_k|M_k, y)d\theta_k \right] \cdot P(M_k|y)$$

- Obviously we need the posterior model probabilities to do this.
- If the number of models is large, it can be difficult to do all the necessary computations.
Reversible Jump MCMC

- Reversible jump MCMC was introduced by Green (1995), *Biometrika*.
- RJMCMC allows one to move between parameter spaces with different dimensions.
- A key application is in model comparison: the chain samples both models and parameters nested within models.
- Let $M_k$, $k = 1, \ldots, K$ be the candidate models and $\theta_k$ the parameter vector for $M_k$.
- RJMCMC sets up a Markov chain on $\{M_k\} \times \{\Theta_k\}$ that satisfies detailed balance.
Detailed balance

- Additional random variables are introduced to ensure dimension mapping.
- One needs a transition matrix for moving between the discrete models, $J(k^*|k)$.
- If one is proposing to move from $k$ to $k^*$, one introduces an auxiliary random variable, $u$, with jumping distribution $J(u|k, k^*, \theta)$.
- Then $(\theta_{k^*}, u^*) = g_{k^*}(\theta_k, u)$ where $d_k + \text{dim}(u) = d_{k^*} + \text{dim}(u^*)$ and $g$ is a deterministic function that relates the parameters of model $k$ to those of model $k^*$. 
RJMCMC algorithm

- Starting in state \((k, \theta_k)\), (i.e., model \(k\)) propose to move to model \(k^*\) with proposal density \(J(k^*|k)\) and propose auxiliary variable from \(J(u|k, k^*, \theta)\).
- Determine the new model parameters, \((\theta_{k^*}, u^*) = g_{k,k^*}(\theta_k, u)\)
- Calculate the acceptance ratio

\[ r = \frac{P(Y|\theta_{k^*}, M_{k^*})P(\theta_{k^*}|M_{k^*})P(M_{k^*})J(k^*|k)J(u^*|k^*, k, \theta^*)}{P(Y|\theta_k, M_k)P(\theta_k|M_k)P(M_k)J(k^*|k)J(u|k, k^*, \theta)}} \left| \frac{\partial g_{k,k^*}(\theta_k, u)}{\partial (\theta_k, u)} \right| \]

One generally mixes proposals to move to new models with proposals for new parameter values within the current model.
Dimension Matching

Consider a simple changepoint model with $y_t \sim \mathcal{N} (\psi_1, 1)$ for $t \le c$ and $y_t \sim \mathcal{N} (\psi_2, 1)$ for $t > c$. What could we use for $g_{k,k^*}$ and $J(u|k, k^*, \theta)$?

To generate a proposal from the simple model, $y_t \sim \mathcal{N} (\psi, 1)$, given the larger model, let

$$\psi^* = (\psi_1 + \psi_2)/2$$

implicitly: $u^* = (\psi_1 - \psi_2)/2$

To go the other way, we need the auxiliary variable to match the dimensions. One proposal might be $u \sim J(u|k, k^*, \theta_k)$ (for some density $J$) with $g_{k,k^*}$ such that

$$\psi_1^* = \psi + u \quad \psi_2^* = \psi - u$$

The trick is to figure out good parameter values in the new model when one proposes to change models. This may be difficult, and mixing in RJMCMC is often an issue.
RJMCMC output

- If you can get it, working with the output is really handy.
- One can calculate the proportion of time in each model to estimate $P(M_k|y)$. This can be used for model averaging or to decide which are competitive models, including BF calculations.
- $P(\theta_k|M_k, y)$ is just estimated from the iterations that are sampled from model $k$. Note that if you have many models, there may be few iterations from a given model, so the chain likely needs to be run for a long time to get adequate convergence and mixing for each model.
Overview

In general, Bayesians are not interested in testing, even when it can be put into a Bayesian framework. Instead, credible intervals and posterior model probabilities are of interest.

There are a few ways to place testing in a Bayesian framework.

Point null hypotheses:

- These are difficult to deal with since for continuous random variables. Why?
  - One possibility is to have the prior put a point mass at $\theta_0$.
  - An alternative is to define an interval of lack of scientific significance and calculate $P(\theta \in [\theta_0 - \epsilon_1, \theta_0 + \epsilon_2]|y)$.
- One-sided tests sometimes produce p-values that are essentially posterior probabilities, e.g., $p \approx P(\theta > 0|y)$. 
Model comparison

Hypothesis testing as model comparison:

- A reasonable approach is to define the null and alternative hypotheses as alternative models and then choose between them based on posterior model probabilities or the Bayes factor.
- In either case, the Bayesian approach does not specify a type I error in advance and there is not the usual asymmetry between null and alternative that one sees in classical testing.
- To differentially weight type I and type II error, one needs to use a decision theoretic framework and include that consideration in the loss function. One can then choose the hypothesis that minimizes posterior risk.
Using the Bayes factor for testing

\[ BF = \frac{P(y|H_0)}{P(y|H_1)} \]

The interpretation is of the ratio of posterior to prior odds of the two hypotheses.

Let

\[ H_0 : \theta \in \Theta_0 \]
\[ H_1 : \theta \in \Theta_1 \]

\[ BF = \frac{P(y|H_0)}{P(y|H_1)} = \frac{\int_{\Theta_0} P(y|\theta, H_0)P(\theta|H_0)d\theta}{\int_{\Theta_1} P(y|\theta, H_1)P(\theta|H_1)d\theta} \]
Comments on BF

- This is analog of the likelihood ratio test, except that we integrate over the unknown parameters, weighted by prior and likelihood, rather than finding the ratio of the maximized likelihoods.
- We don’t need nested models.
- For a simple vs. simple test, the BF is the likelihood ratio test, but without the $p$-value.
- Remember that we cannot use improper priors and non-informative, but proper, priors may not be a good choice.
- Because null and alternative models are on equal footing, with a large sample size, one might get a very small $p$-value but have the BF vote in favor of the null as being more likely than the alternative. In the Bayesian setting, the alternative is treated as a legitimate model, not just as a direction of departure from the null.
Overview

1. Decide on the prior - $P(\theta)$
   - $P(\theta)$ expresses what is known about $\theta$ prior to observing $y$

2. Decide on the likelihood - $P(y|\theta)$
   - describes the process giving rise to the data in terms of unknown $\theta$

3. Apply Bayes’ Theorem to derive the posterior - $P(\theta|y)$
   - $P(\theta|y)$ expresses what is known about $\theta$ after observing $y$

4. Inference statements are derived from the posterior distribution, $P(\theta|y)$
   - determine functions of parameters and/or predictive quantities that address scientific questions
   - point estimates, interval estimates, plots of posterior, probability of hypotheses
   - there is no distinction in principle between estimation and prediction; all unknown quantities are treated as random variables
Graphically...