

Chapter 9:

Bayesian Assessment: Hypothesis and Models

April 5, 2006

Introduction

In classical hypothesis testing, a null hypothesis $H_0 : \theta \in \theta_0$ and an alternative hypothesis $H_1 : \theta \in \theta_1$ are specified.

The choice between these hypotheses is governed by looking at the distribution of a test statistic $T(\mathbf{Y})$ which is a function of data \mathbf{Y} , under H_0 , and the so called *p-value* for the test, defined as

$$\Pr[T(\mathbf{Y}) \text{ at least as extreme as the value observed} | H_0] \quad (1)$$

H_0 will be accepted (rejected - in which case H_1 is accepted) if the *p-value* is large (small) enough, or one may quote the *p-value* and leave things there.

What one is really interested in is not (1), but rather

$$\Pr[H_0 | T(\mathbf{Y}) = t(\mathbf{y})] \quad (2)$$

This is conceptually more straightforward and is the way a Bayesian formulates the problem.

Here a brief introduction to Bayesian hypothesis testing is given, with emphasis on the Bayes factor.

An excellent review about the Bayes factor is given by Kass and Raftery (1995)

Simple Hypotheses

In the simple hypothesis testing problem the likelihoods are completely specified because the parameters are given particular values.

Thus assume that under H_0 the data vector \mathbf{y} has likelihood

$$p(\mathbf{y} | \boldsymbol{\theta} = \boldsymbol{\theta}_0)$$

and under H_1 ,

$$p(\mathbf{y} | \boldsymbol{\theta} = \boldsymbol{\theta}_1)$$

It is not necessary that the likelihoods under the two hypotheses belong to the same family.

One could write that under H_1 , $p(\mathbf{y} | \boldsymbol{\Theta} = \boldsymbol{\Theta}_1)$. We will use the same family of likelihoods for simplicity of notation.

Let w be a binary random variable such that

$$w = \begin{cases} 0 & \text{if } H_0 \text{ is true,} \\ 1 & \text{if } H_1 \text{ is true.} \end{cases}$$

The prior probabilities associated with the hypotheses are

$$p(H_0) = \Pr(w = 0)$$

and

$$p(H_1) = \Pr(w = 1)$$

with

$$p(H_0) + p(H_1) = 1.$$

Now, using Bayes theorem, the posterior probability associated with H_0 is

$$\Pr(H_0|y) = \Pr(w = 0|y) = \frac{\Pr(w = 0)p(y|w = 0)}{p(y)} \quad (3)$$

and similarly

$$\Pr(w = 1|y) = \frac{\Pr(w = 1)p(y|w = 1)}{p(y)} \quad (4)$$

In these expressions,

$$p(y) = \Pr(w = 0)p(y|w = 0) + \Pr(w = 1)p(y|w = 1)$$

is the marginal probability of the data, assumed nonzero.

The ratio

$$\frac{\Pr(w = 0|y)}{\Pr(w = 1|y)}$$

is called the *posterior odds ratio* of H_0 to H_1 , and

$$\frac{\Pr(w = 0)}{\Pr(w = 1)}$$

is called the *prior odds ratio*. The quantity

$$B = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\Pr(w = 0|y)}{\Pr(w = 1|y)} \frac{\Pr(w = 1)}{\Pr(w = 0)} \quad (5)$$

is known as the *Bayes factor* in favour of H_0 .

The Bayes factor can sometimes be interpreted as the "odds for H_0 to H_1 that are given by the data".

This is a valid interpretation when the hypotheses are simple, for then, from (5)

$$B = \frac{p(\mathbf{y}|w = 0)}{p(\mathbf{y}|w = 1)} \quad (6)$$

which is the *likelihood ratio* of H_0 to H_1 .

Composite hypotheses

We consider again two mutually exhaustive hypotheses

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\theta}_0 \text{ and } H_1 : \boldsymbol{\theta} \in \boldsymbol{\theta}_1$$

with prior probabilities as before.

Now, the equivalent to expressions (3) and (4)

$$\Pr(w = 0|\mathbf{y}) = \frac{\Pr(w = 0)p(\mathbf{y}|w = 0)}{p(\mathbf{y})}$$

we have

$$\Pr(w = 0|\mathbf{y}) = \frac{\Pr(w = 0) \int_{\boldsymbol{\theta} \in \boldsymbol{\theta}_0} p(\boldsymbol{\theta}|w = 0)p(\mathbf{y}|w = 0, \boldsymbol{\theta})d\boldsymbol{\theta}}{p(\mathbf{y})} \quad (7)$$

and

$$\Pr(w = 1|\mathbf{y}) = \frac{\Pr(w = 1) \int_{\boldsymbol{\theta} \in \boldsymbol{\theta}_1} p(\boldsymbol{\theta}|w = 1)p(\mathbf{y}|w = 1, \boldsymbol{\theta})d\boldsymbol{\theta}}{p(\mathbf{y})} \quad (8)$$

The terms $p(\boldsymbol{\theta}|w = 0)$ and $p(\boldsymbol{\theta}|w = 1)$ are the prior probabilities for $\boldsymbol{\theta}$ under H_0 and H_1 , respectively.

Note that the second term in the numerator of the right hand sides is

$$\int_{\boldsymbol{\theta} \in \boldsymbol{\theta}_i} p(\boldsymbol{\theta}|w = i)p(\mathbf{y}|w = i, \boldsymbol{\theta})d\boldsymbol{\theta} = p(\mathbf{y}|w = i), \quad i = 0, 1$$

which is an *averaged likelihood* with the prior pdf of $[\boldsymbol{\theta}|w = i]$ serving as the weighting function. The Bayes factor now is

$$\begin{aligned} B &= \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\int_{\boldsymbol{\theta} \in \boldsymbol{\theta}_0} p(\boldsymbol{\theta}|w = 0)p(\mathbf{y}|w = 0, \boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \boldsymbol{\theta}_1} p(\boldsymbol{\theta}|w = 1)p(\mathbf{y}|w = 1, \boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{p(\mathbf{y}|w = 0)}{p(\mathbf{y}|w = 1)} \end{aligned} \quad (9)$$

Intuitively, the Bayes factor provides a measure of whether the data have increased or decreased the odds on H_0 relative to H_1 . However, when the hypotheses are composite, the Bayes factor cannot be regarded as a measure of the relative support for the hypotheses provided solely by the data.

Some comments concerning (9) are in order.

1. The first one is that the models specified by the two hypotheses do not need to be nested.
2. Secondly, the first line of (9) shows that problems arise with the Bayes factor if the prior probabilities $p(\theta|w = i)$, $i = 0, 1$, are not proper.

One may wish to assign a distribution to these priors, reflecting vague prior knowledge.

For example, a uniform distribution bounded between 0 and c_i would yield

$$p(\theta|w = i) = \frac{1}{c_i}, \quad i = 0, 1$$

such that

$$\int_0^{c_i} p(\theta|w = i) d\theta = 1$$

Then, using (9) the Bayes factor becomes

$$B = \frac{c_1}{c_0} \frac{\int_{\theta \in \theta_0} p(\mathbf{y}|w = 0, \theta) d\theta}{\int_{\theta \in \theta_1} p(\mathbf{y}|w = 1, \theta) d\theta}$$

which depends on the ratio c_1/c_0 .

Thus in this example the Bayes factor would depend on the bounds chosen of the uniform prior for $p(\theta|w = i)$.

Berger and Pericchi (1996) introduce the idea of *intrinsic Bayes factors* in an attempt to circumvent the dependence on the prior and in order to allow use of non-informative prior distributions in general.

Computing the Bayes factor using *MCMC*

Newton and Raftery (1994) show how the Bayes factor can be computed using standard *MCMC* output. Consider the posterior distribution of θ under model i

$$p(\theta|\mathbf{y}, w = i) = \frac{p(\mathbf{y}|\theta, w = i)p(\theta|w = i)}{p(\mathbf{y}|w = i)}$$

Then

$$\frac{p(\boldsymbol{\theta}|w = i)}{p(\mathbf{y}|w = i)} = \frac{p(\boldsymbol{\theta}|\mathbf{y}, w = i)}{p(\mathbf{y}|\boldsymbol{\theta}, w = i)}$$

Integrating both sides with respect to $\boldsymbol{\theta}$ yields

$$\frac{1}{p(\mathbf{y}|w = i)} \int p(\boldsymbol{\theta}|w = i) d\boldsymbol{\theta} = \int \frac{1}{p(\mathbf{y}|\boldsymbol{\theta}, w = i)} p(\boldsymbol{\theta}|\mathbf{y}, w = i) d\boldsymbol{\theta}$$

With a proper prior distribution, $\int p(\boldsymbol{\theta}|w = i) d\boldsymbol{\theta} = 1$ and one obtains

$$[p(\mathbf{y}|w = i)]^{-1} = E[p^{-1}(\mathbf{y}|\boldsymbol{\theta}, w = i)]$$

or

$$[p(\mathbf{y}|w = i)] = E^{-1}[p^{-1}(\mathbf{y}|\boldsymbol{\theta}, w = i)] \quad (10)$$

A Monte Carlo estimate of (10) is

$$\hat{p}(\mathbf{y}|w = i) = \frac{1}{\frac{1}{n} \sum_{j=1}^n p^{-1}(\mathbf{y}|\boldsymbol{\theta}^{(j)}, w = i)} \quad (11)$$

where n is the number of Gibbs samples and $\boldsymbol{\theta}^{(j)}$ is the j^{th} MCMC draw from the posterior distribution $[p(\boldsymbol{\theta}|\mathbf{y}, w = i)]$.

Programming the Newton-Raftery Bayes factor

A little caution must be exercised in the actual computation of (11) to avoid over/underflows. One strategy could be as follows. Let

$$\begin{aligned} m &= \frac{1}{n} \sum_{j=1}^n p^{-1}(\mathbf{y}|\boldsymbol{\theta}^{(j)}, w = i) \\ &= \frac{1}{n} \sum_{j=1}^n S_i^{(j)} \end{aligned}$$

where $S_i^{(j)} = p^{-1}(\mathbf{y}|\boldsymbol{\theta}^{(j)}, w = i)$.

One can store in a file for each Gibbs iterate, $\ln S_i^{(j)}$. Then, since

$$\exp(x) = \exp(x - c + c) = \exp(x - c) \exp c,$$

we can write m in the following form:

$$m = \frac{1}{n} \sum_{j=1}^n \exp(\ln S_i^{(j)} - c) \exp c$$

where c is the largest value of $\ln S_i^{(j)}$. Taking logarithms yields

$$\ln m = \ln \left[\frac{1}{n} \sum_{j=1}^n \exp(\ln S_i^{(j)} - c) \right] + c$$

and

$$\ln[\hat{p}(\mathbf{y}|w = i)] = -\ln m$$

Rule of Thumb

With this setup, if we interpret model 1 as the null model, then:

- If $B(\mathbf{x}) \geq 1$ then model 1 is supported
- If $1 > B(\mathbf{x}) \geq 10^{-1/2}$ then minimal evidence against model 1.
- If $10^{-1/2} > B(\mathbf{x}) \geq 10^{-1}$ then substantial evidence against model 1.
- If $10^{-1} > B(\mathbf{x}) \geq 10^{-2}$ then strong evidence against model 1.
- If $10^{-2} > B(\mathbf{x})$ then decisive evidence against model 1.

Other remark

Some examples of determining the Bayes Factor in WinBugs for a variable mean can be found in Congdon (example 2.2); and more complex models in Congdon Chapter 10.

You also may want to use Carlin and Chib's technique for computing Bayes Factors for competing non-nested regression models reported in Journal of Royal Statistical Society. Series B. vol 57:3 1995.

This technique is implemented in the Pines example in BUGS, and is reported on the Winbugs website under the new examples section.

Deviance Information Criteria:

The deviance information criterion (DIC) has been recently introduced as a means of comparing models .

The DIC uses the posterior expectation of the log likelihood as a measure of model fit.

For a particular model M , the DIC is defined as

$$DIC = 2\bar{D} - D(\bar{\theta}_M), \quad (12)$$

where

$$\begin{aligned} \bar{D} &= -2 \int [\log p(\mathbf{y}|\theta_M)] p(\theta_M|\mathbf{y}, M) d\theta_M \\ &= E_{\theta_M|\mathbf{y}}[D(\theta_M)], \end{aligned} \quad (13)$$

is the posterior expectation of the so-called deviance

$$D(\theta_M) = -2 \log p(\mathbf{y}|\theta_M).$$

The second term in the right hand side of (12) is the deviance evaluated at the posterior mean of the parameter vector θ_M .

In order to motivate (12), consider an expansion of the deviance around the posterior mean $\bar{\theta}_M$, to obtain

$$\begin{aligned} D(\theta_M) &\approx -2 \log p(\mathbf{y}|\bar{\theta}_M) - 2 \left[\frac{\partial \log p(\mathbf{y}|\theta_M)}{\partial \theta_M} \right]_{\theta_M=\bar{\theta}_M} (\theta_M - \bar{\theta}_M) \\ &\quad - (\theta_M - \bar{\theta}_M)' \left[\frac{\partial^2 \log p(\mathbf{y}|\theta_M)}{\partial \theta_M \partial \theta_M'} \right]_{\theta_M=\bar{\theta}_M} (\theta - \bar{\theta}_M). \end{aligned}$$

Taking expectations with respect to the posterior distribution of θ gives the expected deviance

$$\bar{D} \approx D(\bar{\theta}_M) + \text{tr} \left\{ \left[-\frac{\partial^2 \log p(\mathbf{y}|\theta_M)}{\partial \theta_M \partial \theta_M'} \right]_{\theta_M=\bar{\theta}} \text{Var}(\theta_M|\mathbf{y}) \right\} \quad (14)$$

where $D(\bar{\theta}_M) = -2 \log p(\mathbf{y}|\bar{\theta}_M)$,

and

$\text{Var}(\theta_M|\mathbf{y})$ is the variance–covariance matrix of the posterior distribution of θ_M .

The second term on the right hand side of (14) is measuring model complexity and is called the “effective number of parameters”.

Expression (12) is a result of combining the posterior expectation of the deviance, given by (13), with the effective number of parameters, which from (4), is given approximately by $\bar{D} - D(\bar{\theta}_M)$.

Thus, DIC has a term that reflects the fit of the model and a second term which introduces a penalty due to the complexity of the model.

The penalty inherent in DIC is more severe than other measures, such as the posterior Bayes Factor (apart from the second term in (11), \bar{D} already includes a penalty factor).

Models having a smaller DIC should be favoured as this indicates a better fit and a lower degree of model complexity.

It has been shown that DIC is related to other model comparison criteria and has an approximate decision-theoretic justification.

DIC is very easily calculated using the MCMC output.

The first term in the DIC is estimated using twice the average of the simulated values of $-\log p(\mathbf{y}|\theta_M)$,

and

the second term is the plug-in estimate of the deviance using the average of the MCMC simulated values of θ_M .

Example: *DIC in the mixed linear model*

Consider a hierarchical model with structure

$$\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \mathbf{e},$$

where

$$\mathbf{y}|\boldsymbol{\theta}, \mathbf{R} \sim N(\mathbf{W}\boldsymbol{\theta}, \mathbf{R})$$

$\boldsymbol{\theta} = [\boldsymbol{\beta}', \mathbf{u}']'$ is typically a vector of “fixed” and “random” effects, and the corresponding known incidence matrix is then $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$.

We can use the following:

$$\boldsymbol{\theta} | \boldsymbol{\mu}_\beta, \boldsymbol{\mu}_u, \mathbf{V}_\beta, \sigma_\beta^2, \mathbf{G}_u \sim N \left(\begin{bmatrix} \boldsymbol{\mu}_\beta \\ \boldsymbol{\mu}_u \end{bmatrix}, \begin{bmatrix} \mathbf{V}_\beta \sigma_\beta^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_u \end{bmatrix} \right).$$

The dispersion parameters $\mathbf{R}, \mathbf{V}_\beta, \sigma_\beta^2, \mathbf{G}_u$, and the location vectors $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\mu}_u$ are assumed known. Then, posterior distribution of $\boldsymbol{\theta}$ is normal, with mean vector

$$\bar{\boldsymbol{\theta}} = \begin{bmatrix} \bar{\boldsymbol{\beta}} \\ \bar{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \frac{\mathbf{V}_{\beta}^{-1}}{\sigma_{\beta}^2} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_u^{-1} \end{bmatrix}^{-1} \times \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} + \frac{\mathbf{V}_{\beta}^{-1}}{\sigma_{\beta}^2}\boldsymbol{\mu}_{\beta} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{G}_u^{-1}\boldsymbol{\mu}_u \end{bmatrix},$$

and variance–covariance matrix

$$\mathbf{C}^{-1} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \frac{1}{\sigma_{\beta}^2}\mathbf{V}_{\beta}^{-1} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}_u^{-1} \end{bmatrix}^{-1}.$$

The deviance is

$$\begin{aligned} D(\boldsymbol{\theta}) &= -2\log p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{R}) \\ &= N\log(2\pi) + \log|\mathbf{R}| + (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}). \end{aligned}$$

Then,

$$D(\bar{\boldsymbol{\theta}}) = N\log(2\pi) + \log|\mathbf{R}| + (\mathbf{y} - \mathbf{W}\bar{\boldsymbol{\theta}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{W}\bar{\boldsymbol{\theta}}),$$

and the expected deviance becomes

$$\begin{aligned} \bar{D} &= N\log(2\pi) + \log|\mathbf{R}| + (\mathbf{y} - \mathbf{W}\bar{\boldsymbol{\theta}})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{W}\bar{\boldsymbol{\theta}}) \\ &\quad + \text{tr}(\mathbf{R}^{-1}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}'). \end{aligned}$$

The effective number of parameters is

$$\begin{aligned} p_D &= \bar{D} - D(\bar{\boldsymbol{\theta}}) \\ &= \text{tr}(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}). \end{aligned}$$

For example, let $\mathbf{R} = \mathbf{I}\sigma_e^2$ and $\mathbf{G}_u = \mathbf{I}\sigma_u^2$, which results in a variance component model. Further, let $\sigma_{\beta}^2 \rightarrow \infty$, to make prior information about $\boldsymbol{\beta}$ vague. Here

$$\begin{aligned} \mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \sigma_e^2 \\ &= \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \sigma_e^2, \end{aligned}$$

and

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \\ &= \mathbf{I}_{p_\beta+p_u} - \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix}. \end{aligned}$$

Hence

$$\begin{aligned} p_D &= \text{tr}(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}) \\ &= \text{tr}(\mathbf{I}_{p_\beta+p_u}) - \text{tr} \left\{ \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix} \right\} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} \text{tr} \begin{bmatrix} \mathbf{0} & \mathbf{C}^{\beta u} \\ \mathbf{0} & \mathbf{C}^{uu} \end{bmatrix} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} \text{tr}[\mathbf{C}^{uu}]. \end{aligned}$$

Note that the prior information about the \mathbf{u} vector results in that the effective number of parameters is smaller than the dimension of θ . ■

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.*, **91**, 109-122.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773-795.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Statist. Soc. B*, **56**, 3-48.