

Computing the Marginal Likelihood

David Madigan

Bayesian Criterion

$$\begin{aligned} p(M_k | D) &\propto p(D | M_k) p(M_k) \\ &= p(M_k) \int p(D | \theta_k, M_k) p(\theta_k | M_k) d\theta_k \end{aligned}$$

- Typically impossible to compute analytically
- All sorts of Monte Carlo approximations

Laplace Method for $p(D | M)$

$$\text{let } l(\theta) = \frac{\log(L(\theta))}{n} + \frac{\log p(\theta)}{n}$$

(i.e., the log of the integrand divided by n)

$$\text{then } p(D) = \int e^{nl(\theta)} d\theta$$

Laplace's Method:

$$p(D) \approx \int \exp[nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2 / (2\sigma^2)] d\theta$$

where $\sigma^2 = -1/l''(\tilde{\theta})$ and

$\tilde{\theta}$ is the posterior mode

Laplace cont.

$$p(D) = \int \exp[nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2 / (2\sigma^2)] d\theta$$
$$\approx \sqrt{2\pi\sigma} n^{-1/2} \exp\{nl(\tilde{\theta})\}$$

- Tierney & Kadane (1986, JASA) show the approximation is $O(n^{-1})$
- Using the MLE instead of the posterior mode is also $O(n^{-1})$
- Using the expected information matrix in σ is $O(n^{-1/2})$ but convenient since often computed by standard software
- Raftery (1993) suggested approximating $\tilde{\theta}$ by a single Newton step starting at the MLE
- Note the prior is explicit in these approximations

Monte Carlo Estimates of $p(D | M)$

$$p(D) = \int p(D | \theta) p(\theta) d\theta$$

Draw iid $\theta_1, \dots, \theta_m$ from $p(\theta)$:

$$\hat{p}(D) = \frac{1}{m} \sum_{i=1}^m p(D | \theta^{(i)})$$

In practice has large variance

Monte Carlo Estimates of $p(D | M)$ (cont.)

Draw iid $\theta_1, \dots, \theta_m$ from $p(\theta | D)$:

$$\hat{p}(D) = \frac{\sum_{i=1}^m w_i p(D | \theta^{(i)})}{\sum_{i=1}^m w_i}$$

“Importance
Sampling”

$$w_i = \frac{p(\theta^{(i)})}{p(\theta^{(i)} | D)} = \frac{\cancel{p(\theta^{(i)})} p(D)}{p(D | \theta^{(i)}) \cancel{p(\theta^{(i)})}}$$

Monte Carlo Estimates of $p(D | M)$ (cont.)

$$\begin{aligned}\hat{p}(D) &= \frac{\sum_{i=1}^m \frac{p(D)}{p(D | \theta^{(i)})} p(D | \theta^{(i)})}{\sum_{i=1}^m \frac{p(D)}{p(D | \theta^{(i)})}} \\ &= \left\{ \frac{1}{m} \sum_{i=1}^m p(D | \theta^{(i)})^{-1} \right\}^{-1}\end{aligned}$$

Newton and Raftery's "Harmonic Mean Estimator"

Unstable in practice and needs modification

$p(D | M)$ from Gibbs Sampler Output

First note the following identity (for any θ^*):

$$p(D) = \frac{p(D | \theta^*) p(\theta^*)}{p(\theta^* | D)}$$

$p(D | \theta^*)$ and $p(\theta^*)$ are usually easy to evaluate.

What about $p(\theta^* | D)$?

Suppose we decompose θ into (θ_1, θ_2) such that $p(\theta_1 | D, \theta_2)$ and $p(\theta_2 | D, \theta_1)$ are available in closed-form...

Chib (1995)

$p(D | M)$ from Gibbs Sampler Output

$$p(\theta_1^*, \theta_2^* | D) = p(\theta_2^* | D, \theta_1^*) p(\theta_1^* | D)$$

The Gibbs sampler gives (dependent) draws from $p(\theta_1, \theta_2 | D)$ and hence marginally from $p(\theta_2 | D)$...

$$\begin{aligned} p(\theta_1^* | D) &= \int p(\theta_1^* | D, \theta_2) p(\theta_2 | D) d\theta_2 \\ &\approx \frac{1}{G} \sum_{g=1}^G p(\theta_1^* | D, \theta_2^{(g)}) \end{aligned}$$

“Rao-Blackwellization”

$p(D | M)$ from Metropolis Output

$\pi(\theta | y) \propto \pi(\theta) f(y | \theta) \dots$ posterior density

$\{\theta^{(1)}, \dots, \theta^{(M)}\}$ generated by Metropolis - Hastings

with proposal density $q(\theta, \theta')$ and acceptance prob :

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{\pi(\theta') f(y | \theta') q(\theta', \theta)}{\pi(\theta) f(y | \theta) q(\theta, \theta')} \right\}$$

Chib and Jeliazkov, JASA, 2001

$p(D | M)$ from Metropolis Output

$$\alpha(\theta, \theta')q(\theta, \theta')\pi(\theta | y) = \alpha(\theta', \theta)q(\theta', \theta)\pi(\theta' | y)$$

$$\int \alpha(\theta, \theta')q(\theta, \theta')\pi(\theta | y)d\theta = \int \alpha(\theta', \theta)q(\theta', \theta)\pi(\theta' | y)d\theta$$

$$\begin{aligned} \pi(\theta' | y) &= \frac{\int \alpha(\theta, \theta')q(\theta, \theta')\pi(\theta | y)d\theta}{\int \alpha(\theta', \theta)q(\theta', \theta)d\theta} \\ &= \frac{E_1\{\alpha(\theta, \theta')q(\theta, \theta')\}}{E_2\{\alpha(\theta', \theta)\}} \end{aligned}$$

E_1 with respect to $[\theta|y]$

E_2 with respect to $q(\theta', \theta)$

$$\approx \frac{M^{-1} \sum_{g=1}^M \alpha(\theta^{(g)}, \theta')q(\theta^{(g)}, \theta')}{J^{-1} \sum_{j=1}^J \alpha(\theta', \theta^{(j)})}$$

Bayesian Information Criterion

$$S_{BIC}(M_k) = 2S_L(\hat{\theta}_k; M_k) + d_k \log n, \quad k = 1, \dots, K$$

(S_L is the negative log-likelihood)

- BIC is an $O(1)$ approximation to $p(D | M)$
- Circumvents explicit prior
- Approximation is $O(n^{-1/2})$ for a class of priors called “unit information priors.”
- No free lunch (Weakliem (1998) example)

Srole (1956): "It's hardly fair to bring a child into the world now the way things look for the future." The data are from the 1993-94 General Social Survey; respondents were given the options of agreeing or disagreeing, and the few who could not choose are excluded from the analysis. The sample of 2,266 valid responses is composed of 44.0

$$\log(n_{ij}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \Theta x_3. \quad (4)$$

In this parameterization, x_1 is a dummy variable that is zero in row 1 and one in row 2, x_2 is a dummy variable that is zero in column 1 and one in column 2, and x_3 is a dummy variable that is one in column 2 of row 2 and zero otherwise. The maximum likelihood estimate of Θ is the logarithm of the observed odds ratio $(n_{11}n_{22})/(n_{12}n_{21})$, so another way to put the question is to ask if Θ is equal to zero.

The sample contains 412 men who agree with the statement, 583 men who disagree, 584 women who agree, and 687 women who disagree. The L^2 for the model of independence is 4.68 with one degree of

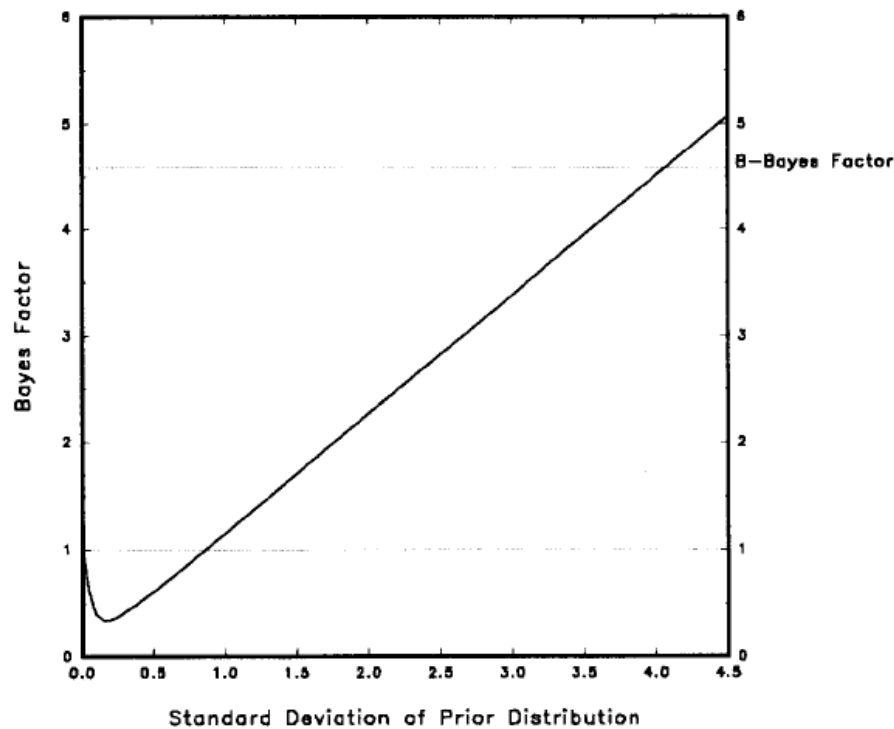


Figure 1: Bayes Factors for Model of No Association in Anomia by Gender Table:
Normal Prior Distribution With Mean Zero

about 4.0 (the exact figure is 4.07). With this prior distribution, the 95 percent range for possible values of the odds ratio would extend from $1/2,914$ to $2,914$, whereas the 50 percent range would extend from $1/14.7$ to 14.7 . In other words, adopting this prior distribution is equivalent to saying that if there is *any* association between the variables, there is a 50 percent chance that the absolute value of the odds ratio will be more than 14.7 or less than $1/14.7$. As discussed above,

Deviance Information Criterion

- Deviance is a standard measure of model fit:

$$D(y, \theta) = -2 \log p(y | \theta)$$

- Can summarize in two ways...at posterior mean or mode:

$$(1) \quad D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

or by averaging over the posterior:

$$(2) \quad D_{avg}(y) = E(D(y, \theta) | y)$$

(2) will be bigger (i.e., worse) than (1)

Deviance Information Criterion

$$p_D^{(1)} = D_{avg}(y) - D_{\hat{\theta}}(y)$$

is a measure of model complexity.

- In the normal linear model $p_D^{(1)}$ equals the number of parameters
- More generally $p_D^{(1)}$ equals the number of unconstrained parameters
- $DIC = D_{avg}(y) + p_D^{(1)}$
- Approximately equal to $E[D(y^{rep}, \hat{\theta}(y))]$

Score Functions on Hold-Out Data

- Instead of penalizing complexity, look at performance on hold-out data
- Note: even using hold-out data, performance results can be optimistically biased

- Pseudo-hold-out Score: $\prod_{i=1}^n [y_i | y_{-i}]$

$$\text{Recall: } \frac{1}{[y_i | y_{-i}]} = \int \frac{1}{[y_i | y_{-i}, \theta]} [\theta | y] d\theta$$

Checks Based on Individual Observations

Consider data y_1, \dots, y_I and parameters θ under the assumed model

Consider these ‘checking functions’

1. the residual: $y_i - E(Y_i)$
2. the standardised residual: $(y_i - E(Y_i)) / \sqrt{V(Y_i)}$
3. the chance of getting a more extreme observation: $\min(p(Y_i < y_i), p(Y_i \geq y_i))$
4. the chance of getting a more ‘surprising’ observation: $p(Y_i : p(Y_i) \leq p(y_i))$
5. the predictive ordinate of the observation: $p(y_i)$

Test Data?

1. *Separate evaluation data available.*

$$p(Y_i|\underline{x}) = \int p(Y_i|\theta)p(\theta|\underline{x})d\theta$$

\uparrow \leftarrow
ith test observation training data

2. *No separate evaluation data available.*

$$p(Y_i|y_{\setminus i}) \quad \text{“cross-validation”}$$

$$\frac{1}{p(y_i|y_{\setminus i})} = \int \frac{1}{p(y_i|y_{\setminus i}|\theta)}p(\theta|y) d\theta$$

See BUGS Manual