Bayesian Model Comparison

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 11, Hilary Term 2009

February 26, 2009

An integral of form

$$I = \int_a^b e^{-\lambda g(y)} h(y)\, dy$$

where $h(y)$ and $g(y)$ are smooth and g has local minimum at $y^* \in (a, b)$ can be approximated as

$$I = e^{-\lambda g(y*)} h(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{ 1 + O\left(\frac{1}{\lambda}\right) \right\}.$$

A more accurate approximation is

$$I = e^{-\lambda \tilde{g}_\lambda(\tilde{y}_\lambda)} \sqrt{\frac{2\pi}{\lambda \tilde{g}''_\lambda(\tilde{y}_\lambda)}} \left\{ 1 + \frac{5\tilde{\rho}_3 - 3\tilde{\rho}_4}{24\lambda} + O\left(\frac{1}{\lambda^2}\right) \right\},$$

where now $\tilde{y}_\lambda$ maximizes $\tilde{g}_\lambda(y) = g(y) - \lambda^{-1} \log h(y)$, and

$$\tilde{\rho}_3 = \frac{\tilde{g}^{(3)}_\lambda(\tilde{y}_\lambda)}{\{\tilde{g}''_\lambda(\tilde{y}_\lambda)\}^{3/2}}, \quad \tilde{\rho}_4 = \frac{\tilde{g}^{(4)}_\lambda(\tilde{y}_\lambda)}{\{\tilde{g}''_\lambda(\tilde{y}_\lambda)\}^{2}}.$$

It holds approximately for large n, that the posterior distribution of θ is

$$\theta \sim \mathcal{N}_d\{\hat{\theta}, j_n(\hat{\theta})^{-1}\} = \mathcal{N}_d(\hat{\theta}, j(\hat{\theta})^{-1}/n\}.$$

A more accurate approximation is obtained from the Laplace approximation to be

$$\begin{aligned}\pi^*(\theta) &= \frac{\exp\{l(\theta)\}\pi(\theta)}{\int_\Theta \exp\{l(\theta)\}\pi(\theta)\,d\theta} \\ &= (2\pi/n)^{-d/2}\exp\{l(\theta) - l(\hat{\theta})\}\frac{\pi(\theta)}{\pi(\hat{\theta})}\left|j(\hat{\theta})\right|^{1/2}\{1 + O(n^{-1})\}.\end{aligned}$$

Note in particular the expression for the normalization constant

$$\int_\Theta f(x\,|\,\theta)\pi(\theta)\,d\theta = (2\pi/n)^{d/2}L(\hat{\theta})\pi(\hat{\theta})\left|j(\hat{\theta})\right|^{-1/2}\{1 + O(n^{-1})\}.$$

We consider a number of competing models $M_j, j = 1, \ldots, m$ for data X; for example M_1 might specify that the expectation of a component X_i of X depends linearly on covariates Y_i, an alternative M_2 may specify that it has a quadratic dependence, whereas a third model M_3 might specify that the expectation does not depend on Y_i at all.

Associated with each of these models are parameter spaces Θ_j and prior distributions $\pi_j(\theta_j)$ as well as prior model probabilities π_j for model M_j being the 'correct' description of affairs.

The posterior probability for model M_j would then satisfy

$$\pi_j^* \propto \int_{\Theta_j} f(x \mid \theta_j, M_j)\pi_j(\theta_j)\, d\theta_j \times \pi_j$$

i.e. it will as usual be proportional to the product of the marginal or *integrated likelihood* $\bar{L}_j$ of model M_j with the *prior model probability,* π_j where

$$\bar{L}_j = f(x \mid M_j) = \int_{\Theta_j} f(x \mid \theta_j, M_j)\pi_j(\theta_j)\, d\theta_j.$$

Comparing two models yields

$$\frac{\pi_j^*}{\pi_k^*} = \frac{f(x \mid M_j)}{f(x \mid M_k)} = \frac{\int_{\Theta_j} f(x \mid \theta_j, M_j)\pi_j(\theta_j)\, d\theta_j}{\int_{\Theta_2} f(x \mid \theta_k, M_k)\pi_k(\theta_k)\, d\theta_k} \frac{\pi_j}{\pi_k}.$$

The factor

$$B_{jk} = \frac{f(x \mid M_j)}{f(x \mid M_k)} = \frac{\int_{\Theta_j} f(x \mid \theta_j, M_j)\pi_j(\theta_j)\, d\theta_j}{\int_{\Theta_2} f(x \mid \theta_k, M_k)\pi_k(\theta_k)\, d\theta_k} = \frac{\bar{L}_j}{\bar{L}_k}.$$

ia known as the *Bayes Factor* in favour of model j over model k. Note that if the Bayesian model is taken to its consequence, this is nothing but the usual likelihood ratio.

Recall that Σ *follows an inverse Wishart distribution* if $K = \Sigma^{-1}$ follows a Wishart distribution, formally expressed as

$$\Sigma \sim \mathcal{IW}_d(\delta, \Psi) \iff K = \Sigma^{-1} \sim \mathcal{W}_d(\delta + d - 1, \Psi^{-1}),$$

i.e. if the density of K has the form

$$f(K \mid \delta, \Psi) \propto (\det K)^{\delta/2-1} e^{-\operatorname{tr}(\Psi K)/2}.$$

The inverse Wishart distributions form a conjugate family for Σ. If the prior distribution of Σ is $\mathcal{IW}_d(\delta, \Psi)$ and $W \mid \Sigma \sim \mathcal{W}_d(n, \Sigma)$, the posterior density of K is

$$\begin{aligned} f(K \mid \delta, \Psi, W) &\propto (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2} \\ &\quad \times (\det K)^{\delta/2-1} e^{-\operatorname{tr}(\Psi K)/2} \\ &= (\det K)^{(n+\delta)/2-1} e^{-\operatorname{tr}\{(\Psi+W)K\}/2}, \end{aligned}$$

and hence the posterior distribution is simply $\mathcal{IW}_d(\delta + n, \Psi + W) = \mathcal{IW}_d(\delta^*, \Psi^*)$.

To calculate the Bayes factor for independence we need the full form of the Wishart density for K:

$$\begin{aligned} & f_d(K \mid \delta, \Psi) \\ &= \quad c(d,\delta)^{-1}(\det\Psi)^{(\delta+d-1)/2}(\det K)^{\delta/2-1}e^{-\operatorname{tr}(\Psi K)/2} \end{aligned}$$

The constant $c(d,\delta)$ is

$$c(d,\delta) = 2^{(\delta+d-1)d/2}(2\pi)^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\{(\delta+d-i)/2\}.$$

The marginal density of W becomes

$$\begin{aligned} f(W \mid \delta, \Psi) &= \int f(W \mid n, K) f(K \mid \delta, \Psi)\, dK \\ &= (\det W)^{(n-d-1)/2} c(d,n)^{-1} c(d,\delta)^{-1} (\det \Psi)^{(\delta+d-1)/2} \\ &\quad \int (\det K)^{(n+\delta)/2-1} e^{-\operatorname{tr}\{K(W+\Psi)\}/2}\, dK \\ &= (\det W)^{(n-d-1)/2} c(d,n)^{-1} c(d,\delta)^{-1} (\det \Psi)^{(\delta+d-1)/2} \\ &\quad \{\det(\Psi + W)\}^{-(\delta+n-1)/2} c(d, n+\delta) \\ &= \frac{(\det W)^{(n-d-1)/2} (\det \Psi)^{(\delta+d-1)/2}}{\{\det(\Psi + W)\}^{(\delta+n-1)/2}} \frac{c(d, n+\delta)}{c(d,n) c(d,\delta)}. \end{aligned}$$

Consider now alternative models M_2 with Σ arbitrary and M_1 with Σ of block diagonal form, i.e. with

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

If the associated prior distributions are for M_2 that $\Sigma \sim \mathcal{IW}_d(\delta, I_d)$ and for M_1 that $\Sigma_{11} \sim \mathcal{IW}_r(\delta, I_r)$, and $\Sigma_{22} \sim \mathcal{IW}_s(\delta, I_s)$, we can now calculate the Bayes factor.

We get

$$
\begin{aligned}
B_{12} &= \frac{f(W_{11} \mid \delta, I_r) f(W_{22} \mid \delta, I_s)}{f(W \mid \delta, I_d)} \\
&= \frac{(\det W_{11})^{(n-r-1)/2} (\det W_{22})^{(n-s-1)/2}}{(\det W)^{(n-d-1)/2}} \\
&\quad \times \left\{ \frac{\det(I_d + W)}{\det(I_r + W_{11}) \det(I_s + W_{22})} \right\}^{(\delta+n-1)/2)} \\
&\quad \times \frac{c(d,n) c(d,\delta) c(r, n+\delta) c(s, n+\delta)}{c(d, n+\delta) c(r,n) c(r,\delta) c(s,n) c(s,\delta)}
\end{aligned}
$$

Note the similarity between the first fraction and Wilks' Λ for independence.

In general the Bayes factor is difficult or impossible to calculate explicitly.
Recall that for competing models M_1 and M_2 with parameters $\theta_1 \in \Theta_1 \in \mathcal{R}^{d_1}$ and $\theta_2 \in \Theta_2 \in \mathcal{R}^{d_2}$ and prior distributions π_1, π_2, the *Bayes factor* B in favour of M_1 over M_2 is

$$B = \frac{f(x_1, \ldots, x_n \mid M_1)}{f(x_1, \ldots, x_n \mid M_2)} = \frac{\int_{\Theta_1} f(x \mid \theta_1, M_1)\pi_1(\theta_1)\, d\theta_1}{\int_{\Theta_2} f(x \mid \theta_2, M_2)\pi_2(\theta_2)\, d\theta_2}.$$

Recall the approximate expression obtained for the Bayesian marginal likelihood using Laplace's method

$$\int_{\Theta} f(x \mid \theta)\pi(\theta)\, d\theta = (2\pi/n)^{d/2} L(\hat{\theta})\pi(\hat{\theta}) \left| j(\hat{\theta}) \right|^{-1/2} \{1 + O(n^{-1})\}.$$

We then get

$$B = (2\pi)^{(d_1-d_2)/2} n^{(d_2-d_1)/2} \frac{L(\hat\theta_1)\pi(\hat\theta_1)}{L(\hat\theta_2)\pi(\hat\theta_2)} \frac{\left|j_2(\hat\theta_2)\right|^{1/2}}{\left|j_1(\hat\theta_1)\right|^{1/2}} \{1 + O(n^{-1})\}.$$

To study the asymptotic behaviour of the Bayes factor we take logarithms and collect terms of similar order to get

$$\begin{aligned} \log B &= n\{\bar l_n(\hat\theta_1) - \bar l_n(\hat\theta_2)\} + \frac{d_2 - d_1}{2}\log n + \log\{\pi(\hat\theta_1)/\pi(\hat\theta_2)\} \\ &\quad -\frac{1}{2}\log\left\{\left|j_1(\hat\theta_2)\right| / \left|j_1(\hat\theta_1)\right|\right\} - \frac{d_2 - d_1}{2}\log(2\pi) + O(n^{-1}). \end{aligned}$$

The dominating terms are those on the first line, as all other terms are of smaller order for $n \to \infty$. Ignoring the latter we get

$$\log B \approx \{l(\hat{\theta}_1) - l(\hat{\theta}_2)\} - \frac{d_1 - d_2}{2} \log n.$$

The right-hand side is the *Bayesian Information Criterion* (BIC). It reflects that, for large n, the Bayes factor will favour the model with highest maximized likelihood (the first term), but will also penalize the model having the largest number of parameters.

The prior distributions π_i do not enter in the expression for BIC which may or may not be seen as an advantage.

Models with a *high* value of BIC would be preferred over models with a low value of BIC.

One can get a more accurate approximation of the Bayes factor by adding terms

$$-\frac{1}{2}\log\left\{\left|j_i(\hat{\theta}_2)\right|\right\}+\frac{d_i}{2}\log(2\pi)$$

but this correction is not increasing with n, so it is most commonly ignored.

For the comparison of two models we get

$$\begin{aligned}\Delta\mathrm{BIC} &= l(\hat{\theta}_1)-l(\hat{\theta}_2)+\frac{d_1-d_2}{2}\log n\\ &= -\log\mathrm{LR}+\frac{d_1-d_2}{2}\log n.\end{aligned}$$

Thus, in comparison with straight maximized likelihood, the simpler model gets preference by entertaining a lower penalty.

In the nested case, if $d_1 < d_2$ the *deviance difference* between the models is $D = -2 \log \mathrm{LR}$ so

$$2\Delta\mathrm{BIC} = D + (d_1 - d_2) \log n.$$

If the true value of the parameter $\theta_0 \in M_1 \subseteq M_2$, the deviance D would under suitable regularity conditions be approximately $\chi^2(d_2 - d_1)$. The penalty term will thus dominate for large values of n, so the simpler model will eventually be chosen.

In this sense, *BIC will asymptotically choose the simplest model which is correct,* often referred to as *consistency* of the BIC.