Bayesian Asymptotics

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 8, Hilary Term 2008

May 7, 2008

For large λ we have the approximation

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

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 $\tilde{y}_\lambda$ maximizes $\tilde{g}_\lambda(y)$ and

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

In the multivariate case we have

$$\begin{aligned} I &= \int_B e^{-\lambda g(y)} h(y)\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) \int_{\mathcal{R}^d} e^{-\lambda (y-y^*)^\top \frac{\partial^2 g(y^*)}{\partial y \partial y^\top}(y-y^*)/2+\dots}\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) (2\pi/\lambda)^{d/2} \left| \frac{\partial^2 g(y^*)}{\partial y \partial y^\top} \right|^{-1/2} \left\{ 1 + O\left(\frac{1}{\lambda}\right)\right\} \end{aligned}$$

and additional accuracy up to $O(\lambda^{-2})$ can be obtained using derivatives of third and fourth order as in the univariate case.

We consider a standard asymptotic setup, involving $X_1, \ldots, X_n, \ldots$ random variables which, conditional on a d-dimensional parameter θ are independent and identically distributed with density $f(x \mid \theta)$, and $\pi(\theta)$ is the prior distribution of the parameter θ.

The posterior density is determined as

$$\pi^*(\theta) = f(\theta \mid x) \propto e^{l(\theta)}\pi(\theta),$$

where $l(\theta) = \log L(\theta)$ is the log-likelihood function. Letting

$$\bar{l}_n(\theta) = l(\theta)/n = \frac{1}{n}\sum_1^n \log f(X_i \mid \theta),$$

the law of large numbers yields that for $n \to \infty$,

$$\bar{l}_n(\theta) \to \mathbf{E}_\theta\{\log f(X \mid \theta)\} = -H(\theta),$$

where $H(\theta)$ is the *entropy* of the density $f(\cdot \mid \theta)$.

Thus the variation in the posterior density

$$\pi^*(\theta) \propto e^{n\bar{l}_n(\theta)}\pi(\theta)$$

will for sufficiently large n be dominated by the contribution from the likelihoood funtion. Expanding $l(\theta)$ around the maximum likelihood estimate $\hat{\theta}$ yields

$$\pi^*(\theta) \propto e^{n\bar{l}_n(\hat{\theta})}\pi(\hat{\theta})e^{-(\theta-\hat{\theta})^\top j_n(\hat{\theta})(\theta-\hat{\theta})/2} \propto e^{-(\theta-\hat{\theta})^\top j_n(\hat{\theta})(\theta-\hat{\theta})/2}$$

where $j_n(\hat{\theta}) = nj(\hat{\theta})$ is the observed information matrix, so, approximately for large n, the posterior distribution of θ is

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

Note this expression makes perfect sense, as $\hat{\theta}$ is not random in the posterior distribution.

A more accurate approximation is obtained by expanding around the posterior mode θ^*_π to get

$$\pi^*(\theta) \propto e^{-(\theta-\theta^*_\pi)^\top j_n(\theta^*_\pi)(\theta-\theta^*_\pi)/2}$$

yielding, approximately for large n, the posterior distribution of θ as

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

Note both differences and similarities to the analogous frequentist results

$$\hat{\theta} \sim \mathcal{N}_d\{\theta, i_n(\theta)^{-1}\} \quad \hat{\theta} \sim \mathcal{N}_d\{\theta, i_n(\hat{\theta})^{-1}\}, \quad \hat{\theta} \sim \mathcal{N}_d\{\theta, j_n(\hat{\theta})^{-1}\},$$

where the two latter needs appropriate interpretation to make perfect sense.

We can obtain an accurate approximation of the posterior distribution by applying Laplace's method to the normalization constant:

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

Note in particular the expression for the normalization constant

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

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

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

Using the approximate expression obtained for the normalization constants, we 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.