Laplace's Method of Integration

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 10, Hilary Term 2009

February 23, 2009

Consider an integral of form

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

where

1. λ is large;
2. $g(y)$ is a smooth function which has a local minimum at y^* in the interior of the interval (a, b);
3. $h(y)$ is smooth.

The integral can be the moment generating function of the distribution of $g(Y)$ when Y has density h, it could be a posterior expectation of $h(Y)$, or just an integral.

When λ is large, the contribution to this integral is essentially entirely originating from a neigbourhood around y^*.

We formalize this by Taylor expansion of the function g around y^*:

$$g(y) = g(y^*) + g'(y^*)(y - y^*) + g''(y^*)(y - y^*)^2/2 + \cdots$$

Since y^* is a local minimum, we have $g'(y^*) = 0$, $g''(y^*) > 0$, and thus

$$g(y) - g(y^*) = g''(y^*)(y - y^*)^2/2 + \cdots$$

If we further approximate $h(y)$ linearly around y^* we get

$$\begin{aligned} I &= \int_a^b e^{-\lambda g(y)} h(y)\, dy \\ &\approx e^{-\lambda g(y*)} h(y^*) \int_{-\infty}^{\infty} e^{-\lambda g''(y^*)(y-y^*)^2/2}\, dy \\ &\quad + e^{-\lambda g(y*)} h'(y^*) \int_{-\infty}^{\infty} (y - y^*) e^{-\lambda g''(y^*)(y-y^*)^2/2}\, dy \\ &= e^{-\lambda g(y*)} h(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}} + 0. \end{aligned}$$

We have exploited that we know the integral and expectation of a Gaussian density with concentration $g''(y^*)\lambda$. The approximation is typically very accurate and satisfies

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

meaning that the relative error

$$\frac{I - A}{A}$$

is $O(\lambda^{-1})$ and thus remains bounded for $\lambda \to \infty$, *even when multiplied with λ*.

Consider the Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

and recall that for integers λ we have

$$\Gamma(\lambda + 1) = \lambda!$$

We get

$$\Gamma(\lambda + 1) = \int_0^\infty t^\lambda e^{-t}\, dt.$$

Substituting $y = t/\lambda$ and letting $g(y) = y - \log y$ we get

$$\Gamma(\lambda + 1) = \lambda \int_0^\infty (\lambda y)^\lambda e^{-\lambda y}\, dy = \lambda^{\lambda+1} \int_0^\infty e^{-\lambda g(y)}\, dy.$$

To use Laplace's method we differentiate twice and get

$$g'(y) = 1 - 1/y, \quad g''(y) = 1/y^2$$

so that $y^* = 1$, $g(y^*) = 1$ and $g''(y^*) = 1$. Laplace's method now yields

$$\begin{aligned} \Gamma(\lambda+1) &= \lambda^{\lambda+1} e^{-\lambda g(y*)} \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} \\ &= \lambda^{\lambda+1/2} e^{-\lambda} \sqrt{2\pi} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} \end{aligned}$$

which is known as *Stirling's formula*.

By expanding the function g further, the error of approximation can be improved for a constant function h so that

$$\begin{aligned}\tilde{I} &= \int_a^b e^{-\lambda g(y)}\, dy \\ &= e^{-\lambda g(y*)}\sqrt{\frac{2\pi}{\lambda g''(y^*)}}\left\{1+\frac{5\rho_3^* - 3\rho_4^*}{24\lambda}+O\left(\frac{1}{\lambda^2}\right)\right\},\end{aligned}$$

where

$$\rho_3^* = \frac{g^{(3)}(y^*)}{\{g''(y^*)\}^{3/2}}, \quad \rho_4^* = \frac{g^{(4)}(y^*)}{\{g''(y^*)\}^2}.$$

In this fashion we can also get *Stirling's improved formula* as

$$\Gamma(\lambda+1) = \lambda^{\lambda+1/2} e^{-\lambda} \sqrt{2\pi} \left\{ 1 + \frac{1}{12\lambda} + O\left(\frac{1}{\lambda^2}\right) \right\}$$

which is remarkably accurate, even for rather small values of λ, as this table of $\log \Gamma(\lambda+1)$ shows:

λ	Exact	Stirling	Improved
2	0.6931472	0.6518048	0.6926268
4	3.1780538	3.1572615	3.1778807
8	10.6046029	10.5941899	10.6045527
16	30.6718601	30.6666508	30.6718456
32	205.1681995	205.1668957	205.1681970

Alternatively, if the variation of h around y^* is not negligible, or a more accurate approximation is desired, one can incorporate h in g as

$$\tilde{g}_\lambda(y) = g(y) - \frac{1}{\lambda}\log h(y)$$

and get the approximation

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

where now $\tilde{y}_\lambda$ maximizes $\tilde{g}_\lambda(y)$, and other quantities are similarly defined.

The multivariate case is completely analogous. Here we again write

$$g(y) = g(y^*) + \frac{\partial g(y^*)}{\partial y}(y - y^*) + (y - y^*)^\top \frac{\partial^2 g(y^*)}{\partial y \partial y^\top}(y - y^*)/2 + \cdots$$

and exploit that the vector of partial derivatives $\frac{\partial g(y^*)}{\partial y}$ must vanish, whereby

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

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

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

The expression for the asymptotic posterior

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

makes perfect sense, as $\hat{\theta}$ is not random in the posterior distribution, whereas θ is.

Contrast this with the standard frequentist result which says that, approximately,

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

This expression does not make sense as written, but is a proxy for the result that

$$nj(\hat{\theta})^{1/2}(\hat{\theta} - \theta) \sim \mathcal{N}_d(0, I),$$

which is identical to the similar Bayesian formulation, just that in the latter θ is random rather than $\hat{\theta}$!

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

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