# Notes on the Laplace approximation

2022-03-16

## 1 Notation

These notes adopt the notation for Bayesian computation introduced in STSCI 4780/5780 Lec11:

$$\begin{aligned} p(\theta|D, M) &= \frac{p(\theta|M)p(D|\theta, M)}{p(D|M)} & (1) \\ &= \frac{\pi(\theta)\mathcal{L}(\theta)}{Z} = \frac{q(\theta)}{Z} = p(\theta), & (2) \end{aligned}$$

where

- $M$ = model specification (context)
- $D$ specifies observed data
- $\theta$ = model parameters, of dimension $m$
- $\pi(\theta)$ = prior pdf for $\theta$
- $\mathcal{L}(\theta)$ = likelihood for $\theta$ (likelihood function)
- $q(\theta) \equiv \pi(\theta)\mathcal{L}(\theta)$ = "quasiposterior"
- $Z = p(D|M) = \int d\theta\, \pi(\theta)\mathcal{L}(\theta)$, the (marginal) likelihood for the model

The main tasks we need to do in Bayesian computation are (1) optimization (e.g., to find the posterior mode), and (2) computing linear functionals equal or proportional to the expectation value of some function, $g(\theta)$, with respect to the posterior PDF for $\theta$. We can write such functionals as follows:

$$I_q[g] \equiv \int d^m\theta\, g(\theta)\, \pi(\theta)\, \mathcal{L}(\theta) = \int d^m\theta\, g(\theta)\, q(\theta). \quad (3)$$

Here are some examples of quantities of interest, and the corresponding choice of $g(\theta)$:

- $g(\theta) = 1 \rightarrow Z = p(D|M)$ (norm. const., model likelihood)
- $g(\theta) = \theta/Z \rightarrow$ posterior mean for $\theta$
- $g(\theta) =$ 'box' $\rightarrow$ probability $\theta \in$ credible region
- $g(\theta) = 1/Z$, integrate over $< m$ params $\rightarrow$ marginal posterior
- $g(\theta) = \delta[\psi - \psi(\theta)]/Z \rightarrow$ propagate uncertainty to $\psi(\theta)$

## 2 Univariate Laplace approximation

We'll focus on the 1-D parameter ($m = 1$) case here. Denote the integrand in $I_q$ as $f(\theta) = g(\theta)\, q(theta)$. The Laplace approximation (LA) approximates the integrand, $f(\theta)$, with a Gaussian function,

$$\tilde{f}(\theta) = A \exp\left[-\frac{1}{2}\frac{(\theta - \hat{\theta})^2}{\sigma^2}\right], \tag{4}$$

where we compute the three parameters, $\hat{\theta}$, $A$, and $\sigma$, as follows:

- $\hat{\theta}$ is the location of the maximum of $f(\theta)$, $\hat{\theta} = \arg\max_\theta f(\theta)$. When implementing the LA analytically, we may find $\hat{\theta}$ by requiring the first derivative of $f$ to vanish: $f'(\hat{\theta}) = 0$ (and verifying that we have found a maximum, not a minimum). When implementing the LA numerically, we typically use an iterative optimizer, rather than trying to solve the vanishing derivative equation.

- $A$ is the value of $f$ at its peak, $A = f(\hat{\theta})$;

- $\sigma$ is found by matching the second derivative (curvature) of $\log \tilde{f}$ and $\log f$ at $\hat{\theta}$. Let $\Lambda(\theta) = \log f(\theta)$. At the peak (where the first derivative vanishes),

$$\Lambda''(\hat{\theta}) = \frac{1}{f(\hat{\theta})} \left.\frac{\mathrm{d}^2 f(\theta)}{\mathrm{d}\theta^2}\right|_{\hat{\theta}}. \tag{5}$$

  For numerical calculations, we can compute this derivative via a second-difference formula (applied either to $\Lambda$ or to $f$). For analytical calculations, it's often convenient to drop any $\theta$-independent factors from $f(\theta)$, writing $f(\theta) = Ck(\theta)$ with the "kernel" function, $k$, capturing the $\theta$ dependence. Then equation (6) takes the form

$$\Lambda''(\hat{\theta}) = \frac{1}{k(\hat{\theta})} \left.\frac{\mathrm{d}^2 k(\theta)}{\mathrm{d}\theta^2}\right|_{\hat{\theta}}. \tag{6}$$

  Matching $\Lambda''(\hat{\theta})$ to the 2nd derivative of the log of the approximating Gaussian function, equation (4), gives

$$\frac{1}{\sigma^2} = -\Lambda''(\hat{\theta}). \tag{7}$$

Once we have the three ingredients— $\hat{\theta}$, $A$, and $\sigma$—we can compute the LA for $I_q[g]$ by integrating $\tilde{f}$ in place of $f$:

$$\begin{aligned} I_q[g] &= \int \mathrm{d}\theta\ f(\theta) \\ &\approx \int \mathrm{d}\theta\ \tilde{f}(\theta) \\ &= A\sigma\sqrt{2\pi}. \end{aligned} \tag{8}$$

For many quantities of interest, $g(\theta)$ will include a factor of $1/Z$, because we need an integral over the full posterior, $p(\theta)$, not just over the unnormalized quasiposterior, $q(\theta)$.

E.g., for computing a posterior mean, $g(\theta) = \theta/Z$. In such cases, we need to first compute an estimate of $Z$, using the LA with $g(\theta) = 1$. The resulting estimate, $\tilde{Z}$, could then be used to define $\tilde{g}(\theta) = \theta/\tilde{Z}$, for subsquent application of the LA. Equivalently, one could take $g(\theta) = \theta$, compute the LA for $I_q[g]$, and divide the resulting estimated integral by $\tilde{Z}$. By a convenient coincidence, it turns out that the leading order error terms in the LA for $g(\theta) = \theta$ and for $Z$ (i.e., $g(\theta) = 1$) *cancel* when computing the ratio of estimates, so the LA for posterior moments is more accurate than one might expect, in terms of its asymptotic behavior (how the overall error falls with sample size).

— *Tom Loredo*