STA 250: Statistics

Notes 11. Laplace Approximation to the Posterior

Book chapters: –

1 Non-conjugate prior and difficulty with posterior computation

While conjugate priors make computation easy, they may not be always appropriate and sometimes they simply do not exist (in a useful way) for the statistical model we want to analyze.

Example (Tennis serves)**.** Consider data $X = (X_1, \cdots, X_n)$ on the first serve success rates of a tennis player from n tournament matches. Consider the model $X_i \overset{\text{IID}}{\sim} g(x_i|\theta) = \theta(\theta+1)x_i^{\theta-1}(1-x_i)$, $x_i \in (0,1)$ with parameter $\theta > 0$. No useful conjugate prior family exists for this model. But it is reasonable to assign θ a $\mathsf{Gamma}(a,b)$ prior. The posterior pdf in θ looks like:

$$\begin{aligned}
\xi(\theta|x) &= \text{const} \times L_x(\theta) \times \xi(\theta) \\
&= \text{const} \times \left\{ \theta^n (\theta+1)^n \prod_{i=1}^n x_i^\theta \right\} \times \theta^{a-1} e^{-b\theta}, \qquad \theta > 0 \\
&= \text{const} \times \theta^{n+a-1} (\theta+1)^n e^{-\{b + \sum_{i=1}^n \log(1/x_i)\}\theta}, \qquad \theta > 0
\end{aligned}$$

which is not a standard pdf.

When $\xi(\theta|x)$ is not of a standard form, it is difficult to summarize its quantiles, or to make a plot (evaluating the normalizing constant is difficult), or even to sample draws from it in order to make prediction.

2 Laplace's technique: normal approximation to posterior for regular models

For any pdf that is smooth and well peaked around its point of maxima, Laplace proposed to approximate it by a normal pdf. It's a simple 2-term Taylor expansion trick on the log pdf. If $\hat{\theta}$ denotes the point of maxima of a pdf $h(\theta)$, then it is also the point of maxima of the log-pdf $q(\theta) = \log h(\theta)$ and we can write:

$$\begin{aligned}
q(\theta) &\approx q(\hat{\theta}) + (\theta - \hat{\theta})\dot{q}(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 \ddot{q}(\hat{\theta}) \\
&= q(\hat{\theta}) + 0 + \frac{1}{2}(\theta-\hat{\theta})^2 \ddot{q}(\hat{\theta}) \qquad [\text{because } \dot{q}(\hat{\theta}) = 0] \\
&= \text{const} - \frac{1}{2}(\theta-\hat{\theta})^2 \ddot{q}(\hat{\theta}) \\
&= \text{const} - \frac{(\theta-\tilde{a})^2}{2\tilde{b}^2}
\end{aligned}$$

with $\tilde{a} = \hat{\theta}$ and $\tilde{b}^2 = \{-\ddot{q}(\hat{\theta})\}^{-1}$ (notice $\ddot{q}(\hat{\theta}) < 0$ because $\hat{\theta}$ is a maxima). But the right hand side of the last display matches the log-pdf of $\mathsf{Normal}(\tilde{a}, \tilde{b}^2)$. Hence the pdf $h(\theta)$ is approximately the $\mathsf{Normal}(\tilde{a}, \tilde{b}^2)$ pdf with $\tilde{a} = \hat{\theta}$ and $\tilde{b}^2 = \{-\ddot{q}(\hat{\theta})\}^{-1}$.

Laplace's approximation is simple and elegant, all one needs is that the log-pdf is smooth at the maximum and peaks well at it so that the quadratic approximation is good. Also, to make it operational, we only need to know the point of maximum $\hat{\theta}$ and the curvature $-\ddot{q}(\theta)$ at this point.

The same technique could be applied to a posterior pdf $\xi(\theta|x) = \text{const} \times L_x(\theta)\xi(\theta)$. The log-pdf in this case is $q(\theta) = \text{const} + \ell_x(\theta) + \log \xi(\theta)$. Typically we will not know the value of the constant term in the front. But it does not affect computing the point of maximum $\hat{\theta}$ and the curvature $-\ddot{q}(\theta)$ at $\theta = \hat{\theta}$. See the example below.

Example (Tennis serves (contd.))**.** For our model,

$$\ell_x(\theta) = \text{const} + n \log \theta + n \log(\theta + 1) + \theta \sum_{i=1}^{n} \log x_i,$$

and

$$\log \xi(\theta) = \text{const} + (a - 1) \log \theta - b\theta,$$

and so

$$q(\theta) = \log \xi(\theta|x) = \text{const} + (n + a - 1) \log \theta + n \log(\theta + 1) - \theta \left\{ b - \sum_{i=1}^{n} \log x_i \right\}$$

with

$$\dot{q}(\theta) = \frac{n + a - 1}{\theta} + \frac{n}{\theta + 1} - \left\{ b - \sum_{i=1}^{n} \log x_i \right\}$$

and

$$\ddot{q}(\theta) = -\frac{n + a - 1}{\theta^2} - \frac{n}{(\theta + 1)^2}.$$

Suppose recorded data shows $n = 20$, $\sum_{i=1} \log X_i = -4.59$. Also suppose we work with $a = 1, b = 1$. So we can find the maxima $\hat{\theta}$ by solving $\dot{q}(\theta) = 0$, i.e, $20/\theta + 20/(\theta+1) - 5.59 = 0$, which is solved at $\hat{\theta} = 6.69$. The curvature at the maximum equals $-\ddot{q}(6.69) = 0.785$. Hence $\xi(\theta|x) \approx \mathsf{Normal}(6.69, 1/0.785) = \mathsf{Normal}(6.69, 1.129^2)$.

3 Finding $\hat{\theta}$

In the above example I solved $20/\theta + 20/(\theta + 1) - 5.59 = 0$ by first re-writing the equation as the quadratic: $20(\theta+1) + 20\theta - 5.59\theta(\theta+1) = 0$ (by multiplying each side with $\theta(\theta+1)$). Such simplifications may not be available in all cases. In general one can use "Newton's method" to find $\hat{\theta}$. In Newton's method, you start with an initial guess $\theta = \theta_0$ and keep iterating:

$$\theta_{t+1} = \theta_t - \frac{\dot{q}(\theta_t)}{\ddot{q}(\theta_t)}, \quad t = 1, 2, \cdots$$

The sequence $\theta_0, \theta_1, \theta_2, \cdots$ eventually converges to the solution $\hat{\theta}$.

4 Quality of the normal approximation to the posterior

Laplace's technique just gives a way to get a bell curve to approximate the posterior pdf. It does not say anything about the quality of the approximation. However, there are general guarantees that such an approximation is actually very good, when the model is a "regular" one, the prior pdf is smooth and the sample size n is large. We state the following result (without technical details) that is an analogue of the asymptotic normality result for MLE (Notes 6).

RESULT 1. (**Bernstein-von Mises Theorem**) Consider the model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$. Under some regularity conditions on the pdfs/pmfs $g(\cdot|\theta)$, including that all of them have the same support, and that for each x_i, $\theta \mapsto \log g(x_i|\theta)$ is twice continuously differentiable, we have that for any prior $\xi(\theta)$ which is positive, bounded and twice differentiable over Θ,

$$\sup_z \left| P(\theta \leq z | X = x) - \Phi\left(\{-\ddot{q}(\hat{\theta})\}^{1/2} (z - \hat{\theta}) \right) \right| \approx 0$$

for all large n.

Under the same regularity condition it turns out that $\hat{\theta} \approx \hat{\theta}_{\text{MLE}}(x)$ and that $-\ddot{q}(\hat{\theta}) \approx I_{\text{OBS}}(x)$.

5 Normal approximation to conjugate posterior

Bernstein-von Mises clearly applies to most of the standard models for which a conjugate prior family exists (among the ones we have seen, binomial, poisson, exponential are regular families, but uniform is not). Therefore for large n, the conjugate posterior too should look like a bell curve. This could also be verified case by case, and in fact one could get easier normal approximation without having to use the Laplace's technique. For example, for the binomial model $X \sim \mathsf{Binomial}(n, p)$ and a $\mathsf{Beta}(a, b)$ prior on $p \in (0, 1)$, the posterior is $\mathsf{Beta}(x + a, n - x + b)$ which should look like a bell curve when n is large. To identify the approximating bell curve $\mathsf{Normal}(m, s^2)$ we could simply match the mean and the variance: $m = (x + a)/(a + b + n)$ and $s^2 = m(1 - m)/(a + b + n + 1)$.