4: Asymptotic and connections to non-Bayesian approaches

09/16/19

Introduction

We examine what happens to posterior distributions when $n \to \infty$. These results help us understand our models better, and they can suggest useful approximations (when computation is too difficult).

Bayesian Consistency

A mathematical framework

1. likelihood we are using/assuming: $p(y \mid \theta)$
2. prior we are using $p(\theta)$
3. the true distribution $f(y) = \prod_{i=1}^{n} f(y_i)$
4. Kullback-Leibler divergence: $0 \leq KL(\theta) = E_f \left[\log \left(\frac{f(y_i)}{p(y_i|\theta)} \right) \right]$
5. θ_0 is the **unique** minimizer of $KL(\theta)$

Bayesian Consistency on finite parameter space

Theorem 1

Suppose there exists θ_0 such that $f(y_i) = p(y_i \mid \theta_0)$ and the parameter space is finite. If $p(\theta_0) > 0$ (prior puts mass on the true value), then

$$p(\theta_0 \mid y) \to 1$$

as $n \to \infty$.

Convergence is with respect to $f(y)$!

Bayesian Consistency

Recall that if $\bar{Y}_n \xrightarrow{p} \mu < 0$, then $\sum_i Y_i \xrightarrow{p} -\infty$.

The y_i are random here! We are keeping parameters fixed. Whenever $\theta \neq \theta_0$,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p(y_i \mid \theta)}{p(y_i \mid \theta_0)}\right) \xrightarrow{p} E_f\left[\log\left(\frac{p(y_i \mid \theta)f(y_i)}{p(y_i \mid \theta_0)f(y_i)}\right)\right]$$
$$= KL(\theta_0) - KL(\theta) < 0$$

1. so $\sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right) \xrightarrow{p} -\infty$
2. so $\log\left(\frac{p(\theta|y)}{p(\theta_0|y)}\right) = \log\frac{p(\theta)}{p(\theta_0)} + \sum_{i=1}^{n} \log\left(\frac{p(y_i|\theta)}{p(y_i|\theta_0)}\right) \xrightarrow{p} -\infty$ if $p(\theta_0) > 0$
3. so $\frac{p(\theta|y)}{p(\theta_0|y)} \xrightarrow{p} 0$ as long as $p(\theta_0) > 0$
4. so $p(\theta_0 \mid y) \xrightarrow{p} 1$ as long as $p(\theta_0) > 0$

Bayesian Consistency when the parameter space is compact

Theorem 2

Suppose there exists θ_0 such that $f(y_i) = p(y_i \mid \theta_0)$ and the parameter space is uncountable and compact. Let $A_\epsilon = \{\theta \in \Theta : \rho(\theta, \theta_0) < \epsilon\}$ be the ϵ-ball about θ_0. For any $\epsilon > 0$, if $p(\theta \in A_\epsilon) > 0$, then

$$p(\theta \in A_\epsilon \mid y) \to 1$$

as $n \to \infty$.

Convergence is with respect to $f(y)$!

Asymptotic Normality: Laplace's Method

These ideas are based on using a Taylor approximation for your posterior distribution.

1. approximations are second-order (quadratic)
2. centered about the **posterior mode** $\hat{\theta}$
3. a better fit when the posterior is unimodal and symmetric
4. assume the mode is in the interior of the parameter space

Asymptotic Normality: Laplace's Method

These ideas are based on using a Taylor approximation for your posterior distribution.

1. approximations are second-order (quadratic)
2. centered about the **posterior mode** $\hat{\theta}$
3. a better fit when the posterior is unimodal and symmetric
4. assume the mode is in the interior of the parameter space

$$
\begin{aligned}
\log p(\theta \mid y) \approx{} & \\
& \log p(\hat{\theta} \mid y) + \overbrace{(\theta - \hat{\theta})' \left[\frac{d}{d\theta} \log p(\theta \mid y) \right] \Big|_{\theta = \hat{\theta}}}^{0} \\
& + \frac{1}{2} (\theta - \hat{\theta})' \left[\frac{d^2}{d\theta^2} \log p(\theta \mid y) \right] \Big|_{\theta = \hat{\theta}} (\theta - \hat{\theta}) \\
& = c - \frac{1}{2} (\theta - \hat{\theta})' \left[-\frac{d^2}{d\theta^2} \log p(\theta \mid y) \right] \Big|_{\theta = \hat{\theta}} (\theta - \hat{\theta})
\end{aligned}
$$

Asymptotic Normality: Laplace's Method

$$\log p(\theta \mid y) \approx c - \frac{1}{2}(\theta - \underbrace{\hat{\theta}}_{\text{mean}})' \underbrace{\left[-\frac{d^2}{d\theta^2} \log p(\theta \mid y)\right]\bigg|_{\theta=\hat{\theta}}}_{\text{precision}} (\theta - \hat{\theta})$$

The **observed posterior information** is

$$\begin{aligned}
&\left[-\frac{d^2}{d\theta^2} \log p(\theta \mid y)\right]\bigg|_{\theta=\hat{\theta}} \\
&= \left[-\frac{d^2}{d\theta^2} \log p(\theta)\right]\bigg|_{\theta=\hat{\theta}} + \sum_{i=1}^{n} \left[-\frac{d^2}{d\theta^2} \log p(y_i \mid \theta)\right]\bigg|_{\theta=\hat{\theta}} \\
&= I(\hat{\theta})
\end{aligned}$$

$\hat{\theta}$ is interior point in the parameter space $\Rightarrow I(\hat{\theta})$ is positive definite.

Asymptotic Normality: Laplace's Method

It's also justified to use the **observed likelihood Fisher Information** $J(\theta) = -E\left(\frac{\mathrm{d}^2 \log p(y|\theta)}{\mathrm{d}\theta^2}\right)$

$$\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log p(\theta \mid y)\right]\bigg|_{\theta=\hat{\theta}}$$
$$= \left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log p(\theta)\right]\bigg|_{\theta=\hat{\theta}} + n\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log p(y_i \mid \theta)\right]\bigg|_{\theta=\hat{\theta}}}_{\text{approx. } J(\hat{\theta})}$$

Asymptotic Normality

So we have, approximately for large n,

$$\theta \mid y_1, \ldots, y_n \sim \text{Normal}\left(\hat{\theta}, I(\hat{\theta})^{-1})\right)$$

or

$$\theta \mid y_1, \ldots, y_n \sim \text{Normal}\left(\hat{\theta}, n^{-1} J(\hat{\theta})^{-1})\right)$$

1. $\hat{\theta}$ is the posterior mode. Using MLE (ignoring prior) can also be justified.
2. $J(\hat{\theta})$ is the observed Fisher Information (of an individual datum's likelihood) evaluated at the posterior mode.
3. This result is known as the Bernstein-von Mises theorem when $\hat{\theta}$ is MLE. (Likelihood dominates prior in large sample)

Asymptotic Normality — Frequentist

Under some regularity conditions (notably that θ_0 is not on the boundary of parameter space and posterior consistency), as $n \to \infty$, the posterior distribution of θ, $p(\theta \mid y)$, approaches Normality with mean θ_0 and variance $(nJ(\theta_0))^{-1}$.

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution
- Data reduction and summary statistics, $\hat{\theta}$ and $I(\hat{\theta})$; useful in hierarchical modeling

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution
- Data reduction and summary statistics, $\hat{\theta}$ and $I(\hat{\theta})$; useful in hierarchical modeling
- Large sample confidence interval

$$I(\hat{\theta})^{1/2}(\theta - \hat{\theta}) \mid y \sim N(0, I)$$

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution
- Data reduction and summary statistics, $\hat{\theta}$ and $I(\hat{\theta})$; useful in hierarchical modeling
- Large sample confidence interval

$$I(\hat{\theta})^{1/2}(\theta - \hat{\theta}) \mid y \sim N(0, I)$$

Issues

- Cautious to use Normal approximation when the sample size is small

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution
- Data reduction and summary statistics, $\hat{\theta}$ and $I(\hat{\theta})$; useful in hierarchical modeling
- Large sample confidence interval

$$I(\hat{\theta})^{1/2}(\theta - \hat{\theta}) \mid y \sim N(0, I)$$

Issues

- Cautious to use Normal approximation when the sample size is small
- Cautious to use Normal approximation when the dimension of θ is high; typically more accurate for conditional and marginal distributions of components of θ

Discussions about Normality approximation

- Estimation of posterior mass via quantiles of χ^2 distribution
- Data reduction and summary statistics, $\hat{\theta}$ and $I(\hat{\theta})$; useful in hierarchical modeling
- Large sample confidence interval

$$I(\hat{\theta})^{1/2}(\theta - \hat{\theta}) \mid y \sim N(0, I)$$

Issues

- Cautious to use Normal approximation when the sample size is small
- Cautious to use Normal approximation when the dimension of θ is high; typically more accurate for conditional and marginal distributions of components of θ
- Convergence to normality of the posterior distribution can be dramatically improved by transformation on θ (example below)

Asymptotic Normality: example

Let $y_i \mid \mu, \theta \sim \mathrm{N}(\mu, \exp(2\theta))$ and $p(\mu, \theta) \propto 1$ with $\theta = \log \sigma$. Then

$$
\begin{aligned}
p(\mu, \theta \mid y) &\propto (2\pi)^{-n/2} \exp(-n\theta) \exp\left[-\frac{1}{2\exp(2\theta)} \sum_i (y_i - \mu)^2\right] \\
&= (2\pi)^{-n/2} \exp(-n\theta) \exp\left[-\frac{1}{2\exp(2\theta)} \left\{n(\mu - \bar{y})^2 + (n-1)s^2\right\}\right]
\end{aligned}
$$

let's approximate this for some practice!

Asymptotic Normality: example

$$
\begin{aligned}
&\frac{d}{d\mu} \log p(\mu, \theta \mid y) \\
&= \frac{d}{d\mu} \left[-\frac{n}{2} \log(2\pi) - n\theta - \frac{1}{2\exp(2\theta)} \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \right] \\
&= -\frac{n(\mu - \bar{y})}{\exp(2\theta)} \overset{\text{set}}{=} 0
\end{aligned}
$$

which means $\hat{\mu} = \bar{y}$

Asymptotic Normality: example

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(\mu, \theta \mid y) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \left[-\frac{n}{2} \log(2\pi) - n\theta - \frac{1}{2\exp(2\theta)} \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \right] \\
&= -n + \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \exp(-2\theta) \overset{\text{set}}{=} 0
\end{aligned}
$$

which means $\hat{\theta} = \log \left\{ \sqrt{\frac{n-1}{n} s^2} \right\}$ after we plug in $\hat{\mu}$

Asymptotic Normality: example

The mean vector is

$$\begin{bmatrix} \hat{\mu} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \log\left\{\sqrt{\frac{n-1}{n}s^2}\right\} \end{bmatrix}$$

Now let's find the observed (posterior) information

Asymptotic Normality: example

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \log p(\mu, \theta \mid y) = -\frac{\mathrm{d}}{\mathrm{d}\mu} \frac{n(\mu - \bar{y})}{\exp(2\theta)}$$
$$= -n \exp(-2\theta)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log p(\mu, \theta \mid y) = \frac{\mathrm{d}}{\mathrm{d}\theta} \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \exp(-2\theta)$$
$$= -2 \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \exp(-2\theta)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu \mathrm{d}\theta} \log p(\mu, \theta \mid y)$$
$$= \frac{\mathrm{d}}{\mathrm{d}\mu} \left\{ n(\mu - \bar{y})^2 + (n-1)s^2 \right\} \exp(-2\theta)$$
$$= 2n(\mu - \bar{y}) \exp(-2\theta)$$

Asymptotic Normality: example

When we plug in the estimates, then the precision matrix is

$$I(\hat{\theta}) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log p(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = \begin{bmatrix} \frac{n^2}{(n-1)s^2} & 0 \\ 0 & 2n \end{bmatrix}$$

so

$$p(\mu, \theta \mid y) \approx \mathrm{N}\left(\begin{bmatrix} \bar{y} \\ \log\left\{\sqrt{\frac{n-1}{n}s^2}\right\} \end{bmatrix}, \begin{bmatrix} \frac{(n-1)s^2}{n^2} & 0 \\ 0 & \frac{1}{2n} \end{bmatrix} \right)$$

Asymptotic Normality: Bioassay experiment

```
w0 <- c(0,0)
optim_res <- optim(w0, bioassayfun, gr = NULL, df1,
                   hessian = T)
w <- optim_res$par
S <- solve(optim_res$hessian)
```

http:
//avehtari.github.io/BDA_R_demos/demos_ch4/demo4_1.html