4: Asymptotic and connections to non-Bayesian approaches

Taylor

University of Virginia

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: $0 \underset{\text{hw q}}{<} KL(\theta) = E_f \left[\log \left(\frac{f(y_i)}{p(y_i|\theta)} \right) \right]$
5. θ_0 is the minimizer of $KL(\theta)$

Bayesian Consistency on finite parameter space

Theorem 1

Suppose there exists θ such that $f(y_i) = p(y_i \mid \theta)$ 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. Also $\theta \neq \theta_0$:

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

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

Theorem 2

Suppose there exists θ such that $f(y_i) = p(y_i \mid \theta)$ 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. Assume 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. Assume the posterior is unimodal and symmetric
4. Assume the mode is in the interior of the parameter space

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

Asymptotic Normality: Laplace's Method

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

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

Asymptotic Normality

So we have, approximately for large n,

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

1. $\hat{\theta}$ is the posterior mode. Using MLE (ignoring prior) is also justified.
2. $J(\hat{\theta})$ is the Fisher Information (of an individual datum's likelihood) evaluated at the posterior mode.
3. This result is known as the Bernstein-von Mises theorem. Proof omitted.

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

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

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 precision matrix

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

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

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: cases of unmet assumptions

We go through some common examples where one of the above assumptions is not met. In these cases, using asymptotics is not allowed.

Asymptotic Normality: cases of unmet assumptions

A *model* is **underidentified** given data y if the likelihood, $p(y \mid \theta)$, is equal for a range of values θ.

A *model* is **weakly identified** given data y if the likelihood, $p(y \mid \theta)$, is close to being equal for a range of values θ.

These can be problematic because $\hat{\theta}$ will not have any specific number/vector θ to which it can converge. These are violations of assumption (3).

Asymptotic Normality: cases of unmet assumptions

$$\begin{bmatrix} u \\ v \end{bmatrix} \Bigg| \rho \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

If v is latent/hidden, then we work with the marginal likelihood $p(u \mid \rho)$:

$$u \mid \rho \sim \text{Normal}(0, 1)$$

Notice that this is free of ρ!

$$p(\rho \mid u) \propto p(u \mid \rho) p(\rho) \propto p(\rho)$$

Here we say the *parameter* is **nonidentified**.

Asymptotic Normality: cases of unmet assumptions

Sometimes it is harder to spot nonidentifiable parameters. It may be the case that $p(y \mid \theta)$ yields the same function in y for two different values of θ. If this is true, then for any particular data set y, $p(y \mid \theta)$ will be equal for these two values of θ.

Example $y \mid \theta \sim \text{Normal}(0, \theta^2)$. Then $p(y \mid \theta) = p(y \mid -\theta)$!

We can fix this easily by restricting the parameter space. The model is no longer underidentified if $\theta \in \mathbb{R}^+$. When this happens, we call this problem **aliasing**.

Asymptotic Normality: cases of unmet assumptions

Another example of **aliasing**. If you look at a histogram of y and it's bimodal, then a possibly suitable model is the **normal mixture model**:

$$
\begin{aligned}
&p(y_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) \\
&= \lambda \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{1}{2\sigma_1^2}(y_i - \mu_1)^2\right] \\
&+(1-\lambda)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{1}{2\sigma_2^2}(y_i - \mu_2)^2\right]
\end{aligned}
$$

Asymptotic Normality: cases of unmet assumptions

When the number of parameters increases with the sample size, the standard asymptotics won't apply. For example, if $p(y_i \mid \theta_i)$ is the likelihood, and θ_i is a different parameter for each datum. This happens with Gaussian Process Models, which we talk about in Chap 21.

Asymptotic Normality: cases of unmet assumptions

Unbounded likelihoods might also be a problem. Assume

$$p(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2}\right].$$

If $y = 0$, then this simplifies to

$$p(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

which goes to ∞ as $\sigma^2 \to 0$. The theoretical probablity of you getting $y = 0$ is obviously 0, but it is possible to get 0s computationally if you have an **underflow** problem. Double precision floating point numbers give you about 15-17 digits of precision.