Chapter 4 :
Asymptotics and connections to non-Bayesian approaches

Kunwoong Kim

September 22, 2020

Seoul National University

Table of Contents

Introduction

- Here, we cover the asymptotic normality of the posterior distribution and their consistency in large samples.
- This provides the connection to non-Bayesian approaches.

Table of Contents

4.1. Normal approximations to the posterior distributions

- Consider a unimodal and symmetric posterior $p(\theta|\mathbf{y})$.

Let $\hat{\theta}$ the mode of the distribution of $\theta|\mathbf{y}$, then by the Taylor's expansion :

$$\log p(\theta|\mathbf{y}) \approx \log p(\hat{\theta}|\mathbf{y}) + (\theta - \hat{\theta})^\top \left[\frac{d}{d\theta}\log p(\theta|\mathbf{y})\right]_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^\top \left[\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{y})\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

Since $\log p(\hat{\theta}|\mathbf{y})$ is constant, $\frac{d}{d\theta}\log p(\theta|\mathbf{y})\Big|_{\theta=\hat{\theta}} = 0$, and $\hat{\theta} \to \theta$

$$p(\theta|\mathbf{y}) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta) := -\frac{d^2}{d\theta^2}\log p(\theta|\mathbf{y}) = -\sum_{i=1}^n \frac{d^2}{d\theta^2}\log p(\theta|y_i)$ is the *observed information*.

4.1. Normal approximations to the posterior distributions

- Under the normal approximation, the posterior is summarized by its mode $\hat{\theta}$ and the curvature of log posterior density $I(\hat{\theta})$.
- Roughly, one can say that $\hat{\theta}$ and $I(\hat{\theta})$ are sufficient statistics.

Example. Normal distribution
Assume a uniform prior for $(\mu, \log \sigma)$.
Let $\mathbf{y} = (y_1, \cdots, y_n) \sim N(\mu, \sigma^2), i.i.d.$
Then, the posterior distribution can be approximated as :

$$p(\mu, \log \sigma | \mathbf{y}) \approx N\left(\begin{pmatrix} \hat{\mu} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix} \right)$$

where $\hat{\mu} = \bar{\mathbf{y}} = \sum_{i=1}^n y_i/n$ and $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \bar{\mathbf{y}})/n$.

Table of Contents

4.2. Large-sample theory

Recall that the posterior distribution is proportional to a multiplication of likelihood and prior.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

4.2. Large-sample theory

If the sample size is large enough, then the likelihood dominates the prior, because :

$$\frac{d^2}{d\theta^2} \log p(\theta|y)\Big|_{\theta=\hat{\theta}} = \frac{d^2}{d\theta^2} \log p(\hat{\theta}) + \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log p(y_i|\theta)\Big|_{\theta=\hat{\theta}}$$

Here, (absolute value of) the term of curvature of the likelihood increases with order n (Appendix B).
Thus if the sample size is large, it dominates the first term of RHS (prior) and else, prior has an impact on the posterior.

Table of Contents

4.3. Counterexamples to the theorems

- Then, what if the prior has an impact of the posterior, even the sample size is large?
- Various counterexamples may exist, here introduces some specific ones.

- Nonidentified

Consider the model

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Assume we only observe u from pair (u, v). Then the parameter ρ is nonidentified.
In other words, since the data supply no information about ρ, the posterior is the same as its prior.

- Aliasing

Let y follows a bimodal Gaussian mixture as the following :

$$\lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(y-\mu_1)^2/2\sigma_1^2} + (1-\lambda) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(y-\mu_2)^2/2\sigma_2^2}$$

This model is not identifiable. Thus we need some assumptions in order to treat the parameter space to be identifiable; for example, $\mu_1 \leq \mu_2$.

Table of Contents

4.4. Frequency evaluations of Bayesian inferences

- Understanding frequentists' estimation as a view of Bayesian.
- The asymptotic properties of estimates from non-Bayesian approaches are also hold for the posterior.

Let $\hat{\theta}$ an estimate (it can be the posterior mean, median or mode) of the true parameter θ_0. Then the following holds under mild regularity conditions and with large sample size.

- Consistency : $\hat{\theta} \to \theta_0$
- Asymptotic unbiasedness : $\big(E(\hat{\theta}|\theta_0) - \theta_0\big)/sd(\hat{\theta}|\theta_0) \to 0$
- Efficiency : $E((\hat{\theta} - \theta_0)^2|\theta_0) \leq E((\theta - \theta_0)^2|\theta_0)$ for all θ.

Table of Contents

4.5. Bayesian interpretations of other statistical models

- What if the number of parameters is large?

Method of inference based on the likelihood alone can be improved if real prior information is availble.
Examples

- Point estimates, confidence intervals
- Hypothesis testing
- Multiple comparisons

Is unbiased estimators good if the sample size is small?

- Minimizing bias often occurs the increases in variance.

Example.

$$\begin{pmatrix} \theta \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 160 \\ 160 \end{pmatrix}, \begin{pmatrix} \sigma & 0.5 \\ 0.5 & \sigma \end{pmatrix} \right)$$

- The posterior mean $E(\theta^{(1)}|y^{(1)}) = 160 + 0.5(y^{(1)} - 160)$ is biased but with repeated sampling $E(y^{(2)}|\theta^{(1)}) = 160 + 0.5(\theta^{(1)} - 160)$, it becomes

$$E(\theta^{(2)}|y^{(2)}) = 160 + 0.5(y^{(2)} - 160)$$

$$\begin{aligned} E(E(\theta^{(2)}|y^{(2)})|\theta^{(1)}) &= 160 + 0.5(E(y^{(2)}|\theta^{(1)}) - 160) \\ &= 160 + 0.25(\theta^{(1)} - 160) \end{aligned}$$

and so on.

- However, $\hat{\theta} = 160 + 2(y - 160)$ is unbiased but with high variance if sample size is small (e.g., if $y = 170$, then $\hat{\theta} = 180$).