Homework 3 for Bayesian Data Analysis

Fan JIN   (2015011506)

April 3, 2018

Question 4.4

- In terms of θ: Under the regularity conditions, we have the asymptotic distribution

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

  where

$$\left[\frac{\mathrm{d}}{\mathrm{d}\,\theta} \log p(\theta|y)\right]_{\theta=\hat{\theta}} = 0,$$

  and

$$I(\theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\,\theta^2} \log p(\theta|y).$$

- In terms of ϕ: Under the regularity conditions, we have the asymptotic distribution

$$p(\phi|y) \to N(\hat{\phi}, [I(\hat{\phi})]^{-1}),$$

  where

$$\left[\frac{\mathrm{d}}{\mathrm{d}\,\phi} \log p(\phi|y)\right]_{\phi=\hat{\phi}} = 0,$$

  and

$$I(\phi) = -\frac{\mathrm{d}^2}{\mathrm{d}\,\phi^2} \log p(\phi|y).$$

It is worth noting that the right hand side is not a normal distribution with fixed mean and variance; instead, the mean $\hat{\theta}$ and variance $[I(\hat{\theta})]^{-1}$ are functions of the sample, and therefore denpend on the sample size. The variance $[I(\hat{\theta})]^{-1}$, moreover, converges to 0 as the sample size increases. In the statement that a nonlinear transform of a normal distribution is no longer normal, however, the topical normal distribution must have fixed parameters. Otherwise, the statement does not hold since its prerequisites are not satisfied.

Question 4.7

The Bayesian posterior estimate is $\mathrm{E}\,(\theta|y)$.

We have the integrated Bayes risk

$$0 \le \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[(\mathop{\mathrm{E}}_{\theta}(\theta|y) - \theta)^2|\theta\right]\right\}$$

$$= \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[(\mathop{\mathrm{E}}_{\theta}(\theta|y))^2|\theta\right]\right\} - \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[\theta \cdot \mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} - \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[\theta \cdot \mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[\theta^2|\theta\right]\right\}$$

$$= \mathop{\mathrm{E}}_{\theta}\left\{\mathop{\mathrm{E}}_{y}\left[(\mathop{\mathrm{E}}_{\theta}(\theta|y))^2|\theta\right]\right\} - \mathop{\mathrm{E}}_{\theta}\left\{\theta \cdot \mathop{\mathrm{E}}_{y}\left[\mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} - \mathop{\mathrm{E}}_{y}\left\{\mathop{\mathrm{E}}_{\theta}\left[\theta \cdot \mathop{\mathrm{E}}_{\theta}(\theta|y)|y\right]\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\theta^2\right\}$$

$$= \mathop{\mathrm{E}}_{y}\left[(\mathop{\mathrm{E}}_{\theta}(\theta|y))^2|\theta\right] - \mathop{\mathrm{E}}_{\theta}\left\{\theta \cdot \mathop{\mathrm{E}}_{y}\left[\mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} - \mathop{\mathrm{E}}_{y}\left\{\mathop{\mathrm{E}}_{\theta}(\theta|y) \cdot \mathop{\mathrm{E}}_{\theta}[\theta|y]\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\theta^2\right\}$$

$$= -\mathop{\mathrm{E}}_{\theta}\left\{\theta \cdot \mathop{\mathrm{E}}_{y}\left[\mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\theta^2\right\}.$$

Assume that it were unbiased, i.e.

$$\int \mathop{\mathrm{E}}_{\theta}(\theta|y) \cdot p(y|\theta_0)\,\mathrm{d}\,y = \theta_0$$

for any θ_0 in the parameter space. Set $\theta_0 = \theta$, and it would give

$$\mathop{\mathrm{E}}_{y}\left[\mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right] = \theta.$$

Thus, the integrated Bayes risk, *if not infinite*, would be

$$-\mathop{\mathrm{E}}_{\theta}\left\{\theta \cdot \mathop{\mathrm{E}}_{y}\left[\mathop{\mathrm{E}}_{\theta}(\theta|y)|\theta\right]\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\theta^2\right\}$$

$$= -\mathop{\mathrm{E}}_{\theta}\left\{\theta \cdot \theta\right\} + \mathop{\mathrm{E}}_{\theta}\left\{\theta^2\right\} = 0.$$

The integrated Bayes risk can be infinite for improper prior distributions, so the result above is not valid at all. In case of proper prior distributions, however, the integrated Bayes risk must be finite and greater than 0 *if not in degenerate problems*, which contradicts the calculation above. This paradox indicates that our assumption is not correct that the Bayesian posterior mean were unbiased.

We therefore conclude that the Bayesian posterior mean cannot be unbiased, giben that the prior distribution is proper (the expectations are finite in our calculation) and the problem is not degenerate (the integrated Bayes risk is not exactly zero).

Question 5.5

For all (i, j), the joint distribution of (θ_i, θ_j) is

$$p(\theta_i, \theta_j) = \int \left[\prod_{k} p(\theta_j|\phi)\right] \cdot p(\phi)\,\mathrm{d}\,\phi \cdot \left[\prod_{k \neq i}^{k \neq j} \mathrm{d}\,\theta_j\right]$$

$$= \int p(\theta_i|\phi)p(\theta_j|\phi) \cdot p(\phi)\,\mathrm{d}\,\phi.$$

From this result, it is clear that θ_i and θ_j are i.i.d. conditioning on ϕ, which means

$$\mathrm{E}(\theta_i|\phi) = \mathrm{E}(\theta_j|\phi) = g(\phi),$$

and

$$\mathrm{E}(\theta_i\theta_j|\phi) = \mathrm{E}(\theta_i|\phi) \cdot \mathrm{E}(\theta_j|\phi),$$

where $g(\phi)$ is a function of the random variable ϕ. Hence,

$$\mathrm{cov}(\theta_i, \theta_j) = \mathrm{E}(\theta_i\theta_j) - \mathrm{E}(\theta_i)\mathrm{E}(\theta_j)$$

$$= \mathrm{E}[\mathrm{E}(\theta_i\theta_j|\phi)] - \mathrm{E}[\mathrm{E}(\theta_i|\phi)] \cdot \mathrm{E}[\mathrm{E}(\theta_j|\phi)]$$

$$= \mathrm{E}[\mathrm{E}(\theta_i|\phi) \cdot \mathrm{E}(\theta_j|\phi)] - \mathrm{E}[\mathrm{E}(\theta_i|\phi)] \cdot \mathrm{E}[\mathrm{E}(\theta_j|\phi)]$$

$$= \mathrm{E}[g(\phi) \cdot g(\phi)] - \mathrm{E}[g(\phi)] \cdot \mathrm{E}[g(\phi)]$$

$$= \mathrm{var}(g(\phi)) \geq 0.$$

Question 5.4a

Since we have no idea which parameter comes from which distribution, the joint distribution of $\theta_1, \cdots, \theta_{2J}$ does not rely on the permutation of indeces:

$$p(\theta_1, \cdots, \theta_{2J}) \propto \sum_p \left[\prod_{k=1}^{J} N(\theta_{p(k)|1,1}) \prod_{k=J+1}^{2J} N(\theta_{p(k)|-1,1}) \right].$$

By definition, $\theta_1, \cdots, \theta_{2J}$ are exchangeable.

Question 5.4b

Proof by contradiction: Assume the distribution of $(\theta_1, \cdots, \theta_{2J})$ could be written as a mixture of i.i.d. components. According to Question 5.5, we would then have

$$\text{cov}(\theta_i, \theta_j) \geq 0.$$

Define random variables (hyperparamters) $\phi_1, \cdots, \phi_{2J}$, half being 1 and the other half being -1. Thus, we have

$$\text{E}(\phi_1|\phi_i = 1) = \cdots = \text{E}(\phi_{i-1}|\phi_i = 1) = \text{E}(\phi_{i+1}|\phi_i = 1) = \cdots \text{E}(\phi_{2J}|\phi_i = 1),$$

and

$$\text{E}(\phi_1 + \cdots + \phi_{i-1} + \phi_{i+1} \cdots \phi_{2J}|\phi_i = 1) = \text{E}((0 - \phi_i)|\phi_i = 1) = -1.$$

Therefore, we have

$$\text{E}(\phi_j \mid \phi_i = 1) = -\frac{1}{2J-1} < 0$$

and likewise

$$\text{E}(\phi_j \mid \phi_i = -1) = \frac{1}{2J-1} > 0.$$

This means

$$\text{cov}(\phi_i, \phi_j) < 0$$

for any $i \neq j$.

The model can then be written as

$$\theta_i|\phi_i \sim N(\phi_i, 1)$$

for any i. By applying the total covariance formula, and noting that $\text{cov}(\theta_i, \theta_j|\phi_i, \phi_j) = 0$ for conditional independence, we have

$$\text{cov}(\theta_i, \theta_j) = \text{E}[\text{cov}(\theta_i, \theta_j|\phi_i, \phi_j)] + \text{cov}[\text{E}(\theta_i|\phi_i, \phi_j) \cdot \text{E}(\theta_j|\phi_i, \phi_j)]$$

$$= \text{E}(0) + \text{cov}(\phi_i, \phi_j) < 0,$$

which contradicts what we concluded in Question 5.5.

Therefore, our assumption cannot be true, and we conclude that the distribution of $(\theta_1, \cdots, \theta_{2J})$ cannot be written as a mixture of i.i.d. components.

Question 5.4c

The de Finetti's theorem holds if we have each sample independently assigned into one of the two categories, whereas in this case we cluster exactly half the samples into one category. As $J \to \infty$,

$$\text{E}(\phi_i|\phi_j) = -\frac{\phi_j}{2J-1} \to 0,$$

which means the correlation between the two samples diminishes. Thus, the difference between the two situations disappears as $J \to \infty$, and therefore this is not a counterexample.