MAST90125: Bayesian Statistical learning

Lecture 5: Properties of Bayesian inference

Prepared by Feng Liu and Guoqi Qian

THE UNIVERSITY OF MELBOURNE

Shrinkage

In the previous lecture, we focused on the mechanics of interpreting the posterior distributions. In this lecture, we will focus more on properties of the posterior.

One important aspect of the transition from the prior to the posterior distribution is that of shrinkage.

- So what is shrinkage?
- Where may you have encountered shrinkage before?

Shrinkage in Bayesian inference

Bayesian inference is probabilistic. In particular, inference based on the posterior is inference based on conditional probability. To understand how this relates to shrinkage, consider

- The law of total expectation:

$$E(\theta) = E_y[E(\theta|y)]$$

- The law of total variance:

$$\mathrm{Var}(\theta) = \mathrm{Var}_y[E(\theta|y)] + E_y[\mathrm{Var}(\theta|y)]$$

Note: You may have also seen this referred to as iterated expectation/variance.

Shrinkage in Bayesian inference

- The law of total expectation implies the prior expectation is equal to the expectation of the posterior expectation taken with respect to y, i.e. averaged over all possible values for y.

- The law of total variance is more important as it implies that the prior variance can be decomposed into two components, both of which must be strictly non-negative. Hence we can write,

  $$\mathrm{Var}(\theta) \geq E_y[\mathrm{Var}(\theta|y)], \text{and } \ \mathrm{Var}(\theta) \geq \mathrm{Var}_y[E(\theta|y)]$$

  meaning that we expect the posterior variance, $\mathrm{Var}(\theta|y)$, on average, to be smaller than the prior variance, $\mathrm{Var}(\theta)$.

- Also we expect the variance of the posterior mean to be less than the prior variance. Hence Bayesian estimation always includes shrinkage.

Example of Shrinkage

- Remember the case of the binomial likelihood with $\text{Be}(\alpha, \beta)$ prior $\Rightarrow$ $p|y \sim \text{Be}(\alpha + y, \beta + n - y)$
- The expectation and variance of a random variable distributed $\text{Be}(\alpha, \beta)$ are

$$E(p) = \frac{\alpha}{\alpha + \beta}, \text{Var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

  - By implication, the posterior expectation and variance is,

$$E(p|y) = \frac{\alpha + y}{\alpha + \beta + n}, \text{Var}(p|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

Example of Shrinkage

- Now to confirm that shrinkage does occur in the case of binomial likelihood with Beta prior, we need to determine $E(y)$ and $\mathrm{Var}(y)$ as functions of α, β. These are:

$$
\begin{aligned}
E(y) &= E_p[E_{y|p}(y)] = nE_p(p) = n\alpha/(\alpha+\beta) \\
\mathrm{Var}(y|p) &= np(1-p) = np - np^2 \\
E_p[\mathrm{Var}(y|p)] &= n(E_p(p) - \mathrm{Var}_p(p) - E_p(p)^2) \\
\mathrm{Var}_p[E_{y|p}(y)] &= n^2\mathrm{Var}_p(p) \\
\mathrm{Var}(y) &= \frac{n\alpha}{\alpha+\beta}\left(1 - \frac{\alpha}{\alpha+\beta}\right) + \frac{n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{n\alpha\beta(\alpha+\beta+1) + n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

Example of Shrinkage

- Taking expectation and variance on both $E(p|y)$ and $\mathrm{Var}(p|y)$, we find

$$
\begin{aligned}
E(p) &= E_y[E(p|y)] = \frac{\alpha + n\alpha/(\alpha+\beta)}{\alpha+\beta+n} = \frac{\alpha(\alpha+\beta)+n\alpha}{(\alpha+\beta)(\alpha+\beta+n)} = \frac{\alpha}{\alpha+\beta} \\
E_y[\mathrm{Var}(p|y)] &= \frac{(\alpha+E(y))(\beta+n-E(y))-\mathrm{Var}(y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\
&= \frac{(\alpha+\frac{n\alpha}{\alpha+\beta})(\beta+n-\frac{n\alpha}{\alpha+\beta})-\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{\frac{\alpha\beta(\alpha+\beta+n)^2}{(\alpha+\beta)^2}-\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\
&= \frac{\frac{\alpha\beta(\alpha+\beta+n)\{(\alpha+\beta+n)(\alpha+\beta+1)-n\}}{(\alpha+\beta)^2(\alpha+\beta+1)}}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{\frac{\alpha\beta(\alpha+\beta+n)(\alpha+\beta+n+1)(\alpha+\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta+n)(\alpha+\beta+1)(\alpha+\beta)} = \frac{\alpha+\beta}{\alpha+\beta+n}\mathrm{Var}(p) \\
&\leq \mathrm{Var}(p)
\end{aligned}
$$

Note on example

- In the example before, we used the marginal expectation and variance of y to determine the presence of shrinkage. From this can you deduce what the prior predictive distribution is?
    - It is Beta-Binomial.
- Next, can you deduce the pattern in shrinkage as $n \rightarrow \infty$?
    - As $n \rightarrow \infty$, the expected posterior variance goes to zero $\Leftrightarrow$ no shrinkage.
- Do you think this result is example specific or more general?
    - Hint: Think about maximum likelihood estimation.

Posterior as $n \to \infty$

- From previously taken subjects, you may be aware that the maximum likelihood estimator is i) consistent and ii) asymptotically normally distributed.
- Do you think this can help our understanding of the posterior distribution? We know that the maximum likelihood estimator $\hat{\theta}_{MLE}$ satisfies,

$$\frac{d \log(p(y_1, \ldots y_n|\theta))}{d\theta}\Big|_{\theta=\hat{\theta}_{MLE}} = \sum_{i=1}^{n} \frac{d \log(p(y_i|\theta))}{d\theta}\Big|_{\theta=\hat{\theta}_{MLE}} = 0 \text{ assuming } |\theta,\ y_i \text{ is } i.i.d.$$

- Now lets do the equivalent to the posterior distribution.

Posterior as $n \to \infty$

- Again assuming given θ, y_i's are $i.i.d.$, the estimator that maximises the posterior, $\hat{\theta}_{MAP}$, should satisfy

$$\begin{aligned}\left.\frac{d\log(p(\theta|y_1,\ldots y_n))}{d\theta}\right|_{\theta=\hat{\theta}_{MAP}} &= \left.\frac{d\{\log(p(y_1,\ldots y_n|\theta))+\log(p(\theta))-\log(p(y))\}}{d\theta}\right|_{\theta=\hat{\theta}_{MAP}}\\ &= \left.\sum_{i=1}^{n}\frac{d\log(p(y_i|\theta))}{d\theta}+\frac{d\log(p(\theta))}{d\theta}\right|_{\theta=\hat{\theta}_{MAP}}\\ &= 0\end{aligned}$$

which looks just like the case for $\hat{\theta}_{MLE}$ except that the information conveyed by the prior is like an additional datapoint. Thus $\hat{\theta}_{MAP}$ is also consistent and asymptotically normal.

Posterior as $n \to \infty$

- However this does not really prove anything about the posterior distribution. To prove something about the posterior, consider a Taylor's expansion,

$$p(\theta|y_1,\dots y_n) = e^{\log(p(\theta|y_1,\dots y_n))}$$

$$= \exp\left(\log(p(\theta|y_1,\dots y_n))|_{\theta=\theta_0} + \sum_{i=1}^{\infty} \frac{d^i \log(p(\theta|y_1,\dots y_n))/d\theta^i\big|_{\theta=\theta_0}(\theta-\theta_0)^i}{i!}\right).$$

which if we choose $\theta_0 = \hat{\theta}_{MAP}$, $p(\theta|y_1,\dots y_n)$ can be written as,

$$\exp\left(\log(p(\hat{\theta}_{MAP}|y_1,\dots y_n)) + \sum_{i=1}^{\infty} \frac{d^i \log(p(\theta|y_1,\dots y_n))/d\theta^i\big|_{\theta=\hat{\theta}_{MAP}}(\theta-\hat{\theta}_{MAP})^i}{i!}\right).$$

Posterior as $n \to \infty$

- We need to focus on the following term,

$$\sum_{i=1}^{\infty} \frac{d^i \log(p(\theta|y_1, \ldots y_n))/d\theta^i\big|_{\theta=\hat{\theta}_{MAP}} (\theta - \hat{\theta}_{MAP})^i}{i!}.$$

Posterior as $n \to \infty$

- We need to focus on the following term,

$$\sum_{i=1}^{\infty} \frac{d^i \log(p(\theta|y_1, \ldots y_n))/d\theta^i\big|_{\theta=\hat{\theta}_{MAP}}(\theta-\hat{\theta}_{MAP})^i}{i!}.$$

- When $i = 1$,

$$\frac{d \log(p(\theta|y_1, \ldots y_n))}{d\theta}\Big|_{\theta=\hat{\theta}_{MAP}}(\theta-\hat{\theta}_{MAP}) = 0 \times (\theta-\hat{\theta}_{MAP}) = 0.$$

Posterior as $n \to \infty$

- We need to focus on the following term,

$$\sum_{i=1}^{\infty} \frac{d^i \log(p(\theta|y_1, \ldots y_n))/d\theta^i|_{\theta=\hat{\theta}_{MAP}}(\theta - \hat{\theta}_{MAP})^i}{i!}.$$

- When $i = 1$,

$$\frac{d \log(p(\theta|y_1, \ldots y_n))}{d\theta}\Big|_{\theta=\hat{\theta}_{MAP}}(\theta - \hat{\theta}_{MAP}) = 0 \times (\theta - \hat{\theta}_{MAP}) = 0.$$

- When $i > 2$ and $k \to \infty$, due to $\hat{\theta}_{MAP}$ is consistent and asymptotically normal,

$$\sum_{i=3}^{k} \frac{d^i \log(p(\theta|y_1, \ldots y_n))}{d\theta^i}\Big|_{\theta=\hat{\theta}_{MAP}}(\theta - \hat{\theta}_{MAP})^i/i! < \frac{Mk}{n^{\frac{3}{2}}} \propto \frac{1}{\sqrt{n}},$$

where $d^i \log(p(\theta|y_1, \ldots y_n))/d\theta^i < M$ for any i.

Posterior as $n \to \infty$

- Since $\hat{\theta}_{MAP}$ is consistent, as $n \to \infty$, the posterior should converge to

$$p(\theta|y_1, \ldots y_n) \quad \to \quad e^{\log(p(\hat{\theta}_{MAP}|y_1,\ldots y_n))} e^{-\frac{(\theta-\hat{\theta}_{MAP})^2}{2(-d^2\log(p(\theta|y_1,\ldots y_n))/d\theta^2|_{\theta=\hat{\theta}_{MAP}})^{-1}}}.$$

which contains the kernel of a normal distribution, indicating that asymptotically

$$p(\theta|y_1, \ldots y_n) \to \mathcal{N}(\hat{\theta}_{MAP}, I(\hat{\theta}_{MAP})^{-1}),$$

where $I(\hat{\theta}_{MAP}) = \frac{-d^2\log(p(\theta|y_1,\ldots y_n))}{d\theta^2}|_{\theta=\hat{\theta}_{MAP}}$.

- This means that asymptotically $E(\theta|y) \to \hat{\theta}_{MAP}$. Moreover, we expect $I(\hat{\theta}_{MAP})$ to increase with n meaning that $E[\text{Var}(\theta|y)] \to 0$ as $n \to \infty$.

Sufficiency principle

- The sufficiency principle states if we have two sets of data $\mathbf{x}$ and $\mathbf{y}$ such that the sufficient statistics $T(\mathbf{x})$ and $T(\mathbf{y})$ are equal, then the inference made about $\boldsymbol{\theta}$ must be the same.
- By the factorisation theorem,

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}|T(\mathbf{y}))g(T(\mathbf{y})|\boldsymbol{\theta})$$

  we know the sufficiency principle should hold for maximum likelihood estimation, since when $T(\mathbf{x}) = T(\mathbf{y})$, $\arg\max_\theta g(T(\mathbf{x})|\boldsymbol{\theta}) = \arg\max_\theta g(T(\mathbf{y})|\boldsymbol{\theta})$.
- However does the sufficiency principle hold for Bayesian inference?
- To check this, consider the posterior $p(\boldsymbol{\theta}|\mathbf{y})$

Sufficiency principle and Bayesian Inference

$$
\begin{aligned}
p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int_{\Theta} p(\mathbf{y}|\theta)p(\theta)d\theta} &= \frac{h(\mathbf{y}|T(\mathbf{y}))g(T(\mathbf{y})|\theta)p(\theta)}{\int_{\Theta} h(\mathbf{y}|T(\mathbf{y}))g(T(\mathbf{y})|\theta)p(\theta)d\theta} \\
&= \frac{g(T(\mathbf{y})|\theta)p(\theta)}{\int_{\Theta} g(T(\mathbf{y})|\theta)p(\theta)d\theta} \\
&= p(\theta|T(\mathbf{y})) \\
&= p(\theta|T(\mathbf{x})) \quad \text{as } T(\mathbf{y}) = T(\mathbf{x}) \\
&= p(\theta|\mathbf{x})
\end{aligned}
$$

- Thus Bayesian inference satisfies the sufficiency principle. Further, conditioning on the data in the posterior distribution is equivalent to conditioning on the sufficient statistics only.

Where have we used/could use the sufficiency principle

- You should use the sufficiency principle in your assignment questions, whenever sensible, in order to simplify algebra, calculation and computation.

- However, we have implicitly used the sufficiency principle in question one of lab two. In the solution, we re-wrote the normal likelihood as a function of the sufficient statistics $\bar{y}, s^2$. Hence, the marginal posterior distribution, assuming a flat prior for μ and $p(\sigma^2) \propto (\sigma^2)^{-1}$, for $\mu|y$ can be written as

$$p(\mu|\bar{y}, s^2) \propto \left(1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2}\right)^{-(n-1+1)/2}$$

and $\sigma^2|y$ as

$$p(\sigma^2|s^2) \propto (\sigma^2)^{-(n-1)/2+1} e^{-\frac{(n-1)s^2}{2\sigma^2}}.$$

Likelihood principle

- The likelihood principle states if two probability models $p(y|\theta, M_1), p(y|\theta, M_2)$ share the same kernel, meaning we can write $L_{M_1}(\theta|y) = c(y)L_{M_2}(\theta|y)$, then any inference about θ should be the same. Note: The frequentist likelihood function $L(\theta|y)$ is a pivoting of the sampling distribution $p(y|\theta)$.

- Bayesian inference obeys the likelihood principle, as

$$
\begin{aligned}
p_{M_1}(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} &= \frac{L_{M_1}(\theta|y)p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L_{M_1}(\theta|y)p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= \frac{c(y)L_{M_2}(\theta|y)p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} c(y)L_{M_2}(\theta|y)p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\
&= p_{M_2}(\boldsymbol{\theta}|\mathbf{y}).
\end{aligned}
$$

- But would this be true of other inferential methods?

Likelihood principle: can it fall down

- The likelihood principle must hold for maximum likelihood, as

$$\arg\max_{\theta} L_{M_1}(\theta|y) = \arg\max_{\theta} c(y)L_{M_2}(\theta|y) = \arg\max_{\theta} L_{M_2}(\theta|y)$$

- But what about small sample hypothesis tests?
- A person tells you they have 12 friends, 9 female and 3 male. We wish to test the hypothesis (choose between models) that friend choice is not influenced by gender, that is $M_1 = H_0 : p_F = 0.5$, against an alternative of bias toward females $M_2 = H_A : p_F > 0.5$. Let $\alpha = 0.05$.
- Two possible sampling models, binomial and negative binomial, satisfy $L_{M_1}(\theta|y) = c(y)L_{M_2}(\theta|y)$ and depend on whether the number of friends total (binomial) or male (negative binomial) is fixed.

Likelihood principle: Negative Binomial vs Binomial

- When assuming a binomial sampling model, the p-value of the test is,

$$\Pr(y \geq 9 | p = 0.5, n = 12) = \sum_{i=9}^{n} \binom{12}{i} 0.5^{i} (1 - 0.5)^{12-i} = 0.0730$$

  - As $0.0730 > 0.05 = \alpha$, there is not sufficient evidence to reject H_0.

- When assuming a negative binomial sampling model, the p-value of the test is,

$$\Pr(y \geq 9 | p = 0.5, r = 3) = \sum_{i=9}^{\infty} \binom{i + 3 - 1}{i} 0.5^{i} (1 - 0.5)^{3} = 0.0327$$

  - As $0.0327 < 0.05 = \alpha$, there is sufficient evidence to reject H_0.

- As the conclusion drawns are contradictory, the likelihood principle does not hold.