Bayesian Statistics
Chapter 7: Model Comparison

Shaobo Jin

Department of Mathematics

Candidate Models

Suppose that we have a set of candidate models

$$\mathcal{M}_k : \ x \sim f_k\left(x \mid \theta_k\right), \qquad \theta_k \in \Theta_k,$$

where $\{f_k\}$ may belong to the same distribution family but have different parameter space, or $\{f_k\}$ may belong to different distribution families.

Example

$$\begin{aligned} H_0 : &\qquad Y \sim N\left(X_1\beta_1, \sigma^2\right) \text{ with conjugate prior} \\ H_1 : &\qquad Y \sim N\left(X_1\beta_1 + X_2\beta_2, \sigma^2\right), \text{ with conjugate prior.} \end{aligned}$$

Or

$$\begin{aligned} H_0 : &\qquad X \sim \text{Binomial}\left(n, p\right), \ p \sim \text{Beta}\left(a_0, b_0\right) \\ H_1 : &\qquad X \sim \text{Poisson}\left(\lambda\right), \ \lambda \sim \text{Gamma}\left(a_1, b_1\right). \end{aligned}$$

Hierarchical Prior

When we compare several candidate models, the model index is also viewed as a parameter. The prior allocation is hierarchical:

1. prior probability for model $\mathcal{M}_k$: $p_k = \mathrm{P}\left(\mathcal{M}_k \text{ is the true model}\right)$,
2. given model $\mathcal{M}_k$, prior $\pi_k\left(\theta_k\right)$ for parameter θ_k in $\mathcal{M}_k$.

The posterior of interest is now

$$\begin{aligned} \mathrm{P}\left(\mathcal{M}_k \mid x\right) &= \frac{p_k \int_{\Theta_k} f_k\left(x \mid \theta_k\right) \pi_k\left(\theta_k\right) d\theta_k}{\sum_j p_j \int_{\Theta_j} f_j\left(x \mid \theta_j\right) \pi_j\left(\theta_j\right) d\theta_j} \\ &\propto \mathrm{P}\left(\mathcal{M}_k\right) f\left(x \mid \mathcal{M}_k\right). \end{aligned}$$

Bayes Factor

We have used the Bayes factor in hypothesis testing. It can also be used in model comparison.

Definition

For two Bayes models $\mathcal{M}_1$ and $\mathcal{M}_2$, the Bayes factor is

$$\begin{aligned} B_{12} &= \frac{\mathrm{P}\left(\mathcal{M}_1 \mid x\right)/\mathrm{P}\left(\mathcal{M}_2 \mid x\right)}{\mathrm{P}\left(\mathcal{M}_1\right)/\mathrm{P}\left(\mathcal{M}_2\right)} \\ &= \frac{\int_{\Theta_1} f_1\left(x \mid \theta_1\right)\pi_1\left(\theta_1\right)d\theta_1}{\int_{\Theta_2} f_2\left(x \mid \theta_2\right)\pi_2\left(\theta_2\right)d\theta_2}, \end{aligned}$$

where

$$m_k\left(x\right) = \int_{\Theta_k} f_k\left(x \mid \theta_k\right)\pi_k\left(\theta_k\right)d\theta_k$$

is the marginal likelihood under model k.

Compute Bayes Factor: Example

It is super important to keep in mind that we cannot use $\propto$ anymore. We must keep track of all normalizing constants!

Example

Suppose that we have randomly chosen n patients and analyzed their blood samples in order to test drug resistance. Let X be the number of patients with positive test result. Two models are under consideration

$$\begin{aligned} \mathcal{M}_1: &\quad X \sim \text{Binomial}\left(n, p\right), \ p \sim \text{Beta}\left(a_0, b_0\right) \\ \mathcal{M}_2: &\quad X \sim \text{Poisson}\left(\lambda\right), \ \lambda \sim \text{Gamma}\left(a_1, b_1\right). \end{aligned}$$

Compute the Bayes factor.

Bayes Factor for Nested Linear Model

Suppose that we want to compare two nested linear regression models

$$\begin{aligned} \mathcal{M}_1 : \quad & Y \sim N_n\left(X_1\beta_1, \sigma^2 I_n\right),\ \left(\beta_1, \sigma^2\right) \sim \text{NIG}\left(a_0, b_0, \tau_0, \Omega_0^{-1}\right), \\ \mathcal{M}_2 : \quad & Y \sim N_n\left(X\beta, \sigma^2 I_n\right),\ \left(\beta, \sigma^2\right) \sim \text{NIG}\left(a_0, b_0, \mu_0, \Lambda_0^{-1}\right), \end{aligned}$$

where $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ such that $X\beta = X_1\beta_1 + X_2\beta_2$.

- Comparing $\mathcal{M}_1$ and $\mathcal{M}_2$ is the same as testing $\beta_2 = 0$.

Example

Compute the Bayes factor for the above nested linear regression models.

Reminder

Keep in mind that using the Bayes factor to conduct model selection and hypothesis testing share many things in common.

- We need to be careful when using improper prior with Bayes factors, as the Bayes factor can be biased towards certain model.
- The Jeffreys-Lindley paradox says that an improper prior cannot be approximated by priors with increasing variances.

Bayesian Information Criterion

Using the marginal likelihood under model k, we call

$$2\log m_k(x) = 2\log\left[\int_{\Theta_k} f_k(x \mid \theta_k)\,\pi_k(\theta_k)\,d\theta_k\right]$$

the exact Bayesian information criterion (BIC) values.

- However, it is rarely computed in practice due to the complexity of the integral.

Definition

The Bayesian information criterion (BIC), aka, Schwartz's criterion, is

$$\mathrm{BIC} = -2\max_{\theta}\log f(x \mid \theta) + p\log(n),$$

where n is the sample size and p is the dimension of θ.

Derivation of BIC

We approximate $m_k\left(x\right)$ by the Laplace approximation and obtain

$$m_k\left(x\right) \approx (2\pi)^{p_k/2}\sqrt{\det\left(\left[\frac{\partial^2 - \log f_k\left(x \mid \hat{\theta}_k\right)}{\partial\theta_k\partial\theta_k^T}\right]^{-1}\right)} f\left(x \mid \hat{\theta}_k\right)\pi_k\left(\hat{\theta}_k\right)$$

where $\hat{\theta}_k$ is the MLE under $\mathcal{M}_k$ such that $\frac{\partial \log f_k\left(x|\hat{\theta}_k\right)}{d\theta_k} = 0$. Hence,

$$-2\log m_k\left(x\right) \approx \underbrace{-2\log f\left(x \mid \hat{\theta}_k\right) + p_k \log n}_{\mathrm{BIC}_k} + O_P\left(1\right).$$

It is interesting to see that the prior vanishes in BIC.

BIC: Example

Example

Suppose that $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$ and $\pi\left(\beta, \sigma^2\right) = \sigma^{-2}$. Find the BIC.

Implications of Approximation

1. Since the Bayes factor is $B_{12} = \frac{m_1(x)}{m_2(x)}$, then

$$2 \log B_{12} \quad \approx \quad \mathrm{BIC}_2 - \mathrm{BIC}_1.$$

   We favor $\mathcal{M}_2$ if B_{12} is small, or equivalently $m_2(x)$ is large. This is similar to favoring $\mathcal{M}$ if BIC_2 is small.

2. BIC penalizes the log-likelihood with $p \log n$.
   - For nested models $\mathcal{M}_1 \subset \mathcal{M}_2$ such that $\theta_1 \subset \theta_2$,

$$\mathrm{P}_{\mathcal{M}_1}\left(-2\log\left[\frac{f\left(x \mid \hat{\theta}_1\right)}{f\left(x \mid \hat{\theta}_2\right)}\right] < c \mid \theta_1\right) \quad \to \quad \mathrm{P}_{\mathcal{M}_1}\left(\chi^2_{p_2 - p_1} < c \mid \theta_1\right),$$

     as $n \to \infty$. Without the penalization term, the probability of choosing the correct model is not 1.

Effect of Penalization

Consider the information criterion of the form

$$\mathrm{IC}_k = -2\sum_{i=1}^{n}\log f_k\left(x_i \mid \hat{\theta}_k\right) + c_k,$$

where $c_k > 0$. We select the candidate model that has the smallest information criterion.

Theorem (Weakly Consistency)

Suppose that there is only one candidate model that minimizes the Kullback-Leibler divergence to the true model. If $c_k = o_P(n)$, then the information criterion selects such closest model with probability approaching 1 as $n \to \infty$.

Both Akaike information criterion (AIC, $c_k = 2p_k$) and BIC ($c_k = p\log n$) are weakly consistent.

Extent of Penalization

Theorem (Consistency)

Suppose that there are several candidate models that minimize the Kullback-Leibler divergence. Let $\mathcal{J}$ be the set of indices of candidate models which all reach the minimum Kullback-Leibler divergence to the true model, and $\mathcal{J}_0$ be the subset of $\mathcal{J}$ containing the smallest dimensions. Suppose that, for any $j_0 \in \mathcal{J}_0$ and $j \in \mathcal{J} \setminus \mathcal{J}_0$, we have

$$P\left(c_j - c_{j_0} \to \infty,\ as\ n \to \infty\right) = 1.$$

Then the information criterion selects the most parsimonious model from the closest models with probability approaching 1.

AIC ($c_k = 2p_k$) does not fulfill the condition of the theorem, whereas BIC ($c_k = p \log n$) is consistent.

Deviance

Definition

The deviance for a given model and given data x is $D(\theta) = -2\log f(x \mid \theta) + \text{constant}$, where the constant is the same in all candidate models, representing the log-likelihood of the perfectly fitted model.

- The difference in the deviance can be used to compare the fit of candidate models to the data such as

$$D\left(\hat{\theta}_1\right) - D\left(\hat{\theta}_2\right) = 2\left[\log f\left(x \mid \hat{\theta}_2\right) - \log f\left(x \mid \hat{\theta}_1\right)\right].$$

- A Bayesian version such deviance difference is

$$D(\theta) - D(\mathrm{E}[\theta \mid x]) = 2\left[\log f(x \mid \mathrm{E}[\theta \mid x]) - \log f(x \mid \theta)\right].$$

Deviance Information Criterion

Let

$$\begin{aligned} p_D &= \mathrm{E}\left[D\left(\theta\right) - D\left(\mathrm{E}\left[\theta \mid x\right]\right) \mid x\right] \\ &= \mathrm{E}\left[D\left(\theta\right) \mid x\right] - D\left(\mathrm{E}\left[\theta \mid x\right]\right) \end{aligned}$$

be the posterior expected value of the deviance difference.

Definition

The deviance information criterion (DIC) is

$$\begin{aligned} \text{DIC} &= \mathrm{E}\left[D\left(\theta\right) \mid x\right] + p_D \\ &= D\left(\mathrm{E}\left[\theta \mid x\right]\right) + 2p_D \\ &= -4\mathrm{E}\left[\log f\left(x \mid \theta\right) \mid x\right] + 2\log f\left(x \mid \mathrm{E}\left[\theta \mid x\right]\right) + \text{constant}. \end{aligned}$$

We choose the model with a smaller value of DIC.

Example: DIC for Linear Model

Example

Consider the linear regression model

$$Y \sim N_n\left(X\beta, \sigma^2 I_n\right), \qquad \left(\beta, \sigma^2\right) \sim \mathrm{NIG}\left(a_0, b_0, \mu_0, \Lambda_0^{-1}\right).$$

Compute the DIC. It is known that, if $\sigma^2 \sim \mathrm{InvGamma}\,(a, b)$, then

$$\begin{aligned} \mathrm{E}\left[\sigma^2\right] &= \frac{b}{a-1}, \text{ if } a > 1 \\ \mathrm{E}\left[\log\sigma^2\right] &= \log\left(b\right) - \psi\left(a\right), \\ \mathrm{E}\left[\sigma^{-2}\right] &= \frac{a}{b}. \end{aligned}$$

where ψ is the digamma function.

DIC and AIC

Suppose that posterior can be well approximated by a normal distribution

$$\theta \mid x \quad \approx \quad N_p \left(\hat{\theta}, \ \left[- \frac{\partial^2 \log f\left(x \mid \hat{\theta}\right)}{\partial \theta \partial \theta^T} \right]^{-1} \right),$$

where $\hat{\theta}$ is the MLE. Then, we can approximate $\mathrm{E}\left[\theta \mid x\right]$ by $\hat{\theta}$.

- We will investigate such normal approximation in a later lecture.

We can show that

$$\text{DIC} \quad \approx \quad -2 \log f\left(x \mid \hat{\theta}\right) + 2p + \text{constant}.$$

which is equivalent to AIC.

Computational Technique

To compute DIC, we need to evaluate the integrals

$$\begin{aligned} \mathrm{E}\left[D\left(\theta\right) \mid x\right] &= \int D\left(\theta\right) \pi\left(\theta \mid x\right) d\theta, \\ D\left(\mathrm{E}\left[\theta \mid x\right]\right) &= D\left(\int \theta \pi\left(\theta \mid x\right) d\theta\right). \end{aligned}$$

The integrals are generally intractable and need to be evaluated numerically.

We can for example draw posterior samples θ_1, ..., θ_T by independent MC or MCMC and approximate the integrals by

$$\begin{aligned} \int D\left(\theta\right) \pi\left(\theta \mid x\right) d\theta &\approx \frac{1}{n} \sum_{t=1}^{T} D\left(\theta_i\right), \\ \int \theta \pi\left(\theta \mid x\right) d\theta &\approx \frac{1}{n} \sum_{t=1}^{T} \theta_i. \end{aligned}$$

Combining Different Models

A different view relative to model selection is to combine the contributions of several models as in ensemble learning. Let Δ be the quantity of interest such as

- average treatment effect of a drug,
- a future value.

Bayesian model selection chooses a model k^* using $\mathrm{P}\left(\mathcal{M}_k \mid x\right)$ and estimates Δ by

$$\hat{\Delta} = \mathrm{E}\left[\Delta \mid x, \mathcal{M}_{k^*}\right].$$

Bayesian model averaging (BMA) takes a weighted average instead as

$$\hat{\Delta} = \sum_k \mathrm{E}\left[\Delta \mid x, \mathcal{M}_k\right] \mathrm{P}\left(\mathcal{M}_k \mid x\right).$$

Posterior of Δ

The posterior of Δ is given by

$$f\left(\Delta \mid x\right) = \sum_{k} f\left(\Delta \mid \mathcal{M}_k, x\right) \mathrm{P}\left(\mathcal{M}_k \mid x\right),$$

where $f\left(\Delta \mid \mathcal{M}_k, x\right)$ is the posterior of Δ under model k. The posterior mean of Δ is

$$\begin{aligned} \mathrm{E}\left[\Delta \mid x\right] &= \int \Delta \sum_{k} f\left(\Delta \mid \mathcal{M}_k, x\right) \mathrm{P}\left(\mathcal{M}_k \mid x\right) d\Delta \\ &= \sum_{k} \mathrm{P}\left(\mathcal{M}_k \mid x\right) \underbrace{\int \Delta f\left(\Delta \mid \mathcal{M}_k, x\right) d\Delta}_{=\mathrm{E}[\Delta \mid x, \mathcal{M}_k]}, \end{aligned}$$

which is the BMA estimator of Δ.

Three Scenarios

The posterior probability

$$\mathrm{P}\left(\mathcal{M}_k \mid x\right) = \mathrm{P}\left(\mathcal{M}_k \text{ is the true model} \mid x\right)$$

can be strange if all candidate models are wrong, especially when we need to specify the prior probability that $\mathcal{M}_k$ is the true model.

1. The $\mathcal{M}$-closed setting means that one of the candidate models is the true data generating process.
2. The $\mathcal{M}$-complete setting means that but the true data generating process can be conceptualized, but it is not one of the candidate models due to, for example, model complexity or lack of information.
3. The $\mathcal{M}$-open setting means that the data generating process cannot be conceptualized and all candidate models are wrong.

Bayesian Stacking

Let $S\left(P,Q\right)$ be a scoring rule to measure the similarity between two probability measure P and Q. Let p and q be the corresponding densities. Then,

$$S\left(P,Q\right) \quad = \quad \int s\left(P,w\right)q\left(w\right)dw,$$

for some function $s\left(\cdot,\cdot\right)$. Bayesian stacking maximizes such similarity

$$S\left(\sum_k w_k f\left(\tilde{x}\mid x,\mathcal{M}_k\right),\ f_{\text{true}}\left(\tilde{x}\mid x\right)\right)$$

with respect to weights $\{w_k\}$ under the restriction that

$$\sum_k w_k = 1,\ 0\leq 1w_k\leq 1,\ \forall k.$$

Scoring Rule

Two commonly used scoring rules are

1. log score: $s(P, x) = \log p(x)$ such that $S(Q, Q) - S(P, Q)$ is the KL divergence.
    - Taking $p(x) = \sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k)$ is the same as maximizing the similarity between the stacked predictive distribution and the true predictive distribution.
2. energy score: $s(P, x) = \frac{1}{2} \mathrm{E}_P \left[\left\| X - \tilde{X} \right\|^{\beta} \right] - \mathrm{E}_P \left[\| X - x \|^{\beta} \right]$, where X and $\tilde{X}$ are two iid random variables, and the expectations are taken with respect to P.
    - If $\beta = 2$, it reduces to $s(P, x) = - \| \mathrm{E}_P [X] - x \|^2$.
    - Maximizing the scoring rule is equivalent to minimizing the squared prediction error.

Leave-One-Out Cross Validation

However, we don't know $f_{\text{true}}\left(\tilde{x} \mid x\right)$ that is needed to evaluate

$$S\left(\sum_k w_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right),\ f_{\text{true}}\left(\tilde{x} \mid x\right)\right).$$

One alternative is to use leave-one-out cross validation as

$$\min_w \frac{1}{n}\sum_{i=1}^n s\left(\sum_k w_k f\left(x_i \mid x_{-i}, \mathcal{M}_k\right),\ x_i\right),$$

where

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int f\left(x_i \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)d\theta_k.$$

The stacked estimate of the predictive density is

$$\hat{f}\left(\tilde{x} \mid x\right) = \sum_k \hat{w}_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right).$$

BMA and Bayesian Stacking

For BMA, it is alleged that, as $n \to \infty$,

- if one of the candidate models is the true model, say $\mathcal{M}_{k^*}$ is the true model,
- or, if all candidate models are misspecified and $\mathcal{M}_{k^*}$ has the smallest Kullback-Leibler divergence to the true model,

then $\mathrm{P}\left(\mathcal{M}_{k^*} \mid x\right) \to 1$.

In contrast to Bayesian model averaging,

- no prior $\mathrm{P}\left(\mathcal{M}_k\right)$ is needed in Bayesian stacking.
- Bayesian stacking is intended for the case where all candidate models are misspecified.

Importance Sampling

It is computationally intensive to compute the leave-one-out (LOO) predictive density

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k$$

for each i, because we have to refit $\mathcal{M}_k$ n times to obtain all $\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)$.

Suppose that, for each k, we fit $\mathcal{M}_k$ using all the data and obtain L draws from the posterior $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$. Then,

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \pi\left(\theta_k \mid x, \mathcal{M}_k\right) d\theta_k,$$

where the importance weight is

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)}.$$

Normalized Importance Sampling

We can rewrite the importance weight as

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \quad \propto \quad \frac{f\left(x_{-i} \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)}{f\left(x \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)}$$

$$\propto \quad \frac{1}{f\left(x_i \mid \theta_k, \mathcal{M}_k\right)}.$$

The normalized importance sampling estimator is

$$\hat{f}^{\text{NIS}}\left(x_i \mid x_{-i}, \mathcal{M}_k\right) \quad = \quad \frac{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right) p\left(x_i \mid \theta_k, \mathcal{M}_k\right)}{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)},$$

where we sample $\theta_k^{(l)}$, $l = 1, ..., L$, from $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$.

However, the importance weights can be unstable if the distribution has a long tail that makes the importance weight very large.

Generalized Pareto Distribution

Theorem

Under suitable conditions on the random variable X, if the threshold u_0 is high enough, the conditional distribution of $X \mid X > u$ converges to a three-parameter generalized Pareto distribution (GPD), as $u \to \infty$. Its density is given by

$$f\left(x; u, \sigma, k\right) \quad = \quad \begin{cases} \frac{1}{\sigma}\left[1 + \frac{k(x-u)}{\sigma}\right]^{-1-1/k}, & \text{if } c \neq 0, \\ \frac{1}{\sigma}\exp\left(\frac{x-u}{\sigma}\right), & \text{if } c = 0, \end{cases}$$

for $y > u$ and $\sigma > 0$.

Pareto Smoothed Importance Sampling

Pareto Smoothed Importance Sampling stabilizes the large importance weights. Without loss of generality, suppose that $\left\{w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}$ has been ordered in increasing order.

- Consider the largest $N = \left\lfloor \min\left(0.2L,\, 3\sqrt{L}\right)\right\rfloor$ importance weights.
- We fit a GPD to $\left(w_i\left(\theta_k^{(L-N+1)}, \mathcal{M}_k\right), ..., w_i\left(\theta_k^{(L)}, \mathcal{M}_k\right)\right)$ with $u = w_i\left(\theta_k^{(L-N)}, \mathcal{M}_k\right)$.
- These N tail importance weights are replaced by

$$\min\left\{F^{-1}\left(\frac{z-1/2}{M}\right),\, \max_i w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}, \qquad z = 1, ..., M,$$

  where F^{-1} is the inverse distribution function of the fitted GPD. The other importance weights are unchanged.