Model Comparison

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

02/19/2025

Multiple hypotheses/models

- Bayesian framework does not treat H_0 and H_a differently
- Methodology may be extended to more than two conclusions
- Instead of "hypotheses", compare evidence for "models"
- For data $\mathbf{y}$, models $M_1, \ldots, M_m$:
  - M_i: $\mathbf{y} \sim p(\mathbf{y} \mid \theta_i, M_i)$, with prior $\theta_i \sim p(\theta_i \mid M_i)$
  - With prior probabilities $P(M_i)$:

$$P(M_1) + \ldots + P(M_m) = 1.$$

Multiple hypotheses/models

- The posterior probability of model i is

$$p(M_i \mid \mathbf{y}) = \frac{P(M_i)p(\mathbf{y} \mid M_i)}{\sum_{j=1}^{m} P(M_j)p(\mathbf{y} \mid M_j)}$$

where

$$p(\mathbf{y} \mid M_i) = \int p(\mathbf{y} \mid \theta_i, M_i)p(\theta_i \mid M_i)\, d\theta_i.$$

Model choice

- Actions $\mathcal{A} = \{M_1, \ldots, M_m\}$
- Under "$0-1$" loss,

$$l(M_i, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq M_i\}}$$

  - Choose M_i with highest posterior probability $P(M_i \mid \mathbf{y})$
- Under "$0-c_i$" loss,

$$l(M_i, d(\mathbf{y})) = c_i \mathbb{1}_{\{d(\mathbf{y}) \neq M_i\}}$$

  - Posterior risk for choosing M_i is

$$\rho(p_\theta, a = M_i) = \sum_{j \neq i} c_j P(M_j \mid \mathbf{y})$$

  - Choose M_i with highest weighted posterior $c_i P(M_i \mid \mathbf{y})$

Example: Placement test

- Reading ability is scaled to have a $\text{Normal}(100, 225)$ distribution over a student population
- A test assesses ability with normal error variance 64.
- Observe the test score y for a student
  - $p(y \mid \mu) = \text{Normal}(\mu, 64)$
  - $p(\mu) = \text{Normal}(100, 225)$
- The posterior distribution for their true ability is
  - $p(\mu \mid \mathbf{y}) = \text{Normal}(22.15 + 0.779\, y\ ,\ 49.83)$

Example: Placement test

- A given student belongs to the
  - remedial learning group if $\mu < 80$ (R)
  - standard learning group if $80 < \mu < 120$ (S)
  - accelerated learning group if $\mu > 120$ (A).
- Assume that a student has score $y = 75$
  - $p(\mu \mid \mathbf{y}) = \text{Normal}(80.56\,,\,49.83)$
- Then, their posterior probability of belonging to each group is
  - $P(R \mid y = 75) = 0.468$
  - $P(S \mid y = 75) = 0.532$
  - $P(A \mid y = 75) \approx 0$

    http://www.ericfrazerlock.com/Model_Comparison_Rcode1.r

Example: Placement test

- Assign loss functions
  - $l(R, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq R\}}$
  - $l(S, d(\mathbf{y})) = 2 \cdot \mathbb{1}_{\{d(\mathbf{y}) \neq S\}}$
  - $l(A, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq A\}}$
- For $y = 75$:
  - $2P(S \mid y = 75) = 1.064 > P(R \mid y = 75) = 0.468$, and
  - $2P(S \mid y = 75) = 1.064 > P(A \mid y = 75) \approx 0$, so
  - So choose the standard group (S).

Example: Placement test

- Decision rule for arbitrary y:

$$d(y) = \begin{cases} R & \text{if } y < 70.4 \\ S & \text{if } 70.4 \leq y \leq 129.6 \\ A & \text{if } y > 129.6 \end{cases}$$

Bayes factors for model comparison

- Recall the Bayes factor for model M_1 over model M_2 is

$$BF = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)}$$

- A likelihood ratio test is based on maximum for each model:

$$\Lambda = \frac{\max_{\theta_1} p(\mathbf{y} \mid \theta_1, M_1)}{\max_{\theta_2} p(\mathbf{y} \mid \theta_2, M_2)}$$

- Under point models $M_1 : \theta = \theta^{(1)}$ and $M_2 : \theta = \theta^{(2)}$:

$$BF = \Lambda = \frac{p(\mathbf{y} \mid \theta^{(1)})}{p(\mathbf{y} \mid \theta^{(2)})}$$

Bayesian Information Criterion

- Let p_i be number of parameters in model M_i
- Let n be the data sample size
- A heuristic for assessing the fit of a model is the *Bayesian Information Criterion* (BIC):

$$BIC(M_i) = -2\log(\max_{\theta_i} p(\mathbf{y} \mid \theta_i, M_i)) + p_i \log n,$$

  - Smaller values are preferred
  - log likelihood, with penalty for the dimension of the model

Bayesian Information Criterion

- Likelihood ratio test usually based on transformed ratio

$$W = -2\log\left[\frac{\max_{\theta_1} p(\mathbf{y} \mid \theta_1, M_1)}{\max_{\theta_2} p(\mathbf{y} \mid \theta_2, M_2)}\right]$$

- The difference in BIC can be expressed in terms of W:

$$\Delta BIC = W - (p_2 - p_1)\log n,$$

  - Δ denotes change (from M_1 to M_2)
  - The likelihood ratio statistic corrected for dimension of each model

Bayesian Information Criterion

- For $\mathbf{y} = y_1, y_2, \ldots y_n$ iid, as $n \to \infty$,

$$-2\log(BF) \approx \Delta BIC$$

  under mild assumptions.

- Derivation:
  https://statproofbook.github.io/P/bic-der.html

- ΔBIC may be easier to compute than the BF

- ΔBIC does not depend on prior distributions

- *BIC* also called *Schwarz information criterion* for G. Schwarz
  - Original article:
    http://projecteuclid.org/euclid.aos/1176344136

Partial Bayes factors

- If $p(\theta_i \mid M_i)$ is improper, then so is

$$p(\mathbf{y} \mid M_i) = \int p(\mathbf{y} \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i$$

  so Bayes factors involving M_i not well defined.

- Possible solution:
  - Assume $p(\theta_1 \mid \mathbf{y}_1)$ is proper for $\mathbf{y}_1 = (y_1, \ldots, y_i)$
  - Find conditional Bayes factor for $\mathbf{y}_2 = (y_{i+1}, \ldots, y_n)$

$$BF(\mathbf{y}_2 \mid \mathbf{y}_1) = \frac{p(\mathbf{y}_2 \mid \mathbf{y}_1, M_1)}{p(\mathbf{y}_2 \mid \mathbf{y}_1, M_2)}$$

  - This is a *Partial Bayes factor*

Example: traffic accidents

- Would like to estimate weekly accident rate at new traffic intersection.
- Each week observe $y \sim \text{Poisson}(\lambda)$ accidents
- M_1:Elicited prior from city planner: $p_1(\lambda) = \text{Gamma}(3, 2)$.
- M_2: Compare with (improper) uniform prior $p_2(\lambda) = 1$.
- Observe data for 5 weeks:
  - $y_1 = 3,\ y_2 = 6,\ y_3 = 2,\ y_4 = 4,\ y_5 = 2$

Poisson-Gamma marginal

- If $y_1, \ldots, y_n \stackrel{iid}{\sim} \text{Poisson}(\lambda)$ and $p(\lambda) = \text{Gamma}(\alpha, \beta)$,

$$P(\mathbf{y}) = \frac{\beta^\alpha \Gamma(\sum y_i + \alpha)}{\Gamma(\alpha) \prod y_i! (\beta + n)^{\sum y_i + \alpha}}$$

Example: traffic accidents

- $p(\mathbf{y} \mid M_2)$ is improper

- Condition on y_1:
  - $p(\lambda \mid M_1, y_1) = \text{Gamma}(y_1 + 3, 3)$
  - $p(\lambda \mid M_2, y_1) = \text{Gamma}(y_1 + 1, 1)$

Example: traffic accidents

- Compute partial Bayes factor, conditioned on y_1:
  - $p(y_2 = 6, y_3 = 2, y_4 = 4, y_5 = 2 | M_1, y_1 = 3) = 0.000133$
  - $p(y_2 = 6, y_3 = 2, y_4 = 4, y_5 = 2 | M_2, y_1 = 3) = 0.000224$
  - The partial BF for M1 over M2 is

$$BF(y_2, y_3, y_4, y_5 \mid y_1) = 0.596$$

  http://www.ericfrazerlock.com/Model_Comparison_Rcode2.r

- Modest evidence that the elicited prior is not better than flat prior

Intrinsic Bayes factors

- Compute n partial Bayes factors:

$$BF(\{y_j\}_{j \neq i} \mid y_i)$$

  for $i = 1, \ldots, n$

- The average of these partial BFs is the *intrinsic Bayes factor*
  - Could take arithmetic or geometric average
  - If $BF(\{y_j\}_{j \neq i} \mid y_i)$ does not exist, condition on larger subsets instead

- The traffic accident example has arithmetic intrinsic Bayes factor 1.64. http://www.ericfrazerlock.com/Model_Comparison_Rcode2.r

Fractional Bayes Factors

- An alternative to intrinsic *BF* is the *fractional Bayes factor*:

$$BF_b = \frac{p(\mathbf{y}, b \mid M_1)}{p(\mathbf{y}, b \mid M_2)}$$

where

$$p(\mathbf{y}, b \mid M_i) = \frac{\int p(\mathbf{y} \mid \theta_i, M_i) p(\theta_i \mid M_i)\, d\theta_i}{\int p(\mathbf{y} \mid \theta_i, M_i)^b p(\theta_i \mid M_i)\, d\theta_i}$$

for $b \in (0, 1)$.

- Often choose $b = 1/n$ if $BF_{1/n}$ is well-defined
- Fractional BF satisfies likelihood principle, intrinsic BF does not.

Fractional Bayes Factors

- Note that

$$p(\mathbf{y}, b \mid M_i) = \int p(\mathbf{y} \mid \theta_i, M_i)^{1-b} p(\theta_i \mid \mathbf{y}, b, M_i)\, d\theta_i$$

  where

$$p(\theta_i \mid \mathbf{y}, b, M_i) \propto p(\mathbf{y} \mid \theta_i, M_i)^b p(\theta_i \mid M_i)$$

- For $\{y_j\}_{j=1}^n$ iid given θ_i,

$$p(\mathbf{y} \mid \theta_i, M_i)^b = \left[\prod_{j=1}^{n} p(y_j \mid \theta_i, M_i)\right]^b$$

  - So $b = 1/n$ gives the geometric mean for the likelihood of one observation.

Example: traffic accidents (continued)

- Note that

$$p(\lambda \mid \mathbf{y}, 1/n, M_2) = \text{Gamma}(\bar{y} + 1, 1),$$

which gives

$$p(\mathbf{y}, 1/n \mid M_2) = \frac{\Gamma(\sum y_i + 1)}{(\prod y_i!)^{\frac{n-1}{n}} \Gamma(\bar{y} + 1)(n)^{\sum y_i + 1}}$$

Example: traffic accidents (continued)

- Similarly,

$$p(\mathbf{y}, 1/n \mid M_1) = \frac{3^{\bar{y}+3}\Gamma(\sum y_i + 3)}{(\prod y_i!)^{\frac{n-1}{n}}\Gamma(\bar{y}+3)(2+n)^{\sum y_i + 3}}$$

- For 5 weeks data, the fractional BF for M_1 over M_2 is

$$BF_{1/5} = 1.28$$

http://www.ericfrazerlock.com/Model_Comparison_Rcode2.r