Bayesian model comparison with log scores

STAT 206
11 Mar 21

Lecture

$D = \underset{\sim}{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $y_i \in \mathbb{R}$

$$M_j : \begin{cases} (\underset{\sim}{\theta}_j \mid [PM]\, B) \sim p(\underset{\sim}{\theta}_j \mid [PM]\, B) \\ (Y_i \mid [SM: M_j]\, \underset{\sim}{\theta}_j\, B) \overset{IID}{\sim} p(y_i \mid [SM: M_j]\, \underset{\sim}{\theta}_j\, B) \\ (i = 1, \ldots, n) \end{cases} \quad (1)$$

$\underset{\sim}{\theta}_j \in \Theta_j \subseteq \mathbb{R}^{k_j}$ PDF

As noted last time, log scores derive from the

Prediction Principle Good models make good predictions; Bad ——— bad ———;

that's how we know a model is good or bad

(+) predictions of data not used in the model-fitting process

Bayesian prediction of a new data value in the setting above involves calculating the posterior predictive distribution $p(y^* \mid \underset{\sim}{y}, M_j, B)$ for a new data value y^*

$$p(y^* | \underset{\sim}{y}, M_j, B) = \int_{\Theta_j} p(y^*, \underset{\sim}{\theta}_j | \underset{\sim}{y}, M_j, B)\, d\underset{\sim}{\theta}_j$$

(PDF) ↑

$$= \int_{\Theta_j} p(y^* | \underset{\sim}{\theta}_j, \cancel{\underset{\sim}{y}}, M_j, B) \cdot p(\underset{\sim}{\theta}_j | \underset{\sim}{y}, M_j, B)\, d\underset{\sim}{\theta}_j$$

$$= \int_{\Theta_j} p(y^* | \underset{\sim}{\theta}_j, M_j, B) \; p(\underset{\sim}{\theta}_j | \underset{\sim}{y}, M_j, B)\, d\underset{\sim}{\theta}_j$$

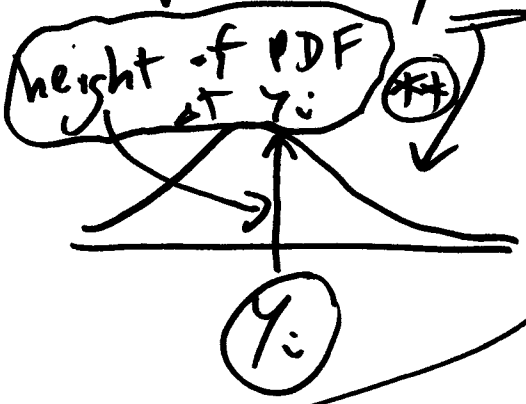
- $p(y^* | \underset{\sim}{\theta}_j, M_j, B)$: sampling PDF for a new data value
- $p(\underset{\sim}{\theta}_j | \underset{\sim}{y}, M_j, B)$: (PDF) posterior distribution for $\underset{\sim}{\theta}_j$ given the data & M_j & B

This looks formidable but is actually easy to simulate from in MCMC (see below)

OK, suppose that we have $p(y^* | \underset{\sim}{y}, M_j, B)$; {how can we use it in model comparison?}

① Let's imagine jack-knifing the data set to get (data not used in the fitting process): define $\underset{\sim}{y}_{[-i]} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{pmatrix}$ i.e. $\underset{\sim}{y}$ with y_i set aside

② Now locate y_i in $p(y_i \mid \underset{\sim}{y}_{[-i]}, \beta)$:



Q: How to make a good comparison between y_i (a scalar) and $p(y_i \mid \underset{\sim}{y}_{[-i]}, \beta)$ (a PDF)?

A: This is called finding a good scoring rule; under reasonable assumptions the optimal choice is $\log\left[p(y_i \mid \underset{\sim}{y}_{[-i]}, \beta)\right]$ = log(** above)

③ Now we compute $LS_{CV}(M_j | y, B) =$

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[p(y_i | y_{[-i]}, M_j, B)\right]$$

In my research this is called the cross-validated (CV) log score

---

④ However, (my) research has shown the intuitively-reasonable fact that even with n only moderate in size

$$p(y_i^* | y_{[-i]}, M_j, B) \doteq p(y_i^* | y, M_j, B)$$

for all y^*

this motivates

$$LS_{FS}(M_j | y, B) = \frac{1}{n}\sum_{i=1}^{n} \log\left[p(y_i | y, M_j, B)\right]$$

full-sample (FS) predictive distribution for a new data point y^*

using MCMC to calculate $LS_{FS}(M_j | \underset{\sim}{y}, B)$

From above $p(y^+ | \underset{\sim}{y}, M_j, B) =$

$$\int_{\Theta_j} p(y^+ | \underset{\sim}{\theta}_j, M_j, B) \cdot \underbrace{p(\underset{\sim}{\theta}_j | \underset{\sim}{y}, M_j, B)}_{\text{PDF}} \, d\underset{\sim}{\theta}_j$$

This is a mixture representation of $p(y^* | \underset{\sim}{y}, M_j, B)$, so (i+) can be approximated to arbitrary MC accuracy as follows:

(A) obtain the MCMC data set for $\underset{\sim}{\theta}_j$ as usual, with M (large) monitoring iterations (B) monitor the deterministic quantity $p(y^* | \underset{\sim}{\theta}_j^{*(m)}, M_j, B)$ (C) Take the mean of the quantities in (B):

$$p(y^* \mid \underset{\sim}{y}, M_j, B) \doteq \frac{1}{M} \sum_{m=1}^{M} p(y^* \mid (\underset{\sim}{\theta_j^*})_m, M_j, B)$$

and now

it's as easy as this: ~~a good model~~ ~~large~~ LS values

$$LS_{PS}(M_j \mid \underset{\sim}{y}, B) = \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{1}{M} \sum_{m=1}^{M} p(y_i^* \mid (\theta_j^*)_m, M_j, B) \right]$$

(Rcode) CRAN seems not to offer much support for log scores, but

(a) LS_{CV} is also called the conditional predictive ordinate (CPO) method &

(b) there are some packages that implement CPO

The LS_{CV} idea goes back to Geisser & Eddy (1979) & Gelfand (1996).

q: where are (GoF) & (model complexity) in LS_{FS}?

A:

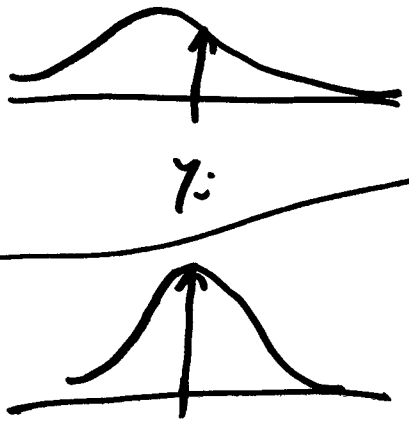
needless spread out (worse LS)

M_2 has extra parameters in it that are not helpful in prediction of future data

Note:

LS_{FS} works with both fixed effects & random effects models

$p(y_i | x, \underline{\underline{M}}, \beta)$



$p(y_i | x, M_2, \beta)$

has extra parameters that (greatly) improve the model fit

M_1 fit to as given data:

random effects, $(\theta_i | \mu, \sigma, \beta) \overset{[SM:N]}{\overset{IID}{\sim}} N(\mu, \sigma^2)$

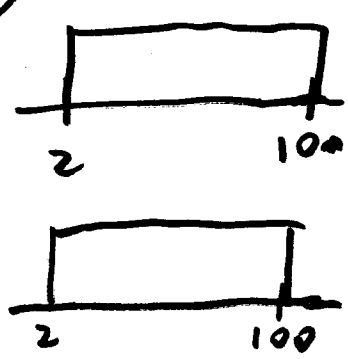
$k=6$

M_2 fit —

$(\theta_i | \mu, \sigma, r) \overset{(SM:\pi)}{\overset{IID}{\sim}} t_r(\mu, \sigma^2)$

$(i=1,\dots,k)$

i.e., r cannot be learned with $k = \square$ this dataset



$p(r)$ ‖ same $p(r|D\dots)$

as $r \to \infty$ $t_r \to N$

Comparison of Bayesian model comparison methods

BIC, other Bayes factors	DIC, AIC, LS
designed with what goal in mind?	
★★★★ achieve frequentist consistency at both M_1 & M_2	favor models that make good predictions OOS (out of sample)
(more complicated) complex	★★★ achieve frequentist consistency at M_2 but not at M_1
BIC, other Bayes factors tend to favor less complex models than {DIC, AIC, LS}	models that predict OOS well tend to be (somewhat) more complex than models that achieve ★★★★

frequentist consistency of Bayesian model comparison methods ↔ wanting to be a well-calibrated Bayesian

(more complex)

assumptions | ① either $M_1 = M_{DG}$ or $M_2 = M_{DG}$

② $n \to \infty$ holding M_1 & M_2 fixed

$$P_F(\text{BIC chooses } M_1 \mid M_1 = M_{DG}) \to 1 \quad (n \to \infty)$$

Bayes factors

if so, BIC / are consistent at the less complex model : true

$$P_F(\text{BIC chooses } M_2 \mid M_2 = M_{DG}) \to 1 \quad (n \to \infty)$$

if so BIC / Bayes factors are consistent at the more complex model : true

However, (LS, DIG, AIC)

$$P_F(\text{LS chooses } M_2 \mid M_2 \doteq M_{DG}) \xrightarrow[n\to\infty]{} 1$$

true

But

$$P_F(\text{LS chooses } M_1 \mid M_1 \doteq M_{DG}) \xrightarrow[n\to\infty]{} 0 \quad (!)$$

However,

(A) My $n = 79$; how do these asymptotic results apply, if at all?

A(A): not always

(B): (B) "$n\to\infty$ holding k_1 & $k_2 > k_1$ fixed" is not realistic for actual data sets: as $n\uparrow$, to decently model real-world complexity need $k_1\uparrow$ & $k_2\uparrow$

better asymptotics: $k_0, n\to\infty$ with $0 < \frac{k}{n} \to G < 1$