Bayesian model selection

Bayesian statistics 10 – choosing between models

Frédéric Barraquand (CNRS, IMB)

18/01/2022

What you probably know

Classical model selection often uses AIC and BIC.

- $\text{AIC} = -2\ln(\mathcal{L}(\theta, y)) + 2p$
- $\text{BIC} = -2\ln(\mathcal{L}(\theta, y)) + p\ln(n)$ where n is sample size.

Here the likelihood is formally $\mathcal{L}(\theta, y) = \mathbb{P}(Y_1 = y_1, ..., Y_n = y_n | \Theta = \theta)$ or $f(Y_1 = y_1, ..., Y_n = y_n | \Theta = \theta)$ for continuous Y.

Choosing best model out of 3 (for instance): argmin (IC_1, IC_2,IC_3)

What you may or may not know

AIC and BIC correspond to two different statistical philosophies (cf. Aho et al. 2014)

AIC	BIC
Asymptotically *efficient* $\approx$ best predicting model when $n \to \infty$	Asymptotically *consistent*: picks *true* model with prob. 1 when $n \to \infty$
Equivalent to LOOCV when $n \to \infty$	Equivalent to Bayes factor $n \to \infty$
Good for *prediction*	Good for *explanation*/*confirmation*
Invented by Hirotugu Aikaike	Gideon Schwartz

In practice these two philosophies often yield the same answers, though not always. There are intermediates such as the Hannan-Quinn Criterion $= -2\ln(\mathcal{L}(\theta, y)) + p\ln(\ln(n))$. A non-technical overview in Aho et al. 2017

Neither AIC nor BIC are Bayesian

These and their differences $(\Delta AIC, \Delta BIC)$ are

- based on the log-likelihoods (or their differences)
- do not depend on priors

In fact Bayesian model selection has classically relied on another tool, the Bayes factor. Initially viewed as a Bayesian alternative to significance testing by Sir Harold Jeffreys in 1935. (the p-value has been around for much much longer but had been popularized by Fisher in 1925).

What is the Bayes factor?

Let's say we have two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ corresponding to two different hypotheses about the data (e.g., humidity affects plant growth in 1 or it doesn't markedly in 2). Following Kass & Raftery (1995)'s derivation. Recall Bayes theorem applied to $Y = (Y_1, ..., Y_n)$ and $\mathcal{M}_1$.

$$\mathbb{P}(\mathcal{M}_1|Y = y) = \frac{\mathbb{P}(Y=y|\mathcal{M}_1)\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(Y=y|\mathcal{M}_1)\mathbb{P}(\mathcal{M}_1)+\mathbb{P}(Y=y|\mathcal{M}_2)\mathbb{P}(\mathcal{M}_2)}$$

The denominator is $\mathbb{P}(Y = y)$ but it will quite conveniently disappear when we divide by

$$\mathbb{P}(\mathcal{M}_2|Y = y) = \frac{\mathbb{P}(Y=y|\mathcal{M}_2)\mathbb{P}(\mathcal{M}_2)}{\mathbb{P}(Y=y|\mathcal{M}_1)\mathbb{P}(\mathcal{M}_1)+\mathbb{P}(Y=y|\mathcal{M}_2)\mathbb{P}(\mathcal{M}_2)}$$

and finally

$$\underbrace{\frac{\mathbb{P}(\mathcal{M}_1|Y = y)}{\mathbb{P}(\mathcal{M}_2|Y = y)}}_{\text{posterior odds}} = \underbrace{\frac{\mathbb{P}(Y = y|\mathcal{M}_1)}{\mathbb{P}(Y = y|\mathcal{M}_2)}}_{\text{BF}_{12}} \underbrace{\frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_2)}}_{\text{prior odds}}$$

The Bayes factor is not as simple as one thinks: priors

But wait, what is $\mathbb{P}(Y = y|\mathcal{M}_i)$ or if Y is continuous, $m(y|\mathcal{M}_i)$? Remember the first course. We defined the marginal density

$$m(y) = \int \underbrace{f(y|\theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}} d\theta$$

in Bayes theorem relating prior $\pi(\theta)$ to posterior $p(\theta|y)$ *for a given model*:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)}$$

Here we have several models, so one marginal density per model

$$m(y|\mathcal{M}_i) = \int \underbrace{f(y|\theta, \mathcal{M}_i)}_{\text{likelihood}} \times \underbrace{\pi(\theta|\mathcal{M}_i)}_{\text{prior}} d\theta$$

How to compute a Bayes factor in practice?

- We actually need to compute the integral including the prior over all possible parameter values, which we usually avoid by using MCMC
- Various techniques to do this, easy for simple models (ANOVA, LM), more complex exercise for complex hierarchical models (esp. non-nested models)

Package `BayesFactor` for simple models. Vignette

In `JAGS`, product space method: ''supermodel" encompassing the models under consideration. See also Tenan et al. (2014).

In R again, `bridgesampling` helps to compute BFs for JAGS and Stan models. Vignette. And extended bridgesampling paper.

A worked example I

Linear regression with 4 predictors (x_1, x_2, x_3, x_4)

```
x0 = 1:100/100
x1 = rnorm(100,0,1) + x0
x2 = rnorm(100,0,1) + x0
cor(x1,x2)
```

```
## [1] 0.1034979
```

```
x3 = rnorm(100,0,1) + x0*0.2 -0.6*x0^2 + 0.3*x0^3
x4 = rnorm(100,0,1) + x0*0.4 -0.6*x0^2
```

True model has 3 predictors in (x_1, x_2, x_3).

```
## Simulating the model
y = x1 + 0.5 * x2 - 1 * x3 + rnorm(100,0,0.05)
## Fitting the model with all covariates
summary(lm(y ~ x1 + x2 + x3 + x4))
```

A worked example II

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.130519 -0.034307 -0.008727  0.035426  0.114507
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -0.003316   0.006224   -0.533    0.595
## x1           1.004423   0.004902  204.881   <2e-16 ***
## x2           0.491438   0.004966   98.965   <2e-16 ***
## x3          -1.008429   0.004904 -205.651   <2e-16 ***
## x4          -0.004396   0.006418   -0.685    0.495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05291 on 95 degrees of freedom
```

A worked example III

```
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9991
## F-statistic: 2.85e+04 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
datareg = data.frame(y,x1,x2,x3,x4)
BF  = regressionBF(y ~ .,data=datareg,whichModels = 'top')
BF
```

```
## Bayes factor top-down analysis
## --------------
## When effect is omitted from x1 + x2 + x3 + x4 , BF is...
## [1] Omit x4 : 141.0018      ±0%
## [2] Omit x3 : 5.608446e-124 ±0%
## [3] Omit x2 : 3.108824e-94  ±0%
## [4] Omit x1 : 7.943542e-124 ±0%
##
## Against denominator:
##   y ~ x1 + x2 + x3 + x4
## ---
## Bayes factor type: BFlinearModel, JZS
```

Asymptotics of the Bayes factor ($n \to \infty$) and connection to BIC

$$\text{BIC} = -2\ln(f(y|\theta, \mathcal{M}_i) + 2\ln(n)p \approx -2\ln(m(y|\mathcal{M}_i))$$

using a Laplace approximation. See e.g. Hobbs and Hooten (eq. 28)

Because of this, a large-sample approximation to the Bayes factor (which can be obtained by simpler max likelihood-based modelling) is

$$BF_{12} = \frac{m(y|\mathcal{M}_1)}{m(y|\mathcal{M}_2)} = \frac{\exp(-\text{BIC}_1/2)}{\exp(-\text{BIC}_2/2)} = \exp\left(-\frac{1}{2}(\text{BIC}_1 - \text{BIC}_2)\right).$$

What does JAGS output by default?

DIC = Deviance Information Criterion. By Spiegelhalter et al. (2002)

Deviance $= -2\ln(\mathcal{L}(\theta, y)) + C$. Deviance is related to the sum-of-squared-errors for linear Gaussian models. Here $D(\theta) = -2\ln(p(y|\theta)) + C$, Bayesian spin on the deviance definition.

DIC starts with the definition of the *effective number of parameters* $p_D = \overline{D(\theta)} - D(\bar{\theta})$ or $\frac{1}{2}\overline{\mathbb{V}(D(\theta))}$ (Gelman et al. BDA).

DIC $= p_D + \overline{D(\theta)} = D(\bar{\theta}) + 2p_D \rightarrow$ similar structure to AIC and BIC. Parameter value $\bar{\theta}$ is the posterior mean.

Another worked example

TD10: Using previous nonlinear models of session 8 – comparing Gompertz (sigmoid) and van Bertalanffy's (simply saturating) growth equations.

Actually a similar example to that of the recent review of Bayesian model selection (in Stan) for fisheries science by Doll & Jacquemin (2019). They include a fish growth example with real data.

Limits of DIC

- sometimes poorly estimating model complexity (p_D)
- inappropriate for mixture models (and generally latent variable models, such as the ones of session 9)
- lack of a direct connection with predictive ability.

See Gelman et al.'s Understanding predictive information criteria for Bayesian models

Other, predictive information criteria

Watanabe-Akaike information criterion (WAIC). In fact, Watanabe (2010) calls it "Widely Applicable Information Criterion" following similar modesty by Akaike.

$$\text{WAIC} = -2\sum_{i=1}^{n} \ln\left(\int f(y_i|\theta)p(\theta|y)d\theta\right) + 2p_D$$

Based on the log-posterior predictive distribution.

More info

- Further explanation in e.g. Hobbs and Hooten (2015)
- JAGS code from O Gimenez here, itself based on recent implementation by M. Plummer.
- R package `loo` for Stan.

Methods to select covariates

If you have very many covariates and combinations thereof, ICs are not always the smartest move (can be bad for many-models comparisons).

Classical penalized regression $\rightarrow$ LASSO. Many Bayesian variants induced by special priors. Variable selection methods include:

- Kuo & Mallick's version of spike-and-slab coefficient priors

$$y_i = \beta_0 + \beta_1\gamma_1 x_1 + \beta_2\gamma_2 x_2 + \beta_3\gamma_3 x_3 + \beta_4\gamma_4 x_4 + \epsilon_i$$

  with $\gamma_i \sim \text{Bernoulli}(q)$.
- Many other ways to specify priors on coefficients summed up in this (not too technical) literature
  - Tenan et al. (2014)
  - O'Hara & Sillanpää (2009)
- General advice on Bayesian (ecological) model selection in
  - Hobbs and Hooten (2015)

Concluding words of caution – no model selection magic

Beware of model selection based on **any** single criterion or technique.

Concluding words of caution – no model selection magic

Beware of model selection based on **any** single criterion or technique.

- Model selection heavily depends on the *quality* of your **model set**. Ideally contains both "good" and "bad" models. Two pitfalls are including only good or only bad models.

Concluding words of caution – no model selection magic

Beware of model selection based on **any** single criterion or technique.

- Model selection heavily depends on the *quality* of your **model set**. Ideally contains both "good" and "bad" models. Two pitfalls are including only good or only bad models.
- You need to check that your models fit. In the data-simulated world this is easy to know (you know the true model). In the real world, prior and posterior predictive checks help a lot. See the guide by Conn et al. (2018) for a recap and more.

Concluding words of caution – no model selection magic

Beware of model selection based on **any** single criterion or technique.

- Model selection heavily depends on the *quality* of your **model set**. Ideally contains both "good" and "bad" models. Two pitfalls are including only good or only bad models.
- You need to check that your models fit. In the data-simulated world this is easy to know (you know the true model). In the real world, prior and posterior predictive checks help a lot. See the guide by Conn et al. (2018) for a recap and more.
- Sometimes a single score based on a variant of least squares (AIC, BIC, DIC, WAIC,...) is not enough. Perhaps some observations are more important. E.g. in a network you want connector nodes to be identified more correctly than others. Temporal or spatial correlations have to be just right. Etc. Simulations under the models resolve some of these issues, but you might want to add your own criteria to optimize.