A Very Brief Overview of Model Comparison Approaches

Vera Bulaevskaya

Lawrence Livermore National Laboratory

June 17, 2016

Model comparison approaches we will discuss:

1. Akaike Information Criterion (AIC)
2. Deviance Information Criterion (DIC)
3. Bayesian Information Criterion (BIC)
4. Bayes Factors
5. Watanabe-Akaike Information Criterion (WAIC)
6. Mixture Models

This is NOT a comprehensive list of model comparison methods!

Akaike Information Criterion (AIC)

$$AIC = -2\ log\ L(\hat{\boldsymbol{\theta}}_{MLE}) + 2k,$$

where

$L(\boldsymbol{\theta})$ is the likelihood function of parameter vector $\boldsymbol{\theta}$

$\hat{\boldsymbol{\theta}}_{MLE}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$,

k is $\#$ of parameters, i.e., the dimension of $\boldsymbol{\theta}$.

Notes on AIC:

- Balances the model's quality of fit (analogous to χ^2 goodness of fit) and parsimony of the model
- Not a measure of whether the model is true, only a relative measure when comparing to AICs for alternative models
- Cannot make use of prior information

Deviance Information Criterion (DIC)

Can be thought of as a Bayesian version of AIC $=$ $-2\ log\ L(\hat{\boldsymbol{\theta}}_{MLE}) + 2k$, with two changes:

1. Replace $\hat{\boldsymbol{\theta}}_{MLE}$ with $\hat{\boldsymbol{\theta}}_{Bayes} = E(\hat{\boldsymbol{\theta}}|\boldsymbol{Y})$ (the posterior mean)
2. Replace the number of parameters k with a bias correction based on data:

$$p = 2\left(log\ L(\hat{\boldsymbol{\theta}}_{Bayes}) - E_{\boldsymbol{\theta}|\boldsymbol{Y}} log\ L(\boldsymbol{\theta})\right)$$

This results in

$$DIC = -2\ log\ L(\hat{\boldsymbol{\theta}}_{Bayes}) + 2p$$

Notes on DIC:

- More natural than AIC from Bayesian point of view
- Unlike AIC, does not require maximizing the likelihood and is easy to calculate using MCMC samples from the posterior
- Requires the mean to be a good summary of the posterior (not the case if the distribution is heavily skewed or bimodal)

Bayes Factors (BF)

Suppose there are two competing models M_1 and M_2, which apriori we believe to be equally likely. Then we can compute the posterior odds of M_1:

$$\frac{p(M_1|\boldsymbol{Y})}{p(M_2|\boldsymbol{Y})} = \frac{p(M_1)}{p(M_2)} \times \frac{p(\boldsymbol{Y}|M_1)}{p(\boldsymbol{Y}|M_2)} = \frac{p(\boldsymbol{Y}|M_1)}{p(\boldsymbol{Y}|M_2)} \stackrel{def}{=} BF,$$

that is, BF is just the ratio of the marginal density of the data under model 1 to that under model 2, and the more BF exceeds 1, the more evidence in favor of M_1 over M_2. Furthermore,

$$BF = \frac{p(\boldsymbol{Y}|M_1)}{p(\boldsymbol{Y}|M_2)} = \frac{\int p(\boldsymbol{Y}|\boldsymbol{\theta_1}, M_1)p(\boldsymbol{\theta_1}|M_1)d\boldsymbol{\theta_1}}{\int p(\boldsymbol{Y}|\boldsymbol{\theta_2}, M_2)p(\boldsymbol{\theta_2}|M_2)d\boldsymbol{\theta_2}}$$

Notes on BF:

- Works well when the set of candidate models is truly discrete
- Marginal distribution of the data under each model must be proper (otherwise, the ratio is not well defined)

Bayesian Information Criterion (BIC)

Computing Bayes factors requires computing the marginal density of the data under a given model:

$$p(\boldsymbol{Y}|M) = \int p(\boldsymbol{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

This integral can be difficult or intractable and is often approximated using the Laplace approximation (Taylor expansion around the MLE), which can be shown to be

$$\begin{aligned} log\ p(\boldsymbol{Y}|M) &\approx log\ p(\boldsymbol{Y}|M, \hat{\boldsymbol{\theta}}_{MLE}) - k/2\ log\ N \\ &= log\ L(\hat{\boldsymbol{\theta}}_{MLE}) - k/2\ log\ N \end{aligned}$$

(where N = sample size). Multiplying the result by -2 gives rise to BIC:

$$BIC = -2\ log\ L(\hat{\boldsymbol{\theta}}_{MLE}) + k\ log\ N$$

BIC cont'd

Notes on BIC:

- Model with minimum BIC is the model with the largest approximate marginal density
- Very similar to AIC, but places much higher penalty on complexity than AIC, particularly for large N
- As $N \to \infty$, the probability that BIC will pick the correct model approaches 1 (asymptotic consistency), unlike AIC, which will favor models that are too complex
- For smaller N, however, it often picks models that are too simple (due to heavy penalty on complexity)
- If all models are assumed to be equally likely, $p(M_i|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|M_i) \approx e^{-\frac{1}{2}BIC_i}$, we can use BIC to estimate $p(M_i|\boldsymbol{Y})$:

$$p(M_i|\boldsymbol{Y}) \approx \frac{e^{-\frac{1}{2}BIC_i}}{\sum_{j=1}^{m} e^{-\frac{1}{2}BIC_j}}$$

Watanabe-Akaike Information Criterion (WAIC)

- Gelman et al. 2013 advocate using cross-validation rather than the criteria discussed above
- However, cross-validation can be computationally expensive
- WAIC is a cheaper approximation to cross-validation

Mixture Models

Given two candidate models M_1 and M_2 with parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$, respectively, model the data as their mixture:

$$\boldsymbol{Y} \sim \delta \cdot p(\boldsymbol{Y}|M_1, \boldsymbol{\theta_1}) + (1 - \delta) \cdot p(\boldsymbol{Y}|M_2, \boldsymbol{\theta_2})$$

with $0 \leq \delta \leq 1$.

The value of δ is the probability that the data are generated according to model M_1.

Thus, the objective here is to obtain the posterior distribution of δ via MCMC (or can use the EM algorithm if want to get its MLE).

See, for example, Kamary et al. (2014).

Bayesian Model Averaging

Rather than select one model, it may be appropriate to compute the weighted average of several models' predictions with weights equal to individual model's posterior probability.

Can be computationally difficult, but several approaches to simplify the computation have been proposed over the years.

References[1]

AIC, BIC, DIC, Bayes factors:

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin. *Bayesian Data Analysis* (chapter 7). Chapman & Hall, 2013.

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning* (chapter 7). Springer, 2009.

R. E. Kass and A.E. Raftery. Bayes Factors. *Journal of American Statistical Association*, 90: 773-795, 1995.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 440, 92-107, 2000.

[1] This is by no means an exhaustive list, but is rather a good starting point on each of the topics.

References cont'd

Model averaging:

J. Hoeting, D. Madigan, A. E. Raftery and C. T. Volinsky. Bayesian model averaging: A Tutorial. *Statistical Science*, 14, 382-417, 1999.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 440, 92-107, 2000.

Mixture models:

K. Kamary, K. Mengersen, C. P. Robert and J. Rousseau. Testing hypotheses via a mixture estimation model. arXiv:1412.2044v2, 2014.