Model selection and multimodel inference

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

29 April 2020

Goals for today

- Understand how to evaluate models via AIC and BIC
- Understand model likelihoods and evidence ratios
- Understand how model averaging can address model uncertainty
- Understand the differences between in-sample and out-of-sample methods

Model selection

There are 2 general approaches to model selection:

1. In-sample
2. Out-of-sample

In-sample model selection

There are 2 general ways to do in-sample selection:

1. Null hypothesis testing (F-test, likelihood ratio test)
2. Regularization (AIC, BIC)

K-L divergence

Recall the relationship between K-L divergence and likelihood

$$
\begin{aligned}
D_{KL} &= \mathrm{E}\left[\log\left(\frac{\mathcal{L}(y;\theta)}{\mathcal{L}(y;\Theta)}\right)\right] \\
&= \mathrm{E}\left(\log \mathcal{L}(y;\theta) - \log \mathcal{L}(y;\Theta)\right) \\
&= \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\theta)\right)}_{\text{entropy}} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}} \\
&= \text{constant} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}}
\end{aligned}
$$

Akaike's information criterion

Akaike's information criterion (AIC) for a given model is

$$D_{KL} = \text{constant} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}}$$

$$\Downarrow$$

$$AIC \approx 2D_{KL}$$

$$\Downarrow$$

$$AIC = 2k - 2\log \mathcal{L}(y;\theta)$$

where k is the number of parameters in the model

Bayesian information criterion

Not long after Akaike developed his information criterion, Gideon Schwarz derived the *Bayesian information criterion* (BIC)

Bayesian information criterion

BIC also has a relationship to K-L divergence

$$D_{KL} = \text{constant} - \underbrace{\mathrm{E}\left(\log \mathcal{L}(y;\Theta)\right)}_{\text{log likelihood}}$$

$$\Downarrow$$

$$BIC \approx 2D_{KL}$$

$$\Downarrow$$

$$BIC = k \log n - 2 \log \mathcal{L}(y;\theta)$$

where k is the number of parameters & n is the sample size

Biases in AIC and BIC

AIC tends to select more complex models, regardless of n

BIC tends to select more simple models, regardless of n

Thus, some people use both in model selection

AIC and BIC for Gaussian distributions

If our model is based on a Gaussian (normal) distribution, we can replace the likelihood term in AIC or BIC

$$\begin{aligned} IC &= \text{constant} - 2\log \mathcal{L}(y;\theta) \\ &= \text{constant} + n\log\hat{\sigma}^2 \end{aligned}$$

where $\hat{\sigma}^2 = \frac{SSE}{n}$ (a *biased* variance estimator)

AIC and BIC for Gaussian distributions

More specifically, we have

$$\begin{aligned} AIC &= 2k - 2\log \mathcal{L}(y;\theta) \\ &= 2k + n\log \hat{\sigma}^2 \end{aligned}$$

$$\begin{aligned} BIC &= k\log n - 2\log \mathcal{L}(y;\theta) \\ &= k\log n + n\log \hat{\sigma}^2 \end{aligned}$$

Differences in AIC or BIC

When using AIC or BIC for model selection, it's easier to think about differences among models

For the i^{th} model in a set of models

$$\Delta IC_i = IC_i - \min IC$$

$\min IC$ is the minimum IC among all of the models in the set

The "best" model will have ΔIC = 0

Differences in AIC or BIC

Given a set of models and their ΔIC values, what is the strength of evidence against a model with a higher ΔIC?

Differences in AIC

Some guidelines for ΔAIC from Burnham & Anderson (2002)

ΔAIC_i	Interpretation
0 - 2	essentially none
4 - 7	considerably less
> 10	substantial

Differences in BIC

Some guidelines for ΔBIC from Kass & Raftery (1995)

ΔBIC_i	Interpretation
0 - 2	not worth more than a bare mention
2 - 6	positive
6 - 10	strong
> 10	very strong

A benchmark for ΔAIC

Why do we use a lower cutoff for ΔAIC = 2?

Consider 2 models (A & B) that have the same AIC, and are otherwise identical other than B having 1 more parameter

$$A : y = \alpha + e$$
$$B : y = \alpha + \mu + e$$

A benchmark for ΔAIC

Why do we use a lower cutoff for ΔAIC = 2?

We can decompose their AIC's and compare the difference in log-likelihoods

$$AIC_B = AIC_A$$
$$2(k+1) - \log \mathcal{L}_B = 2k - \log \mathcal{L}_A$$
$$2k + 2 - \log \mathcal{L}_B = 2k - \log \mathcal{L}_A$$
$$\log \mathcal{L}_B = \log \mathcal{L}_A + 2$$

Thus, adding 1 parameter to an otherwise identical model should increase its log-likelihood by at least 2 units

Relative rankings

Using information criteria to rank our models works well, but the “best” model may not be any good in practice

Thus, we also need to consider other measures of goodness-of-fit, predictive ability (eg, R^2, $MSPE$)

QUESTIONS?

Uncertainty in our analysis

We have been focused on 2 types of uncertainty (variance):

1. parameter (how good are our estimates of $\hat{\beta}$)
2. sampling (how noisy are the data; how big is σ^2)

Model uncertainty

There is a 3^{rd} form of uncertainty that is also important

1. parameter (how good are our estimates of $\hat{\beta}$)
2. sampling (how noisy are the data; how big is σ^2)
3. model (how do we know this is the correct model)

Model likelihood

Let's consider a way to assign a weighting to a given model

Recall that we defined the likelihood of some parameters given some data to be

$$\mathcal{L}(\theta \mid y) = \mathcal{L}(y; \theta)$$

Model likelihood

Let's consider a way to assign a weighting to a given model

Recall that we defined the likelihood of some parameters given some data to be

$$\mathcal{L}(\theta \mid y) = \mathcal{L}(y; \theta)$$

We can similarly define the likelihood of a model f given the data as

$$\mathcal{L}(f \mid y) = \mathcal{L}(y; f)$$

Model likelihood

More formally, given $\Delta_i = AIC_i - \min AIC$

$$\mathcal{L}(y; f_i) \propto \exp\left(-\frac{1}{2}\Delta_i\right)$$

Akaike weights

Because the model likelihoods are all relative (just as with other likelihoods), we can create a set of normalized *Akaike weights* that sum to 1

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{s=1}^{S} \exp\left(-\frac{1}{2}\Delta_i\right)}$$

- w_i is the weight of evidence in favor of model i being the "best" model *in the set*

Evidence ratios

We can compute the evidence in favor of model j relative to model i as the ratio of their likelihoods

$$ER_{ij} = \frac{\mathcal{L}(y; f_i)}{\mathcal{L}(y; f_j)}$$

Evidence ratios

Given a set of Akaike weights, this evidence is also given by the ratio of model weights

$$ER_{ij} = \frac{\mathcal{L}(y; f_i)}{\mathcal{L}(y; f_j)} = \frac{w_i}{w_j}$$

Evidence ratios

Most often, we are interested in the ER_{ij} between the best model and others in the set

$$\begin{aligned} ER_{1j} &= \frac{\exp\left(-\frac{1}{2}0\right)}{\exp\left(-\frac{1}{2}\Delta_j\right)} \\ &= \frac{1}{\exp\left(-\frac{1}{2}\Delta_j\right)} \\ &= \exp\left(\frac{1}{2}\Delta_j\right) \end{aligned}$$

Evidence ratios

The relationship between ΔAIC and the evidence ratio is exponential

$$\begin{aligned} \Delta &\rightarrow ER \\ 2 &\rightarrow 2.7 \\ 4 &\rightarrow 7.4 \\ 8 &\rightarrow 55 \\ 16 &\rightarrow 2981 \end{aligned}$$

Ambivalence

Some people fret when they cannot definitively select a best model within their set

This is not a defect of the information criterion, but rather that the data are *ambivalent* concerning model structures

Data must be fixed

Note: when using information criteria, the data *must be fixed* across all models

That is, our inference is conditional on the data in hand

Multimodel inference

Given uncertainty in which model is the “best”, we can use *multimodel inference* to average our predictions over all models in a set

This is equivalent to the National Weather Service’s *ensemble predictions*

Model averaging

We can use Akaike weights to average the parameters or predictions from several models

For a given parameter θ, it's model averaged estimate is

$$\bar{\hat{\theta}} = \sum_{i=1}^{S} w_i \hat{\theta}_i$$

and S is the total number of models in the set

Model averaging

If a given parameter θ does not appear in all models, we can use an indicator function to compute the average estimate

$$\bar{\hat{\theta}} = \frac{\sum_{i=1}^{S} I(f_i) w_i \hat{\theta}_i}{\sum_{i=1}^{S} I(f_i) w_i}$$

$$I(f_i) = \begin{cases} 1 & \text{if } \theta \text{ is in } f_i \\ 0 & \text{otherwise} \end{cases}$$

Model selection

Selecting among possible models begins with a *reasonable* set based on first principles

This set of models

- may represent different hypotheses about the process of interest
- may represent different combinations of predictors
- should be finite

Model selection

Selecting among possible models begins with a *reasonable* set based on first principles

This set of models

- may represent different hypotheses about the process of interest
- may represent different combinations of predictors
- sound reasonable to others as well

Model selection

- Be **very** wary of “data mining” (“everything but the kitchen sink”)
- Use your knowledge of the system to choose predictors judiciously
- Pay attention to possible collinearity among predictors

QUESTIONS?

Out-of-sample model selection

We have seen several in-sample approaches to model selection

Let’s check out some options for out-of-sample selection

Model validation

A common Q is, “How well does a model predict new data?”

To answer this, we can use $n - q$ data points to fit the model and reserve q data points for model validation

There are several measures for evaluating out-of-sample predictions

Scale-dependent measures

Their scale depends on the units of the data

Should not be used when comparing across different data sets

Scale-dependent measures

Mean squared prediction error (MSPE)

Mean squared prediction error (MSPE) is perhaps the most common

$$MSPE = \frac{\sum_{i=1}^{q}(y_i - \hat{y}_i)^2}{q}$$

Scale-dependent measures

Root mean squared prediction error (RMSPE)

RMSPE is the square root of MSPE

$$RMSPE = \sqrt{MSPE}$$

It has the advantage of being on the same scale as the data

Scale-dependent measures

Mean absolute error (MAE)

$$MAE = \frac{\sum_{i=1}^{q} |y_i - \hat{y}_i|}{q}$$

Percentage-based measures

They have the advantage of being scale-independent

They can be used to compare models across different data sets

However, they are extremely skewed when *any* $y_i \approx 0$

Percentage-based measures

Mean absolute percentage error (MAPE)

$$MAPE = \frac{\sum_{i=1}^{q} |p_i|}{q}$$

$$p_i = 100 \frac{y_i - \hat{y}_i}{y_i}$$

Percentage-based measures

Root mean square percentage error (RMSPE)

$$RMSPE = \sqrt{\frac{\sum_{i=1}^{q} p_i^2}{q}}$$

$$p_i = 100\frac{y_i - \hat{y}_i}{y_i}$$

Model validation

There is a long, winding road littered with numerous critiques of these different methods

My advice is to consider “what happens if I’m wrong?” and consider the biases of the different methods

Cross-validation

Another form of out-of-sample selection is *cross-validation*

There are 2 types:

1. Exhaustive

Exhaustive methods use *all possible combinations* of fitting and testing data

Cross-validation

Another form of out-of-sample model selection is *cross-validation*

There are 2 types:

1. Exhaustive
2. Non-exhaustive

Non-exhaustive methods do not use all possible combinations of fitting and testing data

Cross-validation

Another form of out-of-sample model selection is *cross-validation*

There are 2 types:

1. Exhaustive
2. Non-exhaustive

Both types rely on some form of model validation method like we just discussed

Cross-validation

Exhaustive

Leave-p-out cross-validation uses $n - p$ data points for fitting the model, and p points for evaluating the fit

If $p > 1$ and n even somewhat large, this can be prohibitively slow because there are $\binom{n}{k}$ combinations

For example, if $p = 3$ and $n = 20$ there are $\binom{20}{3} = 1140$ different permutations

Cross-validation

Exhaustive

Leave-one-out cross-validation uses $n - 1$ data points for fitting the model, and 1 point for evaluating the fit

This results in n models being fit

Cross-validation

Non-exhaustive

k-fold cross-validation is a hybrid approach where the data are randomly partitioned into k equal sized groups

One of the k sub-samples is retained for validation while the remaining $k - 1$ groups are used for fitting

This process is then repeated k times, with each of the k sub-samples used exactly once for validation

The k results can then be averaged to produce a single estimate

Summary

Here we have seen the difference between in-sample and out-of-sample model selection

- using AIC and BIC for model selection
- using model weights in evidence ratios to compare one model to another
- using model averaging to address model uncertainty
- using exhaustive and non-exhaustive cross-validation