NRES_798_15_201501

Model selection

“All models are wrong but some are useful” – George Box

- Models are abstractions
- Most biological systems too complex to be defined exactly by a model

- Is a model “good”, is a model “good enough”?
- Is model A better than model B?

Parsimonious Models

- A model is parsimonious if it has the following desirable properties:

  1. Fits the data well (i.e. can explain the data).
  2. Has few parameters, so the at the explanation of the data is not overly complicated.
  3. There is confidence in the estimates of the parameters, and hence, confidence in the model's explanation of the data.

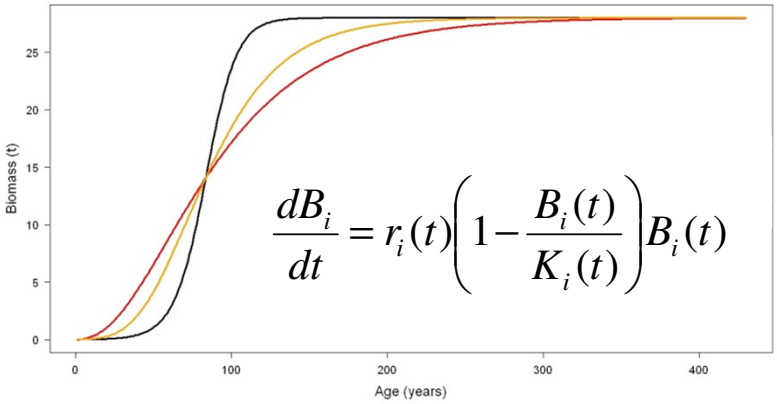
Identifying parsimonious models is referred to as model selection

Ecological (statistical) models

- Models aim to capture some component of the real world

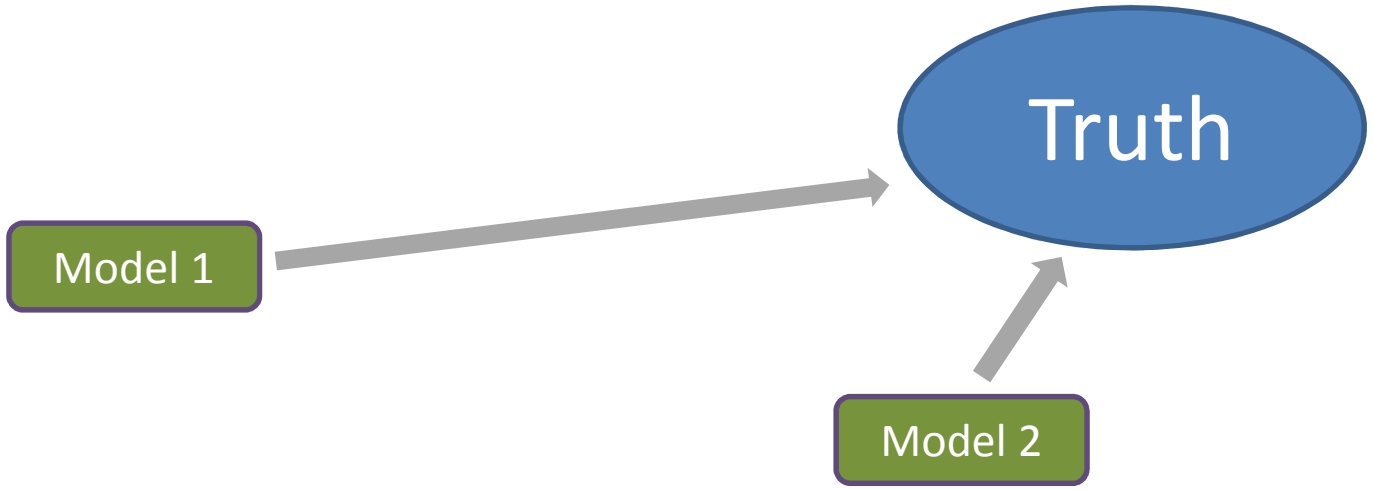
Truth
(complex)

Model
(simple)



Ideal model is the
"true" model

Ecological (statistical) models



- Complex models can approximate the truth well or poorly depending on their **parameter** values

- We need a measure of “distance” from the truth

Distance from truth: discrete system

- True model *f*
- True probability of observing event *i* (of *K* possible events) is p_i

- Test model *g*
- Test model probability of observing event *i* is π_i
- Information lost when assuming model *g* instead of *f* is given by:

$$I(f,g) = \sum_{i=1}^{K} p_i \ln\left(\frac{p_i}{\pi_i}\right)$$

- Kullback-Leibler information (distance)
- Kullback 1959: discrepancy between two distributions

Kullback-Leibler distance (information)

$$I(f,g) = \sum_{i=1}^{K} p_i \ln\left(\frac{p_i}{\pi_i}\right)$$

- If our model was the truth (i.e. $\pi_i = p_i$ for all i) then the distance $I = 0$
- I can be interpreted as the “information” lost when the model M is used to approximate the truth T
- K-L is one possible definition of “distance”, there are numerous others that could have been chosen

Continuous K-L distance

- A proposed model M predicts the **probability density** of seeing outcome *x* as *g(x)*
  - i.e. *f(x)* and *g(x)* are probability density functions

$$\boldsymbol{I(f,g) = \int_{x} f(x) ln\left(\frac{f(x)}{g(x)}\right) dx}$$

Burnham and Anderson (2002) pages 50-64

K-L distance

$$I(f,g) = \sum_{i=1}^{K} p_i \ln\left(\frac{p_i}{\pi_i}\right)$$

- To calculate *I* we need to know the **true** set of probabilities p_i
  - In practice we don't know them!
- We also need to know the parameter values of our model so we can calculate π_i
- If we wish to use model *g* to approximate the truth (*f*) and make predictions, then the best parameters of our model are those that minimize *I*
  - i.e. the structure of the model (*g*) must be the same as the true model (*f*), and the parameterization of *g* must be correct (minimize *I*)

K-L distance

- In reality, we do not know the truth, and therefore cannot use the Kullback-Leibler distance to find the best parameters for a model.

- Instead model parameters must be estimated from our data (i.e. model fit to the data)

- We can **estimate** K-L distance using maximum log-likelihoods
  - (maximum likelihood parameter estimation (MLE))

Maximum Likelihood Estimation (MLE)

- Likelihood: x data outcomes, θ parameters

$$L(\theta|x) = P(x|\theta)$$

- Likelihood of θ given x

- Maximizing the likelihood value therefore gives the parameter values that allow the mode to best approximate the data

Estimating K-L distance

- Akaike (1973) proposed K-L distance as a basis for model selection
  - For a given model there exists a set of parameter values that minimizes the K-L distance
  - We do not know this set of values, instead, we estimate them using maximum likelihood estimation (MLE)
  - Akaike found that MLEs bias the approximation of K-L distance
  - The more parameters the stronger the bias
    - We should therefore correct for this bias

AIC

- A relative estimation of K-L distance for a give model, referred to as the Akaike Information Criterion (AIC) is:

$$AIC = -2LL(\hat{p}|d, model) + 2K$$

Where:

$\hat{p} = \{\widehat{p_1}, \dots, \widehat{p_K}\}$ is the set of K model parameters estimated using MLE

$d = \{d_1, \dots, d_N\}$ is the set of N data

K is the number of parameters

$\mathrm{LL}(\hat{\theta})$is the maximum log-likelihood of the model given the data

AIC

$$AIC = -2LL(\hat{p}|d, model) + 2K$$

- AIC should be interpreted as:
  - Twice the expected K-L distance, minus some unknown **constant** (C, model specific), if the experiment or data collection were repeated many times.

- It is important to remember that AIC is an **estimate** of the K-L distance, and as such the model with the lowest AIC may not actually have the lowest K-L distance

- However, the model with the smallest AIC value is more likely to be the most parsimonious model from the set of candidate models based on K-L distance

AIC_c

- AIC works well when there is:
  - Lots of data (i.e. there are a large number of potential outcomes)
  - The proposed data can potentially fit the data well

- When there is not a lot of data (i.e. N/K is small) a "corrected" AIC works better (AIC_c)
  - The rule of thumb is use AIC_c if the number of independent data points (N) per parameter in the model is less than 40.

$$AIC_c = -2LL(\hat{p}|d, model) + 2K + \frac{2K(K+1)}{N-K-1}$$

Applying AIC and AIC_c

- AIC estimates K-L distance for each model but the estimate includes an unknown constant (C) (which depends on the model and the "truth" which is unknown.
- Because of this, the absolute magnitude of the AIC values are meaningless!

- Differences in AIC values among models are important

$$\Delta(M_m) = AIC_c(M_m) - \min_i[AIC_c(M_i)]$$

- Where min[…] refers to the minimum AIC value of all models considered

AIC Δ-values

Model	AIC	Δ	Rank
m1	277.58	3.4	3
m2	288.38	14.2	6
m3	274.18	0	1
m4	279.88	5.7	4
m5	275.93	1.75	2
m6	285.78	11.6	5

Applying AIC Δ-values

- Rules of thumb
  - Models with a Δ-value greater than 10 are very unlikely to be the model with the lowest K-L distance and can be disregarded.
  - Models with a Δ-value between 4 and 7 are less likely to be the model with the lowest K-L distance but should not e disregarded.
  - Models with Δ-values less than 2 are all likely to be the model with the lowest K-L distance.

- Important:
  - For the Δ-values to provide useful information at least one of the proposed models must describe the data reasonably well
  - i.e. AIC cannot choose a good model if all the candidate models are poor

AIC weights

- An alternative approach is to calculate AIC weights for each model

$$w(M_m) = \frac{\exp(-\Delta(M_m)/2)}{\sum_{g=i}^{G} \exp(-\Delta(M_g)/2)}$$

  - Where G is the number of proposed models (G weights sum to 1)

  - Weight $w(M_m)$ can be interpreted as:
    - The "weight of evidence" that model M_m is the model, out of those proposed, having the lowest K-L distance.

    - NOT, the probability that the model has the lowest K-L distance

Other information criterion

- Bayesian information criterion (BIC)

$$\mathrm{B}IC = -2LL(\hat{p}|d, model) + K \cdot \ln(n)$$

  - More strongly penalized number of parameters

- Deviance information criterion (DIC)
  - Modification of AIC for hierarchical model comparison
  - Often used with Bayesian models

Aho, K.; Derryberry, D.; Peterson, T. (2014), "Model selection for ecologists: the worldviews of AIC and BIC", *Ecology* **95**: 631–636

AIC in R

- AIC(model, k=2)
- AIC(model1,model2,k=2)
  - K is the penalty per parameter to be used
    (k = 2 in classic AIC)

$$AIC = -2LL(\hat{p}|d, model) + 2K$$

  - model object must have logLink method (i.e. log-likelihood can be calculated)

- extractAIC(model,k=2)
  - Model objects from “lm”,”aov”,”glm”,”coxph”,”suvreg”

AIC in R

- Mod1 <- lm(bmass ~ pop + age + pop*age + height^2)
- Mod2 <- lm(bmass ~ pop + age + pop*age)

- AIC(Mod1, Mod2,k=2)

- Model selection
  - 4 terms, 1 interaction
  - 9 models