Model Selection

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

1. The Problem of Overfitting
2. Likelihood-Ratio Tests
3. Information Criteria
4. IC Considerations

The Problem of Overfitting

Adding Model Structure

We started by fitting simple linear regression models of the form:

$$y_i = \beta_0 + \beta x_i + \varepsilon_i$$

We then extended this to multiple linear regression of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$$

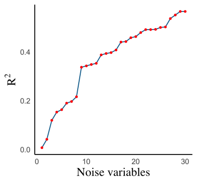
Last lecture we covered linear mixed effects model that add additional structure to account for correlations within groups:

$$\boldsymbol{y_i} = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

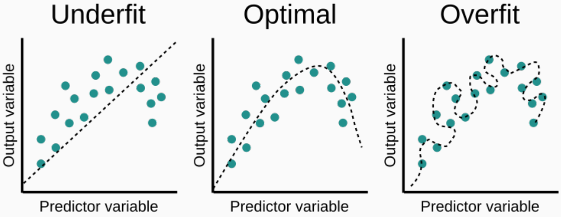
The Problem of Overfitting

Lecture by lecture we've been increasing the complexity of our models

But more complexity does not necessarily mean improved performance, or that the parameters are meaningful.



This puts us in a situation where we need to strike an optimal balance between having too many or too few features in our models.



Overfitting and prediction in action

Walters & Ludwig (1981) simulated fish population dynamics using complex age-structured model (different births and deaths for fish of different ages).

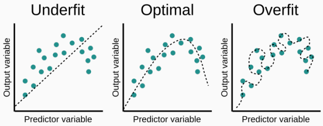
When data were realistically sparse and noisy a simple model without age structure resulted in the best predictions, even though they knew for a fact that there were age differences (!)

Why? Data contain a fixed amount of information (limited by resources, time, instrument precision, etc.) and as we estimate more and more parameters that information gets spread thinner and thinner.

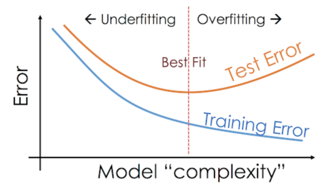
For Walters & Ludwig, spreading the information across age classes meant each age class' dynamics were included, but poorly estimated.

The Problem of Overfitting Cont.

An **underfit** model fails to accurately predict the data that were used to fit the model, and test datasets or future conditions.



An **overfit** model gives a very low prediction error on the dataset used to fit the model, but has a very high prediction error on test data.



This happens because you're fitting the noise not the signal.

Adding complexity

In practice, adding features/complexity to a model usually boils down to answering the question:

Does adding an extra parameter improve the fit sufficiently to justify the additional complexity?

... or seen another way:

Is there enough information in my data to support the additional complexity that comes with an extra parameter?

Note: Excluding a parameter does not necessarily mean an effect does not exist. It does mean we can't estimate it in a meaningful way from the information we have on hand.

Model selection

The goal of **modelling** is to identify the simplest model possible that captures all of the most important features of our data/system.

The goal of **model selection** is to know whether adding another parameter to our model not only improves the performance, but improves it by some specific amount (i.e., not just marginal gains) in order to minimise the risk of overfitting. If the more complex model doesn't pass this threshold, then it is rejected in favour of the simpler model.

Model selection highly controversial, but extremely important. We're going to cover two approaches:

1. Likelihood-ratio tests
2. Information criteria

Likelihood-Ratio Tests

Likelihood-Ratio Test

The likelihood-ratio test compares a pair of **nested** models based on the ratio of their likelihoods.

$$\lambda_{\mathsf{LR}} = -2\ln\left[\frac{\mathcal{L}(\text{Reduced model})}{\mathcal{L}(\text{Full model})}\right]$$

The likelihood-ratio test statistic is often expressed as a difference between the log-likelihoods

$$\lambda_{\mathsf{LR}} = -2(\ln[\mathcal{L}(\text{Reduced})] - \ln[\mathcal{L}(\text{Full})])$$

Nested Models

A ‘simple’ model is nested in a ‘complex’ model if the full model reduces to the simpler model when parameters are set to some fixed values (usually 0, 1, or ∞).

For example:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

if we set β_2 to 0, then:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

So the models are nested

$y = \frac{ax}{1+(\frac{a}{b}x)}$ (Beverton-Holt)

if we set $b \to \infty$, then:

$y = ax$ (Linear model)

Again the models are nested

Nested Models

A ‘simple’ model is nested in a ‘complex’ model if the full model reduces to the simpler model when parameters are set to some fixed values (usually 0, 1, or ∞).

For example:

$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$ is not nested in $y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$

because if we set β_1 to 0, then:

$y_i = \beta_0 + \varepsilon_i$

Both are simple linear models, and might even be based on the same data, but they are not nested.

Nested distributions

Both the deterministic and stochastic components of models can be nested.

For example:

The Poisson distribution is nested in the negative binomial distribution.

The binomial distribution is nested in the beta-binomial distribution

In other words, likelihood ratio tests can be used to identify both deterministic *and* stochastic components of your model (we'll cover non-gaussian models later, but keep this in mind).

The test statistic

So how does being able to quantify λ_{LR} help us identify the best model structure?

According to Wilks' theorem, as the sample size n approaches ∞, the test statistic λ_{LR} will be chi-squared distributed with degrees of freedom equal to difference in the number of parameters between the two models.

This implies that we can compare λ_{LR} to the χ^2 value corresponding to a desired statistical significance threshold (usually $\alpha = 0.05$) as an approximate statistical test.

Likelihood ratio test in R

```
library(nlme)

data <- read.csv("Ant_Richness.csv")

FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

FIT_Reduced  <- gls(num_sp ~ latitude, data = data, method = "ML")

lambda <- -2*(FIT_Reduced$logLik - FIT$logLik)

pchisq(lambda, df = 1, lower.tail=FALSE)

0.001948725

anova(FIT, FIT_Reduced)

            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
FIT             1  4 16.11769 20.48186 -4.058844
FIT_Reduced     2  3 23.71490 26.98803 -8.857452 1 vs 2 9.597217  0.0019
```

This is telling us that the extra parameter is resulting in a significant improvement to the model, so the extra complexity is worth the cost.

Limitations of LRTs

With only a small number of nested models to compare, likelihood ratio tests are quick easy, and statistically efficient.

As the complexity of the full model increases, the number of pairwise LRTs increases dramatically and soon becomes unwieldy.

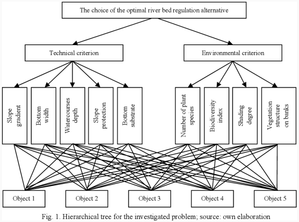


Fig. 1. Hierarchical tree for the investigated problem; source: own elaboration

(Hachoł *et al.*, 2017)

If some of the models aren't nested in one another, there is no way to compare them using LRTs.

Information Criteria

Information Criteria

Information criterion (IC) approaches (sometimes called ‘Information Theoretic’; IT approaches) can compare all models at once, avoiding the need to cary out multiple pairwise comparisons, and do not require nested models.

In practice, all IC methods reduce to finding the model that minimises some ‘criterion’ that is the sum of a term based on the likelihood and a penalty term

$$IC \approx \mathcal{L}(\text{model}) + \text{penalty term}$$

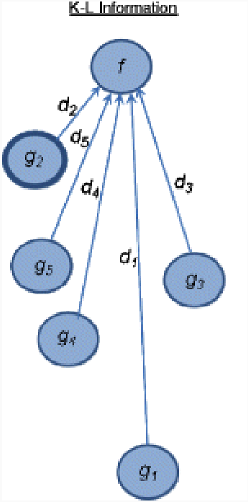
Kullback-Leibler Information

IC approaches are based on Kullback-Leibler information.

K-L information represents the information lost when model g_i is used to approximate full reality (f), or the distance between model g_i and reality.

The goal is then to select the model that minimises K-L information loss (i.e., the model that's closest to reality, f).

The problem is that K-L information loss cannot be computed or estimated.



(Burnham *et al.*, 2011)

Akaike Information Criterion

In the early 1970s, Hirotogu Akaike made a major breakthrough when he found a formal relationship between K-L information and maximum likelihood (Akaike, 1998).

He focused on the double expectation of the second term of the K-L information and found that, for large sample sizes, this can be estimated simply as $\ln(\mathcal{L}) - K$, where K is the total number of estimable parameters in the model.

Akaike multiplied both terms by -2 to get his now famous:

$$AIC = -2\ln(\mathcal{L}) + 2K$$

Note: The term $-2\ln(\mathcal{L})$ is well known among statisticians as the 'deviance', a goodness-of-fit statistic for a statistical model.

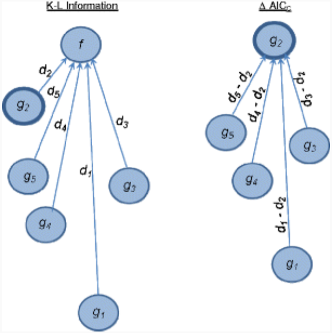
Akaike Information Criterion Cont.

K-L information represents the distance between model g_i and reality.

Because we can't estimate 'reality', we instead rely on AIC (or other IC).

We don't know how far the models are from the truth, but their relative positions should be the same, so we can now rank them amongst one another.

Low K-L information means closer to reality, so with AIC, the lower the value the better.



(Burnham *et al.*, 2011)

Small sample size bias correction

Akaike's derivations/approximations were based on large sample sizes.

These are not necessarily valid when sample sizes are small.

When the sample sizes are small, there is a good chance that AIC will overfit and select models that have too many parameters.

A small sample size bias correction for AIC was derived that increases the penalty term and is more frequently used in practice. This criterion is denoted as AICc to make it distinct from AIC, and is given by:

$$\mathrm{AICc} = \mathrm{AIC} + \tfrac{2k^2+2k}{n-k-1}$$

Question: What happens when $n \rightarrow \infty$?

AICc Model Selection

The actual AICc values are not particularly interesting in themselves. It's the differences, or the ΔAICc values, that are the key to ranking the models.

When picking among a set of potential models, we select the one with the lowest AICc value of the bunch.

The ΔAICc values for any given model are linked to the evidence ratio for the best model as $\exp[-\frac{1}{2}\Delta\text{AICc}]$.

For example: A model with a ΔAICc of 2 is ∼2.7 times less likely than the best fit model, and a model with a ΔAICc of 50 is 72 billion times less likely

ΔAICc	**Evidence Ratio**
2	2.7
4	7.4
6	20.1
8	54.6
9	90
10	148.4
11	244
12	403
13	665
14	1,097
15	1,808
20	22,026
50	72 billion

(Burnham *et al.*, 2011)

Relationship between ΔAIC and LRTs

The likelihood-ratio can be expressed as a difference between the log-likelihoods:

$$\begin{aligned}\lambda_{\mathsf{LR}} &= -2(\log \mathcal{L}(\theta_1) - \log \mathcal{L}(\theta_2)) \\ &= -2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_2)\end{aligned}$$

The ΔAIC between a pair of models is

$$\begin{aligned}\Delta\mathrm{AIC} &= \mathrm{AIC}_{\theta_1} - \mathrm{AIC}_{\theta_2} \\ &= (-2 \log \mathcal{L}(\theta_1) + 2K_1) - (-2 \log \mathcal{L}(\theta_2) + 2K_2) \\ &= -2 \log \mathcal{L}(\theta_1) + 2K_1 + 2 \log \mathcal{L}(\theta_2) - 2K_2 \\ &= -2 \log \mathcal{L}(\theta_1) + 2 \log \mathcal{L}(\theta_2) - 2(K_2 - K_1) \\ \Delta\mathrm{AIC} &= \lambda_{\mathsf{LR}} - 2(K_2 - K_1)\end{aligned}$$

If $K_1 = K_2$, then LRT = ΔAIC.

AICc model selection in R

```
library(nlme)

data <- read.csv("Ant_Richness.csv")

FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

k <- 4 #Intercept, 2 params, and the variance

-2*FIT$logLik + 2*(k)

[1]16.11769

AIC(FIT)

[1]16.11769

AIC(FIT) + (2*k^2 + 2*k)/(nrow(data) - k - 1)

[1] 18.47063

library(MuMIn)

 AICc(FIT)

[1] 18.47063
```

AICc model selection in R cont.

Let's compare all possible models

```
FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

FIT_el  <- gls(num_sp ~ elevation, data = data, method = "ML")

FIT_lat  <- gls(num_sp ~ latitude, data = data, method = "ML")

INTERCEPT <- gls(num_sp ~ 1, data = data, method = "ML")

AICc(FIT); AICc(FIT_el); AICc(FIT_lat); AICc(INTERCEPT)

[1] 18.47063
[1] 24.67657
[1] 25.04824
[1] 31.07926
```

This favours the full model. But by how much?

```
AICc(FIT_el) - AICc(FIT)
[1] 6.205939

1/exp(-(1/2)*6.205939)
[1] 22.26397
```

```
AICc(INTERCEPT) - AICc(FIT)
[1] 12.60863

1/exp(-(1/2)*12.60863)
[1] 546.9268
```

In a paper we would report that the next best model had a ΔAICc of ∼6.2, or was ∼22 times less likely, and the intercept only model had a ΔAICc of ∼12.6 and was >546 less likely than the full model.

AICc model selection in R cont.^{2}

This process is automated by the `dredge()` function in the `MuMIn` package.

```
FIT <- gls(num_sp ~ latitude + elevation, data = data, method = "ML")

dredge(FIT)

Global model call: gls(model = num_sp ~ latitude + elevation, data = data, method = "ML")
---
Model selection table
  (Intrc)     elvtn   lattd df  logLik AICc delta weight
4  11.120 -0.001373 -0.2018  4  -4.059 18.5  0.00  0.922
2   2.489 -0.001613          3  -8.672 24.7  6.21  0.041
3  12.390           -0.2388  3  -8.857 25.0  6.58  0.034
1   2.114                    2 -13.224 31.1 12.61  0.002
Models ranked by AICc(x)
```

Other IC

Since the advent of AIC, a number of other IC methods have been developed. In essence they all reduce to the combination of some term based on the likelihood and a penalty term based on model complexity.

$$IC \approx \mathcal{L}(\mathrm{model}) + \mathrm{penalty\ term}$$

The Bayesian information criterion (BIC) BIC was developed by Gideon Schwarz as an alternative to AIC with a larger penalty term:

$$\mathrm{BIC} = -2\ln(\mathcal{L}) + k\ln(n)$$

The Quasi-AIC (QAIC) is an alternative that corrects for overdispersion:

$$\mathrm{QAIC} = \frac{-2\ln(\mathcal{L})}{\hat{c}} + 2K$$

When $\hat{c} = 1$ there is no overdispersion and QAIC = AIC.

IC Considerations

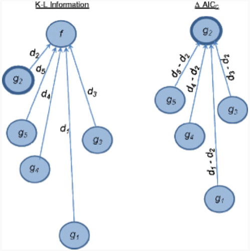
Model Criticism

Because we can’t estimate ‘reality’, IC values only provide relative information (i.e., one model is always going to win out over other models).

Just because a model is the best fit out of a pool of candidates doesn’t mean it’s any good.

After identifying a candidate model it’s important to check all of the assumptions, and test the performance to make sure it’s function reasonably well.

For example, comparing with the intercept only model tells us how much of an improvement our model is over simply looking at the mean.



(Burnham *et al.*, 2011)

Likelihoods and comparable data

IC are based on a model's likelihood. The likelihood is the probability of some model given data $\{x\}$

$$\mathcal{L}(\theta|x)$$

Models don't need to be nested, but if there are any differences in the datasets, the likelihoods, and therefore the IC, are not comparable (this can be a real problem in practice depending on if your data have `NA` values, and how they get handled in `R`).

No clear winner?

Sometimes IC based model selection leads to a clear ‘winner’, other times the differences between the top models are miniscule e.g. ΔAICc of 0.1 (evidence ratio of ~ 1)

What would you do in this situation?

This is where most of the controversy in model selection comes from, and we’ll cover some options for handling that situation next lecture.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In: *Selected papers of hirotugu akaike*. Springer, pp. 199–213.

Burnham, K.P., Anderson, D.R. & Huyvaert, K.P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65, 23–35.

Hachoł, J., Hämmerling, M. & Bondar-Nowakowska, E. (2017). Applying the analytical hierarchy process (ahp) into the effects assessment of river training works. *Journal of Water and Land Development*, 35, 63 – 72.

Walters, C.J. & Ludwig, D. (1981). Effects of measurement errors on the assessment of stock–recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 38, 704–710.