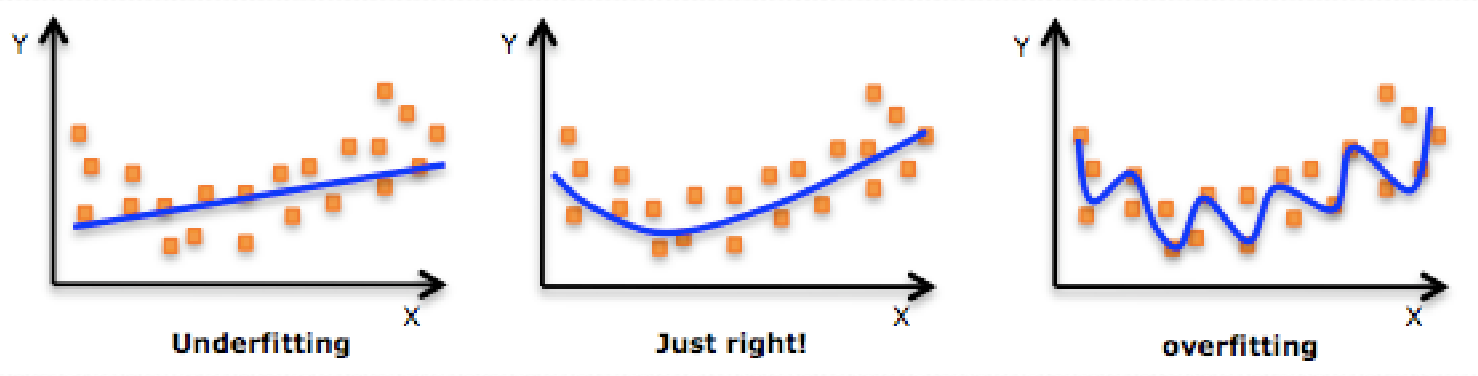
Topic 11: Null hypothesis significance testing (NHST) versus information theoretic (IT) approach and model selection

"A model is always an abstraction and thus strictly always wrong!"

...but some models are useful and our goal must be to search for them

- The more complex a model is, the better it fits the data and residual plots…
- …but at what point do we stop adding complexity?

- There is no unique answer to this question



Null hypthesis significance testing (NHST)

NHST expects mean = 0 or difference = 0

Complaints:

- Arbitrary judgement concerning“statistical significance“
- Confusing scientific hypothesis with statistic hypothesis (running after a nice p-value is not really a scientific hypothesis and a null hypothesis is a statistical test rather like a tool but not what is driving us to do science)
- Null can be biologically trivial
- Only single hypothesis – no multi-hypotheses
- P-value focus often leads to forgetting about effect size (coefficients) and model strength (R^2)

Information theoretic (IT) approach

- Allows to define several plausible alternative hypothesis
- Strength of evidence can be obtained for each hypothesis -> ranking

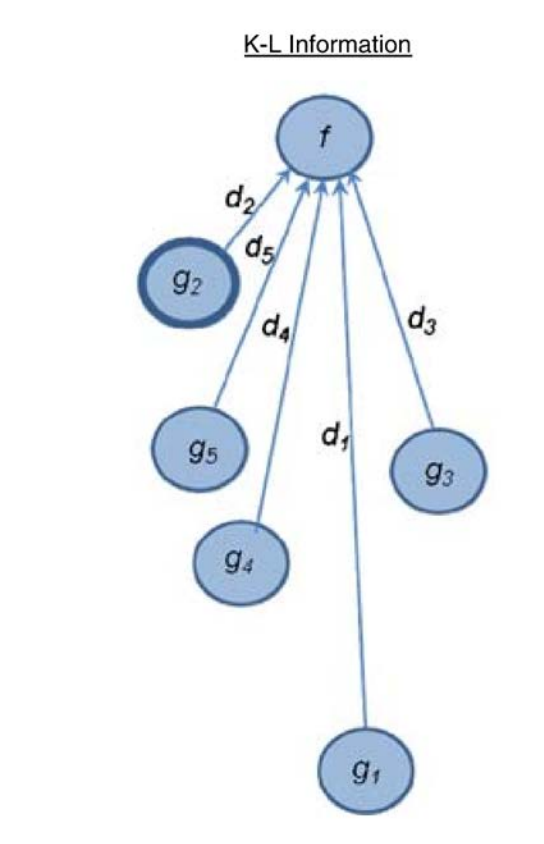
Which model is the
- least bad (best) or
- most plausible or
- most efficient or
- most likely to be truthful model

among all models in the list?

Information theoretic (IT) approach

- Based on Kullback-Leibler's Information Theory (1951)
- offers an estimate of the relative information lost when a given model is used to represent the process that generated the data

- Akaike 1973 adapts this for statistical theory: AIC Akaike information criterion



Akaike information criterion AIC

- AIC is calculated using the number of fitted parameters in the model(k), and either the maximum likelihood estimate for the model (L) or the residual sum of squares of the model (RSS).

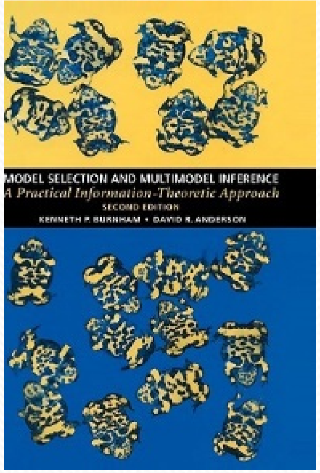
$$\mathrm{AIC} = -2\ln(L) + 2k$$

- AICc = second order bais correction for AIC (Hurvich and Tsai 1989) corrects for small sample size and converges to AIC with larger sample size – so always use AICc

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

Model selection and inference based on AIC

- Burnham and Anderson 1998 and 2002: fundamental book for model selection and inference



Behav Ecol Sociobiol (2011) 65:23–35
DOI 10.1007/s00265-010-1029-6

REVIEW

AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons

Kenneth P. Burnham • David R. Anderson •
Kathryn P. Huyvaert

Δ AICc - Differences

- Report the difference, Δ delta, between the lowest AICc (best model) and the next AICc (model i)
- The larger Δ, the less plausible the model
- Δ0-2: substantial support for model i = model i is equally good as the best model
- Δ2-7: some support (but considerably less)
- Δ >10: essentially no support

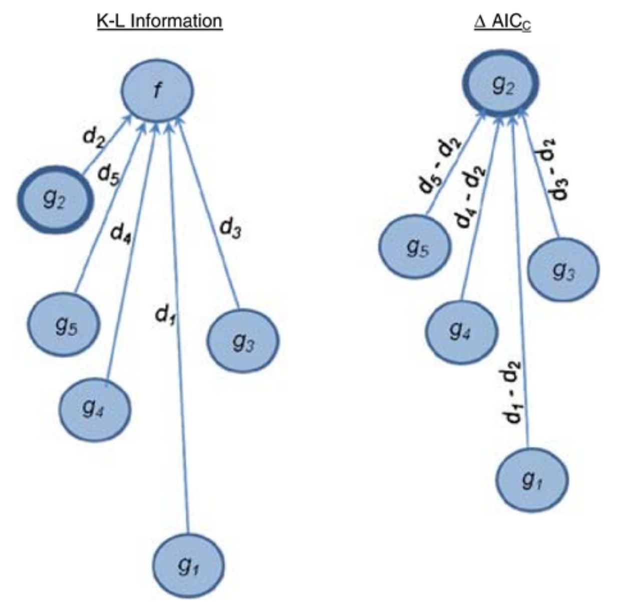


Fig. 1 Kullback–Leibler information is shown (at left) as the distances (d_i) between full reality (f) and the various models (g_i). The Δ values (right) provide the estimated distance of the various models to the best model (in this case, model g_2). These values are on the scale of *information* irrespective of the scale of measurement or type of data. The Δ values are simple to compute, allow a quick ranking of the models, and are the key to multimodel inference

Burnham et al. 2011

Example for R output

Model selection table

	(Intrc)	X1	X2	X3	X4	df	logLik	AICc	delta	weight
4	52.58	1.468	0.6623			4	-28.156	69.3	0.00	0.566
12	71.65	1.452	0.4161		-0.2365	5	-26.933	72.4	3.13	0.119
8	48.19	1.696	0.6569	0.2500		5	-26.952	72.5	3.16	0.116
10	103.10	1.440			-0.6140	4	-29.817	72.6	3.32	0.107
14	111.70	1.052		-0.4100	-0.6428	5	-27.310	73.2	3.88	0.081
15	203.60		-0.9234	-1.4480	-1.5570	5	-29.734	78.0	8.73	0.007
16	62.41	1.551	0.5102	0.1019	-0.1441	6	-26.918	79.8	10.52	0.003
13	131.30			-1.2000	-0.7246	4	-35.372	83.7	14.43	0.000
7	72.07		0.7313	-1.0080		4	-40.965	94.9	25.62	0.000
9	117.60				-0.7382	3	-45.872	100.4	31.10	0.000
3	57.42		0.7891			3	-46.035	100.7	31.42	0.000
11	94.16		0.3109		-0.4569	4	-45.761	104.5	35.21	0.000
2	81.48	1.869				3	-48.206	105.1	35.77	0.000
6	72.35	2.312		0.4945		4	-48.005	109.0	39.70	0.000
5	110.20			-1.2560		3	-50.980	110.6	41.31	0.000
1	95.42					2	-53.168	111.5	42.22	0.000

Models ranked by AICc(x)

AICc weights and evidence ratio

- Report the **weight** of each AICc, which is the likelihood of each model out of all calculated ones, i.e. the probability that model i is best among the listed models
- The evidence ration can be calculated for any two models i and j:
  weight model i / weight model j
  e.g. 0.566/0.119 = 4.756
  i.e. the empirical support for model i is about 4.8 times higher than for model j
  (don't use the term "significant")

Δ_j	Evidence ratio
2	2.7
4	7.4
6	20.1
8	54.6
9	90.0
10	148.4
11	244
12	403
13	665
14	1,097
15	1,808
20	22,026
50	72 billion

What to do with many „best" models – small Δ AICc values?

- Often several models have a very similar AICc, weight and thus likelihood

-> Solution: Multimodel inferences with model averaging

In general inference can (should) be made on all models in the a priori set!

All models or a subset of models will be averaged and predictor variables will be ranked based on their **relative importance!**

Model selection

In the past and still stepwise model selection is used, based on AIC ranking or p-values. (the latter is the worst!)

And **none** of the stepwise methods (backward, forward, both, or blockwise inclusions/exclusions) can be **recommended**!!
Whittingham et al., 2006

Example for R output

```
Model-averaged coefficients:
(full average)
             Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept) 64.693128  22.235479   22.462414   2.880  0.00398 **
X1           1.455798   0.203668    0.219304   6.638  < 2e-16 ***
X2           0.505758   0.268371    0.271702   1.861  0.06268 .
X4          -0.147870   0.252517    0.255126   0.580  0.56219
X3          -0.004302   0.168612    0.175632   0.024  0.98046

(conditional average)
            Estimate Std. Error Adjusted SE z value Pr(>|z|)
(Intercept) 64.69313   22.23548    22.46241   2.880  0.00398 **
X1           1.45580    0.20367     0.21930   6.638  < 2e-16 ***
X2           0.62503    0.12026     0.12917   4.839  1.3e-06 ***
X4          -0.47601    0.22152     0.23094   2.061  0.03929 *
X3          -0.02153    0.37671     0.39244   0.055  0.95624
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Relative variable importance:
                     X1   X2   X4   X3
Importance:          1.00 0.81 0.31 0.20
N containing models:    5    3    3    2
```

Important: do not use the p-values of averaged coefficients!!!

Multimodel inference

- in R e.g. with package MuMIn

Difference between full and conditional average:

- Full average: in cases where high model selection uncertainty exists (i.e. the best AIC model is not strongly weighted) – typically in all-subset modelling
- Conditional average: in cases when there is strong support for the best model (weight>0.9)

- Report the averaged estimates + adjusted SE and the relative importance

How to create a model set a priori?

- Hard thinking!!! What does makes sense? Each model is a hypothesis – which alternative models are useful?
- This can be a long process and ideally happens before the data collection!
- Include the null model!

Alternative:

- all-subset modelling: refers to computing all possible model combinations – often critisized

Table 3 A model set for the example examining ecological factors and extra-pair paternity in a hypothetical bird species

Model description	Model notation
Male body size ('body')	$\beta_0+\beta_{2i}X_{2i}$
Food availability ('food')	$\beta_0+\beta_4X_4$
Male dominance ('status')	$\beta_0+\beta_5X_5$
Territory quality ('territory')	$\beta_0+\beta_7X_7$
Body+food	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4$
Body+status	$\beta_0+\beta_{2i}X_{2i}+\beta_5X_5$
Body+territory	$\beta_0+\beta_{2i}X_{2i}+\beta_7X_7$
Food+status	$\beta_0+\beta_4X_4+\beta_5X_5$
Food+territory	$\beta_0+\beta_4X_4+\beta_7X_7$
Body+food+status	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4+\beta_5X_5$
Body+food+territory	$\beta_0+\beta_{2i}X_{2i}+\beta_4X_4+\beta_7X_7$
Body×status	$\beta_0+\beta_{2i}X_{2i}+\beta_5X_5+\beta_{2i,5}(X_{2i}*X_5)$
Body×territory	$\beta_0+\beta_{2i}X_{2i}+\beta_7X_7+\beta_{2i,7}(X_{2i}*X_7)$
Food×territory	$\beta_0+\beta_4X_4+\beta_7X_7+\beta_{4,7}(X_4*X_7)$
Intercept only	β_0

A great guide to model selection

Behav Ecol Sociobiol (2011) 65:13–21
DOI 10.1007/s00265-010-1037-6

REVIEW

A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion

Matthew R. E. Symonds · Adnan Moussalli

JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2010.02210.x

REVIEW

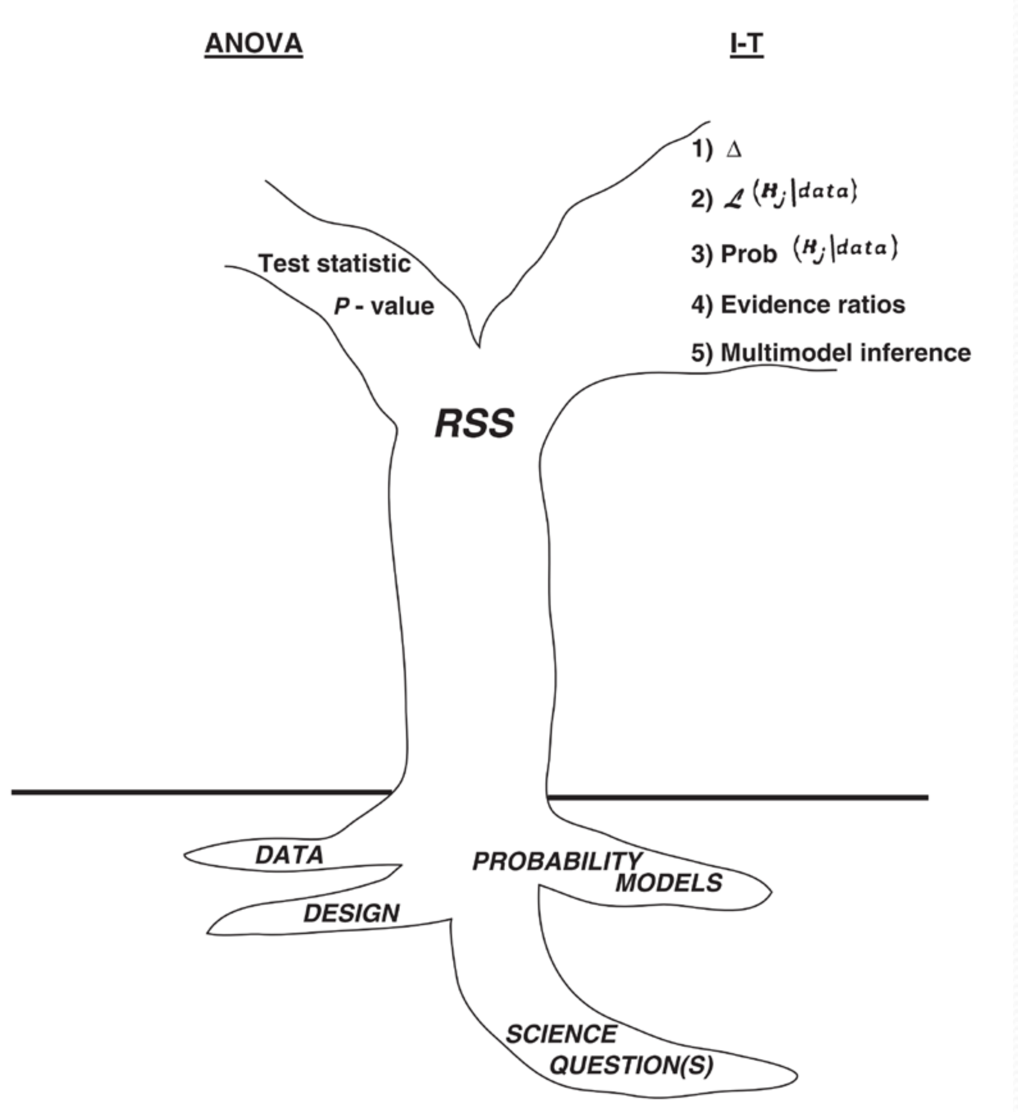
Multimodel inference in ecology and evolution: challenges and solutions

C. E. GRUEBER, S. NAKAGAWA, R. J. LAWS & I. G. JAMIESON

Department of Zoology, University of Otago, Dunedin, New Zealand

- In the appendix contains a worked example for multimodel inference including interepretation.
- Only take care with the proposed functions to calculate Confidence Intervalls and to male prediction – this is not perfectly correct (see Corrigendum of the paper)
- Currently, it is only possible with Bayesian methods to take into account all uncertainties.

NHST versus IT



When to use what?

NHST

- For planned controlled experiements
- Randomized and replicated treatments
- Few treatments (>3)
- BUT: report coefficients, not just p and CIs and R^2

- Ideally we have simple experiments

IT

- For more complicated experiments (e.g. mixed effects, nested)
- Regression models
- Observational data
- Also report coefficients, CIs and and R^2

- Reality often much more complicated

Further information

Model selection with AICc
https://www.youtube.com/watch?v=7XAHjm6Vy5k&t=148s

Other interesting topics of the same author (but I did not check it‘s reliability):
https://www.youtube.com/channel/UCExBFDFe1XbNbtooyYXaxPg