Cary Institute of Ecosystem Studies

Columbia University in the City of New York

Day 6

Model Selection and Multimodel Inference

Topics

- Model selection: how do you choose between (and compare) alternate models?
- Multimodel inference: how do you combine information from more than 1 model?

Topics

- Model comparison.
- Evidence ratios.
- Likelihood ratio tests.
- Akaike Information Criterion
- Akaike weights.
- Multimodel inference

Comparing alternative models

- We don't ask "Is the model right or wrong?" We ask "Do the data support one model more than a competing model?"

- Strength of evidence (support) for a model is relative:
  - Relative to other models: As models improve, support may change.
  - Relative to data at hand: As the data improve, support may change.

Bias and Uncertainty in Model Selection

- Model Selection Bias: Chance inclusion of meaningless variables in a model will produce a biased underestimate of the variance, and a corresponding exaggeration of the precision of the model
(the problem with "fishing expeditions")

- Model Selection Uncertainty: The fact that we are using data (with uncertainty) to both estimate parameters and to select the best model necessarily introduces uncertainty into the model selection process

See discussion on pages 43-47 of Burnham and Anderson

Comparing alternative models: methods

- Likelihood ratio tests
  - Limited to comparisons between two models
- Akaike's Information Criterion (AIC)
  - Can be used to simultaneously assess many models

Remember: you can only directly compare alternate models applied to exactly the same dataset…

Recall the Likelihood Principle…

"Within the framework of a statistical model, a set of data supports one statistical hypothesis better than the other if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis".

(Edwards 1972)

But remember parsimony..

- A more complex model (more parameters) is expected to have higher likelihood, so we need some way to penalize models with higher numbers of parameters..

Likelihood ratios

- The likelihood ratio L[A(x)] /L[B(x)] is a measure of the strength of evidence favoring model (hypothesis) A over model (hypothesis) B.

- Issues:
  - What constitutes a "big" difference?
  - How do you penalize a model if it uses more parameters?

Likelihood ratio tests (LRT)

- LRT follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between models A and B.

$$LRT = 2[\log L(x \mid \theta_A) - \log L(x \mid \theta_B)]$$

$$LRT \approx \chi^2, \text{ with } df = \text{difference in number of parameters}$$

Remember: if the two models have the same number of parameters, just use likelihood to compare them…

df	Critical value of χ2 with p=0.05	Critical log likelihood difference
1	3.84	1.92
2	5.99	3.00
3	7.81	3.91
4	9.49	4.74
5	11.07	5.54
6	12.59	6.30
7	14.07	7.03
8	15.51	7.75
9	16.92	8.46
10	18.31	9.15

Limitations of Likelihood Ratio Tests

- Can only compare a pair of models at a time… (gets clumsy when you have a larger set of models)
- Requires that you use a traditional frequentist "p-value" as your basis for judging between models…

A more general framework for model comparison: Information theory

- "Reality" = "Truth" = Unknowable (or at least too much trouble to find…)
- Models are approximations of reality, and we'd like to know how "close" they are…
- The "distance" between a model and reality is defined by the "Kullback-Leibler Information" (K-L distance)
- Unfortunately, K-L distance can only be directly computed in hypothetical cases where reality is known..

See Chapter 2 of Burnham and Anderson for discussion and details…

Interpretation of Kullblack-Leibler Information

- Information entropy = information content of a random outcome
- Minimizing KL is the same as maximizing entropy.
- We want a model that does not respond to randomness but does respond to information.
- We maximize entropy subject to the constraints of the model used to capture information in the data.
- By maximizing entropy, subject to a constraint, we leave only the information supported by the data. The model does not respond to noise

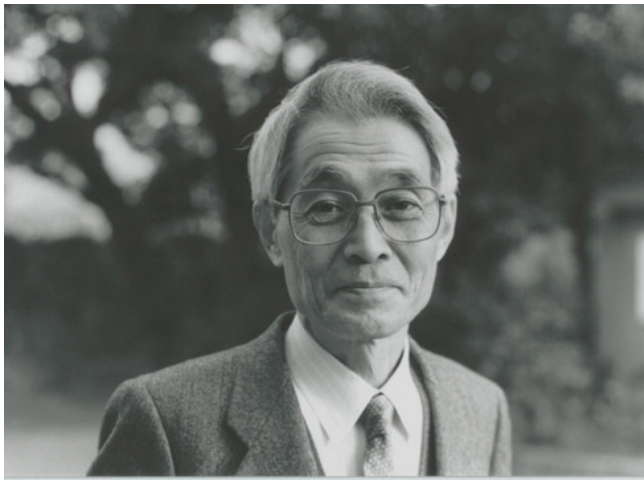
Akaike's contribution (1973)

- Akaike (1973) proposed "an information criterion" (AIC) (but now often called an Akaike Information Criterion) that relates likelihood to K-L distance, and includes an explicit term for model complexity…

$$AIC = -2\,ln(\,L(\hat{\theta}\,|\,y\,)) + 2K$$

This is an estimate of ***the expected, relative distance between the fitted model and the unknown true mechanism that generated the observed data***.

K=number of estimated parameters



Akaike's Information Criterion

- AIC has a built in penalty for models with larger numbers of parameters.

$$AIC = -2\,ln(L(\hat{\theta} \mid y)) + 2K$$

- Provides implicit tradeoff between bias and variance.

AIC

- We select the model with **smallest** value of AIC (i.e. closest to "truth").
- AIC will identify the best model in the set, even if all the models are poor!
- It is the researcher's (your) responsibility that the set of candidate models includes well founded, realistic models.

AIC for small samples

- Unless the sample size *(n)* is large with respect to the number of estimated parameters *(K)*, use of AICc is recommended.

$$AIC_c = -2\,ln(L(\theta \mid y)) + 2K\left(\frac{n}{n-K-1}\right)$$

- Generally, you should use AICc when the ratio of n/K is small (less than ~ 40), based on K from the global (most complicated) model.
- Use AIC or AICc consistently in an analysis rather than mix the two criteria.

Some Rough Rules of Thumb

- Differences in AIC (Δi's) can be used to interpret strength of evidence for one model vs. another.
- A model with a Δ value within 1-2 of the best model has substantial support in the data, and should be considered along with the best model.
- A Δ value within only 4-7 units of the best model has considerably less support.
- A Δ value > 10 indicates that the worse model has virtually no support and can be omitted from further consideration.

Comparing models with different PDFs

- LRTs and AIC can be used as one basis for selecting the "best" PDF for a given dataset and model,
- But more generally, an examination of the distribution of the residuals should guide the choice of the appropriate PDF
- There will be cases where different PDFs are appropriate for different models applied to the same dataset
  - Example: neighborhood competition models where residuals shift from lognormally to normally distributed as the models are improved by additional terms

Strength of evidence for alternate models: Akaike weights

Akaike weights (w_i) are ***the weight of evidence in favor of model i being the actual best model*** for the situation at hand given that one of the N models must be the best model ***for that set of N models.***

$$w_i = \frac{exp(-0.5\Delta_i)}{\sum_{r=1}^{N} exp(-0.5\Delta_r)}$$

where

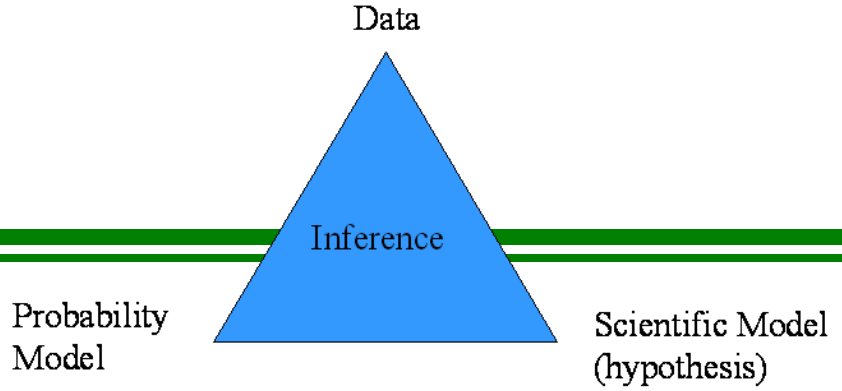
$$\Delta_i = AIC_i - AIC_{min}$$

Akaike weights for all models combined should add up to 1.

Uses of Akaike weights

- "Probability" that the candidate model is the best model.
- Relative strength of evidence (evidence ratios).
- Variable selection—which independent variable has the greatest influence?
- Model averaging.

An example...



The Data:

x_i = measurements of DBH on 50 trees

y_i = measurements of crown radius on those trees

The Scientific Models:

$y_i = \beta x_i + \varepsilon$ [1 parameter (β)]

$y_i = \alpha + \beta x_i + \varepsilon$ [2 parameters (α, β)]

$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon$ [3 parameters (α, β, γ)]

The Probability Model:

ε is normally distributed, with mean = 0 and variance estimated from the observed variance of the residuals...

Back to the example…..

	Likelihood	No. parameters
Model 1: Radius = B*DBH	**130.42**	1
Model 2: Radius = A + B*DBH	**108.01**	2
Model 3: Radius = A + B*DBH + C*(DBH^2)	**107.99**	3

	AIC	Akaike weight
Model 1: Radius = B*DBH	264.84	0.00
Model 2: Radius = A + B*DBH	222.02	0.73
Model 3: Radius = A + B*DBH + C*(DBH^2)	223.97	0.27

Akaike weights can be interpreted as the estimated probability that model *i* is the best model for the data at hand, given the set of models considered. Weights > 0.90 indicate that robust inferences can be made using just that model.

Akaike weights and the relative importance of variables

- For nested models, estimates of relative importance of predictor variables can be made by summing the Akaike weights of variables across all the models where the variables occur.

- Variables can be ranked using these sums.

- The larger this sum of weights, the more important the variable is.

Example: detecting density dependence

TABLE 1. Seven example species representing major taxa with long-term (≥14 years) time series abundance data.

Common name	Scientific name	Taxon	q	AIC_c model weight					wt DD (%)
				RW	EX	RL	GL	TL	
Whooping Crane	*Grus americana*	BIR	57	0.213	**0.430**	0.146	0.163	0.049	35.8
Grizzly bear	*Ursus arctos*	MAM	38	**0.524**	0.262	0.092	0.095	0.027	21.4
Snapping turtle	*Chelydra serpentina*	REP	14	**0.451**	0.124	0.166	0.230	0.029	42.4
Atlantic salmon	*Salmo salar*	FIS	110	0.017	0.006	0.367	**0.459**	0.150	97.6
Desert locust	*Schistocerca gregaria*	INS	104	0.002	0.001	0.005	**0.964**	0.028	99.7
Spiny lobster	*Panulirus interruptus*	MIN	62	0.096	0.035	0.121	**0.607**	0.141	86.9
Blue grama	*Bouteloua gracilis*	PLA	22	0.096	0.037	0.090	0.047	**0.729**	86.7

Notes: Abbreviations for taxa are: birds, BIR; mammals, MAM; reptiles, REP; fish, FIS; insects, INS; marine invertebrates, MIN; and plants, PLA; q is the median number of yearly transitions. Shown are relative strengths of evidence for five a priori population dynamics models (Akaike's Information Criterion [AIC_c] weight) under density-independent (random walk [RW], exponential [EX]) and density-dependent (Ricker-logistic [RL], Gompertz-logistic [GL], and θ-logistic [TL]) growth. The sum of AIC_c weights for the density-dependent models represents the combined percentage weight for those models (wt DD). The binary outcomes (yes [Y] or no [N]) for the selection of density dependence using AIC_c, Bayesian Information Criterion (BIC), cross-validation (C-V; Turchin 2003), R (Bul; Bulmer 1974), randomization (Ran; Pollard et al. 1987), and parametric bootstrap likelihood ratio test (PBLR; Dennis and Taper 1994) are shown, as is whether a lagged density-dependent response was detected by AIC_c and C-V. The values in boldface show the model with the highest AIC_c weight per taxon.

Source: Brook, B.W. and C.J.A. Bradshaw. 2006. Strength of evidence for density dependence in abundance time series of 1198 species. Ecology 87:1445-1451.

Ambivalence about selecting a best model to use for inference…

The inability to identify a single best model is not a defect of the AIC method. It is an indication that the ***data are not adequate to reach strong inference.***

What is to be done??

MULTIMODEL INFERENCE AND MODEL AVERAGING

Multimodel Inference

- If one model is clearly the best (w_i>0.90) then inference can be made based on this best model.
- Weak strength of evidence in favor of one model suggests that a different dataset may support one of the alternate models.
- Designation of a single best model is often unsatisfactory because the "best" model is highly variable.
- We can compute a weighted estimate of the parameter and the predicted value using Akaike weights.

Akaike Weights and Multimodel Inference

						AIC	Akaike weights
Model 2: Radius = A + B*DBH						**222.02**	0.73
Model 3: Radius = A + B*DBH + C*(DBH^2)						**223.97**	0.27

- Estimate parameter values for the models with at least some measurable support.
- Estimate weighted average of parameters across those models.
- Only applicable to linear models.
- For non-linear models, we can generate weighted averages of the predicted response value for given values of the predictor variables.

Akaike Weights and Multimodel Inference

						AIC	Akaike weights
Model 2: Radius = A + B*DBH						**222.02**	0.73
Model 3: Radius = A + B*DBH + C*(DBH^2)						**223.97**	0.27

Estimate of parameter $A = (0.73*1.04) + (0.27*1.31) = 1.11$

Multimodel Inference: An example

$$RG = MaxRG * e^{-1/2\left[\frac{(g/G_o)}{G_b}\right]^2} e^{-1/2\left[\frac{\ln(DBH_t/X_o)}{X_b}\right]^2} e^{-C\left[\sum_{i=1}^{s}\sum_{j=1}^{n}\lambda_s\frac{(DBH_{ij})^{\alpha}}{(dist_{ij})^{\beta}}\right]^D (DBH_t)^{\gamma}}$$

- Neighborhood models of tree growth:
  - Can we use MMI to improve parameter estimates for individual terms in the model? (not easily, given non-linearities in this model)
  - Can we use MMI to improve predictions of growth using a weighted suite of alternate models? (yes, but is it worth the effort?)

See: Papaik, M. J., and C. D. Canham. 2006. Multi-model analysis of tree competition along environmental gradients in southern New England forests. Ecological Applications **16**:1880-1892.

Summary: Steps in Model Selection

- Develop candidate models based on biological knowledge.
- Take observations (data) relevant to predictions of the model.
- Use data to obtain MLE of parameters of the alternate models.
- Evaluate strength of evidence for alternate models using AIC and Akaike weights.
- …Multimodel Inference?

Do you agree with Burnham and Anderson that MMI is generally preferable to "best-model inference"?