Model Selection and Model Averaging in Phylogenetics

Johan Nylander

November 27, 2005

1 Key Concepts

- Model selection - to use the data to select a model - should be an integral part of inference [4].
- The data generating or "true" model (f) has an infinite number of parameters and is unreachable.
- The best approximate model (g): best descriptive model given the limited sample size. Finding the best g is (or can be) the goal of model selection.
- A more parameter-rich model has a higher potential than a less parameter rich model: less *discrepancy due to approximation*. However, a more parameter-rich model tends to perform farther below its potential than a less parameter rich model caused by the *discrepancy due to estimation* [20].
- Parsimonious trade off between error (decreases with additional parameters) and variance (increases with additional parameters).
- To help us with the trade off: apply a model selection criterion.

2 Model Selection Criteria in Phylogenetics

2.1 Likelihood

- Changing the model changes the likelihood - (which is proportional to) the probability of data, given the parameters and the model [5].

- Maximized log Likelihood is biased upward as an estimator of the target model. The bias is proportional to the number of parameters [4].

2.2 Likelihood Ratio Testing [7]

$$\delta = -2(\ln L_0 - \ln L_1)$$

- Basic idea: Is the increase in likelihood significant?
- δ is asymptotically χ^2_n distributed, with df n, the difference in number of free parameters between models.
- Only for nested models (model L_0 must be a special case of L_1).
- Mixed χ^2_n distributions when one parameter is in its limit (e.g., GTR vs. GTR+G) [19].

Applications: *Modeltest* [13], *MrModeltest2* [15]

2.3 AIC - Akaike Information Criterion [1, 4]

$$AIC_i = -2\ln(L) + 2p$$

- L: Max. log Likelihood for model i, p: number of parameters.
- Estimates the expected Kullback-Leibler (K-L) distance: information lost when model g is used to approximate f.
- Min AIC is the best K-L model in the set of competing models.
- No accept or reject (not a strict test).
- Applies to nested and non-nested models.
- AICc - takes sample size in to account.
- Must be based on the maximum likelihood - problematic(?).

Applications: *Modeltest* [13], *MrModeltest2* [15], *MrAIC.pl* [16], *FindModel* [8], etc.

2.4 BIC - Bayesian Information Criterion [18, 4]

$$BIC = -2\ln(L) + p\ln(n)$$

- L: Max. Likelihood, p: parameters, n: sample size.
- Not sure what the sample size in phylogenetics really is (open area for research!)
- Min. BIC is the best model in the set of competing models.
- No accept or reject (not a strict test).
- Applies to nested and non-nested models.
- Designed for a different purpose than AIC: AIC selects the best K-L model (the g that minimizes the K-L distance to f) in the set of candidate models while BIC will select f when sample size increases (IF f is in the set!).

Applications: *MrAIC.pl* [16], *Modeltest* [13].

2.5 Bayes Factors [10]

- The goal is to calculate posterior probabilities of different models given the same data.
- Using Bayes rule, the posterior probability of model M_k (k=1,2) given data D is:

  $$P(M_k \mid D) = \frac{P(D \mid M_k)\,P(M_k)}{P(D \mid M_1)\,P(M_1) + P(D \mid M_2)\,P(M_2)}$$

  and the ratio of posterior odds to the prior odds is

  $$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)}{P(D \mid M_2)}\frac{P(M_1)}{P(M_2)}$$

  and we see that the transformation between the posterior and prior odds is done by multiplying with

  $$BF_{12} = \frac{P(D \mid M_1)}{P(D \mid M_2)}$$

  which is the *Bayes factor* [10].
- If we allow the prior model probabilities to be equal, we can get the ratio of model probabilities from the ratio of model likelihoods $P(D|M_k)$. The model likelihoods (or integrated likelihoods, or predictive likelihoods) are obtained by *integrating* (not maximizing) over the parameter space.

- The model likelihood is the denominator of Bayes rule which is difficult to calculate but can be approximated using Markov chain Monte Carlo.
- Applies to nested and non-nested models.
- No accept or reject (not a strict test).
- Accounts for uncertainty in parameter estimates.

$2 \log B_{10}$	B_{10}	Evidence against M_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Applications: *MrBayes* [17]

3 Model Accuracy

A model can be "best" according some criterion, but is it adequate?

3.1 Penelized Likelihood - Decision Theory

- Additional to AIC/BIC, uses an *ad-hoc* loss function, e.g., branch length variance.
- Minimize posterior risk, R_i of choosing a wrong model:

$$R_i = \sum_{j=1}^{M} l_{ij} P(M_j \mid D) = 1 - P(M_i \mid D)$$

- where l_{ij} is 1 if model i is chosen when model j is correct and $l_{ij} = 0$ if the correct model is chosen. M is the set of competing models.

Applications: *DT-ModSel* [12]

3.2 Parametric Bootstrap [6]

- When the χ^2 approximation to δ in the LRT does not hold (e.g., non-nested models).
- Simulate data under a null hypothesis (null model).
- Formally, If one has a parametric distribution $P(\theta)$ which is well defined except for the parameter θ, instead of drawing with replacement from the original sample, we can draw from the distribution $F(\hat{\theta})$, where $\hat{\theta}$ is the estimation obtained from the original sample.
- Test: How extreme is the observed data under the (null) model?

3.3 Posterior Predictive Checks

- Bayesian version of the parametric bootstrap.
- An adequate model would be good in predicting future data
- Simulate new data $D\prime$ from the posterior distribution $P(D \mid \theta, \tau)$.
- An adequate model would have a high $P(D\prime \mid D)$, the posterior probability of $D\prime$ given the original data D.
- A multinomial test statistic $T(D)$, is used for comparing the observed and generated distributions of site patterns. [3].

Applications: *MAPPS* [3]

4 How Much Better is the Best Model?

4.1 Akaike weights [4]

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta AIC_i\right)}{\sum_{j=1}^{R}\exp\left(-\frac{1}{2}\Delta AIC_j\right)}$$

- $\Delta AIC_i = AIC_i - AIC_{MIN}$,where AIC_{MIN} is set to the best K-L model in the set R of competing models.
- A shortcut for estimating the pobability of models.

- Can be used for Occams's window [11][14].

Applications: *Modeltest, MrModeltest, MrAIC.pl*

5 Model Averaging

- Basing inference on a single model selected on the basis of data ignores model uncertainty. By doing this, we tend to underestimate the uncertainty in parameter estimates and overestimate the strength of support for hypothesis.
- One way to account for model uncertainty is to allow all models to contribute to inference by averaging:

$$\hat{\theta}_{MA} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

where $\hat{\theta}_i$ is the parameter estimate under model i in the set R of candidate models, and w_i is the weight for model i.

5.1 Bayesian Model Averaging

- Takes *both* model selection uncertainty and parameter uncertainty in to account.
- Allows all models to contribute in proportion to their probability:

$$\hat{\theta}_{BMA} = \sum_{i=1}^{R} P\left(M_i \mid D\right) \hat{\theta}_i$$

where the weight for model i is the posterior probability of the model given the data; $P\left(M_i \mid D\right)$.

- Bayesian model averaging can be accomplished:
  1. By marginal likelihood estimation and Bayes Factors [14][2].
  2. By Reversible jump MCMC [9].

References

[1] Akaike, H. 1973. Information theory as an extension to the maximum likelihood principle. Pages 267-281 *in* Second international symposium on information theory (Petrov, B. N., and F Csaki, eds.) Akademiai Kiado, Budapest.

[2] Beier, B.-A., J. A. A. Nylander, M. W. Chase, and M. Thulin. 2004. Phylogenetic relationships and biogeography of the desert plant genus Fagonia (Zygophyllaceae), inferred by parsimony and Bayesian Model Averaging. Mol. Phylogen. Evol. 33:91-108.

[3] Bollback, J. P. 2003. Bayesian model adequacy and choice in phylogenetics.. Mol. Biol. Evol. 19:1171-1180.

[4] Burnham, K. P. and D. R. Anderson, 2002. Model selection and multi model inference. A practical information-theoretic approach. Springer, New York.

[5] Edwards, A.W.F. 1992. Likelihood. Expanded Edition. John Hopkins Univ. Press. London.

[6] Efron, B. 1979. Bootstrap methods: Another look at the Jackknife. Ann. Statist. 7:1-26.

[7] Felsenstein, J. 2004. Inferring phylogenies. Springer, New York.

[8] FindModel. Los Alamos National Laboratory. URL:http://hcv.lanl.gov./content/hcv-db/findmodel/findmodel.html

[9] Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123-1133.

[10] Kass, R.E. and A.E. Raftery, 1995. Bayes Factors. J. Am. Stat. Assoc. 90:773-795.

[11] Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model selection uncertainty in graphical models using Occam's window. J. Am. Stat. assoc. 89:1535-1546.

[12] Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2004. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674-683.

[13] Posada, D. and K. A. Crandall, 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14:817-818.

[14] Nylander, J. A. A. 2004. Bayesian Phylogenetics and the evolution of Gall wasps. Ph.D Dissertation, Uppsala University.

[15] Nylander, J.A.A. 2004. MrModeltest2. Software distributed by the author. Evolutionary Biology Centre, Uppsala University.

[16] Nylander, J. A. A. 2004. MrAIC.pl. Software distributed by the author. Evolutionary Biology Centre, Uppsala University.

[17] Ronquist, F. and J. P. Huelsenbeck. 2003. MRBAYES 3:Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

[18] Schwartz, G 1978. Estimamting the dimension of a model. Ann. Stat. 6:461-464.

[19] Whelan, S. and N. Goldman. 1999. Distributions of Statistics used for the comparison of models of sequence evolution in phylogenetics. Mol. Biol. Evol. 16:1292-1299.

[20] Zucchini, W. 2000. An introduction to model selection. J. Math. Psych. 44:41-61.