Statistics on Trees

Heiko Schmidt / Greg Ewing

CIBIV

June 8, 2007

Bootstrapping (Efron, 1979)

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability, i.e., mean and standard deviation.
- We want to be able to do the same with trees.
- The bootstrap is a general statistical method that can be used in this case.
  - Nonparametric bootstrap, just re-samples the alignment.
  - Parametric bootstrap uses model parameters to generate replicate data.
- Bayesian methods usually get this for “free” because we already have a large set of trees that represent potions in the posterior density.

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.
- Studies indicate that it’s quite conservative.

Cons

- Results have no convenient interpretation. ie 50% support does not mean 50% probability.
- Some strong assumptions are imposed on the data. ie iid.
- Relies on the fact that the data sample we are using is representative of entire “population” of data.

Bootstrap flow

- Estimate a ML tree and the model parameters θ.
- From the data/or estimate generate replicate data sets.
- For each replicate data set estimate a replicate ML tree.
- Combine the replicate ML trees into some kind of consensus tree.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
  - A site, or column in the alignment is picked at random.
  - This column of sequence data is placed into the replicate alignment.
  - Some columns will appear more than once in the replicate alignment.
  - Other columns will not appear at all.
- Requires that the data is i.i.d. across sites.

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A
A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A
A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A
A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	T	A	C	C	C	T	T	T	T

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.
  - Using these estimated parameters and the estimated ML tree simulate a new replicate data set.
  - Estimate a new ML tree and parameters θ'.
  - In some cases model parameters can be fixed.
- Parametric bootstraps do not make any extra assumptions about the data over the model.

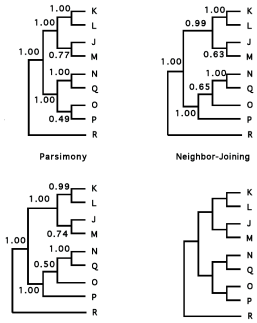
Combining the trees

- **50% majority rule consensus** is conservative and no two branches/splits can be conflicting.
- **Extended consensus rules** can vary slightly in implementation.
- In particular the **extended majority rule consensus** (default in PHYLIP's consense) can have splits in the final tree that conflict with splits that are more frequent.

Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.
- It is not a probability.
- Generally it appears to be somewhat conservative.
- On the other hand it is not uncommon to see high bootstrap support for the wrong tree.
- One interpretation is that the bootstrap attempts to measure sampling variance. (Swofford, et al 1996)

Example Support of a known tree



Hills et al, 1992. Bacteriophage T7 DNA sequences with a known phylogeny.

Bayesian Support Values: an aside

- Can be interpreted as Probabilities. That is the probability that this node is in the "true" tree given the data.
- Cannot be directly compared to bootstrap values.
- But not independent from bootstraps either.
- Bayesian Support $\neq$ Bootstrap support.

Hypothesis testing

- What question do I want to answer?
  - Say should I use the JC model or the GTR model?
  - Or perhaps, Is tree A statistically different from tree B?
- It's important to note that you should know the null hypothesis/hypotheses **before** you "collect" the data.

Testing with Likelihood (LRT, KH, SH)

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- E.g., a clocklike tree is nested in his non-clocklike variant, JC69 is nested in GTR.
- In such a situation we can consider the likelihood of both models.
- The Null Hypothesis H_0: Both models are equally good.
- The alternative Hypothesis H_A: The more complicated model is better.
- Note that the more complicated model always has an equal or higher likelihood (since we have more parameters to tweek).
- Here, we can use a (log-)**likelihood ratio test** (LRT).

LRT

Log Likelihood ratio test

$$\lambda = -2 \log \frac{L_0}{L_1} = 2(\log L_1 - \log L_0)$$

L_0: likelihood of the less parameter-rich model (Null hypothesis, H_0)
L_1: likelihood of the more parameter-rich model (alternative, H_1)

- λ is asymptotically distributed to the χ^2 distribution with the appropriate degrees of freedom.
- The degrees of freedom are the difference in number of parameters between the two models, e.g., JC69 – HKY85 have 4 parameters, GTR – GTR-Γ have 1 parameters difference.
- We calculate λ and check if it's outside our P-value range on the χ^2 distribution.

Testing Tree Topologies

- **Only nested models** can be tested:
  One model (H_0, Null-model, constraint model) is nested in another model (H_A, alternative, unconstraint model) if the model H_0 can be produced by restricting parameters in model H_A.
- two differrent topologies are not nested.
- Thus, LRT cannot be used on different topologies, because the assumption of the χ distribution does not fit.
- Hence, other (bootstrap-bases) methods have been devised to determine the distribution of log-likelihood differences for testing (e.g., KH or SH test).

Basic Idea:

- Compute log-likelihood values $L_1, \ldots, L_N$ for your trees $T_1, \ldots, T_N$.
- Draw bootstrap samples i from the alignment, re-estimate the log-likelihood values $L_x^{(i)}$ for each tree T_x and for each sample i.
- Adjust the log-likelihoods with the mean by setting $\tilde{L}_x^{(i)} = L_x^{(i)} - \bar{L}_x^{(i)}$ (Centering)
- Use the differences between the $\tilde{L}_x^{(i)}$ to determine the distribution of differences $\delta^{(i)} = \tilde{L}_y^{(i)} - \tilde{L}_z^{(i)}$.
- Use the distribution of $\delta^{(i)}$ to test your trees.

Time Saving: RELL

- The re-optimization to get the log-likelihood values $L_x^{(i)}$ is very time consuming.
- Hence, often the site-likelihoods are used fixed.
- During the bootstrap the already estimated site-log-likelihoods are sampled and added to produce $L_x^{(i)}$.
- The resampling of estimated log-likelihoods (RELL) has been shown to be often sufficient to produce the distribution of log-likelihood differences.

Original Kishino and Hasegawa test (KH test)

- This test was devised to test whether two *a priory* chosen trees (e.g., from a Markov Chain) are equally well supported by the dataset.
- H_0: the expected $\delta = L_1 - L_2 = 0$.
  H_A: the expected $\delta = L_1 - L_2 \neq 0$.
  KH assumes that the ML tree is not among the trees.

Mis-use of the Kishino and Hasegawa test (KH test)

- Often, instead two *a priory* chosen trees, one tree is tested against the ML tree T_{ML}.
- That means, $\delta = L_{ML} - L_1$ is rarely negative.
- Hence, δ has to be tested in a single-sided regime.
- H_0: the expected $\delta = L_1 - L_2 = 0$. H_A: the expected $\delta = L_1 - L_2 > 0$.

Multiple trees (Shimodaira and Hasegawa test - SH test)

- The SH test offers a correct way to test a set of trees, which may be chosen *a posteriory* after ML analysis.
- H_0: All trees including T_{ML} are equally supported. H_A: Some or all trees T_x are not equally well supported.
- The SH test assumes, that the ML tree T_{ML} is among the trees.

The SH procedure:

- Compute log-likelihood values and the differences $\delta_x = L_{ML} - L_x$ for your trees.
- Draw bootstrap samples i from the alignment (with RELL) to gain log-likelihood values $L_x^{(i)}$ for each tree T_x.
- Adjust the log-likelihoods with the mean over the samples i by setting $\tilde{L}_x^{(i)} = L_x^{(i)} - \bar{L}_x^{(i)}$ (Centering)
- For each sample i, find $\tilde{L}_{ML}^{(i)}$ over all topologies T_x.
- and compute $\delta_x^{(i)} = \tilde{L}_{ML}^{(i)} - \tilde{L}_x^{(i)}$.
- For each tree T_x, test whether δ_x is a plausible sample from the distribution of $\delta_x^{(i)}$ (over all replicates i).
- We use a single sided test, since $\tilde{L}_{ML}^{(i)} \geq \tilde{L}_x^{(i)}$.

Summary

- Likelihoods gives a strong statistical framework for hypothesis testing.
- Proper experimental design and proper use of tests is required.
- One should alway be aware of the hypothesis a test assesses and should make sure that this answers the question asked.
- Testing tree topologies can be used to assess whether two competing hypotheses are really substantially different. If they are not, one cannot be prefered over the other.