Lecture 27. Phylogeny methods, part 7 (Bootstraps, etc.)

Joe Felsenstein

Department of Genome Sciences and Department of Biology
A non-phylogeny example of the bootstrap

Bootstrap replicates

(unknown) true value of $\theta$

(unknown) true distribution

empirical distribution of sample

Distribution of estimates of parameters

Bootstrap sampling from a distribution (a mixture of two normals) to estimate the variance of the mean
Bootstrap sampling

To infer the error in a quantity, \( \theta \), estimated from a sample of points \( x_1, x_2, \ldots, x_n \) we can

- Do the following \( R \) times (\( R = 1000 \) or so)
- Draw a “bootstrap sample" by sampling \( n \) times with replacement from the sample. Call these \( x_1^*, x_2^*, \ldots, x_n^* \). Note that some of the original points are represented more than once in the bootstrap sample, some once, some not at all.
- Estimate \( \theta \) from the bootstrap sample, call this \( \hat{\theta}_k^* \) (\( k = 1, 2, \ldots, R \))
- When all \( R \) bootstrap samples have been done, the distribution of \( \hat{\theta}_i^* \) estimates the distribution one would get if one were able to draw repeated samples of \( n \) points from the unknown true distribution.
Bootstrap sampling of phylogenies

Original Data

sequences

Bootstrap sample #1

sequences

Bootstrap estimate of the tree

bootstrap sample #1

sample same number of sites, with replacement

sequences

Bootstrap estimate of the tree, #1

bootstrap sample #2

sample same number of sites, with replacement

sequences

Bootstrap estimate of the tree, #2

(and so on)
More on the bootstrap for phylogenies

- The sites are assumed to have evolved independently given the tree. They are the entities that are sampled (the $x_i$).
- The trees play the role of the parameter. One ends up with a cloud of $R$ sampled trees.
- To summarize this cloud, we ask, for each branch in the tree, how frequently it appears among the cloud of trees.
- We make a tree that summarizes this for all the most frequently occurring branches.
- This is the *majority rule consensus tree* of the bootstrap estimates of the tree.
Majority rule consensus trees

Trees:

How many times each partition of species is found:

- AE | BCDF 3
- ACE | BDF 3
- ACEF | BD 1
- AC | BDEF 1
- AEF | BCD 1
- ADEF | BC 2
- ABDF | EC 1
- ABCE | DF 3

Majority–rule consensus tree of the unrooted trees:
An example of bootstrap sampling of trees

232 nucleotide, 14-species mitochondrial D-loop analyzed by parsimony, 100 bootstrap replicates
Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a “multiple-tests" problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods
Other resampling methods

- Delete-half jackknife. Sample a random 50% of the sites, without replacement.

- Delete-$1/e$ jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).

- Reweighting characters by choosing weights from an exponential distribution.

- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1

- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth

- (to correct for correlation among adjacent sites) (Künsch, 1989)

Block-bootstrapping – sample $n/b$ blocks of $b$ adjacent sites.
Delete half jackknife on the example
Calibrating the jackknife

Exact computation of the effects of deletion fraction for the jackknife
(suppose 1 and 2 are conflicting groups)

\[ n \text{ characters} \]
\[ n(1-\delta) \text{ characters} \]

We can compute for various n’s the probabilities of getting more evidence for group 1 than for group 2

A typical result is for \( n_1 = 10, n_2 = 8, n = 100 \):

<table>
<thead>
<tr>
<th></th>
<th>Bootstrap</th>
<th>Jackknife</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta = 1/2 )</td>
<td>( \delta = 1/e )</td>
<td></td>
</tr>
<tr>
<td>( \text{Prob}(m_1 &gt; m_2) )</td>
<td>0.6384</td>
<td>0.5923</td>
</tr>
<tr>
<td>( \text{Prob}(m_1 = m_2) )</td>
<td>( \frac{1}{2} \text{Prob}(m_1 = m_2) )</td>
<td>( \frac{1}{2} \text{Prob}(m_1 = m_2) )</td>
</tr>
<tr>
<td>( \text{Prob}(m_1 &gt; m_2) )</td>
<td>0.7230</td>
<td>0.7587</td>
</tr>
<tr>
<td>( \text{Prob}(m_1 = m_2) )</td>
<td>0.6807</td>
<td>0.6755</td>
</tr>
</tbody>
</table>
Parametric bootstrapping

The Parametric Bootstrap (Efron, 1985)

Suppose we have independent observations drawn from a known distribution:

\[
\begin{align*}
    x_1, x_2, x_3, \ldots, x_n & \rightarrow \theta
\end{align*}
\]

and a parameter, \( \theta \), calculated from this.

To infer the variability of \( \theta \):

Use the current estimate, \( \hat{\theta} \).

Use the distribution that has that as its true parameter.

Sample \( R \) data sets from that distribution, each having the same sample size as the original sample.

And take the distribution of the \( \hat{\theta}_i \) as the estimate of the distribution from which it is drawn.
The parametric bootstrap for phylogenies

original data → estimate of tree → computer simulation

- data set #1 → T₁
- data set #2 → T₂
- data set #3 → T₃
- data set #100 → T₁₀₀

estimation of tree
An example of the parametric bootstrap
Likelihood ratio confidence limits on Ts/Tn ratio

for the 14-species primate data set
Likelihoods in tree space – a 3-species clock example

Lecture 27. Phylogeny methods, part 7 (Bootstraps, etc.) – p.16/30
The constraints for a molecular clock

Constraints for a clock

\[ v_1 = v_2 \]
\[ v_4 = v_5 \]
\[ v_1 + v_6 = v_3 \]
\[ v_3 + v_7 = v_4 + v_8 \]
Testing for a molecular clock

To test for a molecular clock:

- Obtain the likelihood with no constraint of a molecular clock (For primates data with $T_s/T_n = 30$ we get $\ln L_1 = -2616.86$
- Obtain the highest likelihood for a tree which is constrained to have a molecular clock: $\ln L_0 = -2679.0$
- Look up $2(\ln L_1 - \ln L_0) = 2 \times 62.14 = 124.28$ on a $\chi^2$ distribution with $n - 2 = 12$ degrees of freedom (in this case the result is significant)
Goldman’s simulation test of Likelihood Ratios

Goldman (1993) suggests that, in cases where we may wonder whether the Likelihood Ratio Test statistic really has its desired $\chi^2$ distribution we can:

- Take our best estimate of the tree
- Simulate on it the evolution of data sets of the same size
- For each replicate, calculate the LRT statistic
- Use this as the distribution and see where the actual LRT value lies in it (e.g.: in the upper 5%?)

This, of course, is a parametric bootstrap.
Two trees to be tested by paired sites tests

Tree I

- Mouse
  - Bovine
    - Gibbon
      - Orang
        - Gorilla
          - Chimp
            - Human

Tree II

- Mouse
  - Bovine
    - Gibbon
      - Orang
        - Gorilla
          - Chimp
            - Human
Differences in log likelihoods site by site

<table>
<thead>
<tr>
<th>Tree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>231</th>
<th>232</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2.971</td>
<td>-4.483</td>
<td>-5.673</td>
<td>-5.883</td>
<td>-2.691</td>
<td>-8.003</td>
<td>...</td>
<td>-2.971</td>
<td>-2.691</td>
</tr>
<tr>
<td>II</td>
<td>2.983</td>
<td>-4.494</td>
<td>-5.685</td>
<td>-5.898</td>
<td>-2.700</td>
<td>-7.572</td>
<td>...</td>
<td>-2.987</td>
<td>-2.705</td>
</tr>
<tr>
<td>Diff</td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>-0.431</td>
<td>...</td>
<td>+0.012</td>
<td>+0.010</td>
</tr>
</tbody>
</table>
Histogram of log likelihood differences

Difference in log likelihood at site
Paired sites tests

- Winning sites test (Prager and Wilson, 1988). Do a sign test on the signs of the differences.
- $z$ test (me, 1993 in PHYLIP documentation). Assume differences are normal, do $z$ test of whether mean (hence sum) difference is significant.
- $t$ test. Swofford et. al., 1996: do a $t$ test (paired)
- Wilcoxon ranked sums test (Templeton, 1983).
- RELL test (Kishino and Hasegawa, 1989 per my suggestion). Bootstrap resample sites, get distribution of difference of totals.
In this example

- Winning sites test. 160 of 232 sites favor tree I. $P < 3.279 \times 10^{-9}$
- $z$ test. Difference of log-likelihood totals is 0.948104 standard deviations from 0, $P = 0.343077$. Not significant.
- $t$ test. Same as $z$ test for this large a number of sites.
- Wilcoxon ranked sums test. Rank sum is 4.82805 standard deviations below its expected value, $P = 0.000001378765$
- RELL test. 8,326 out of 10,000 samples have a positive sum, $P = 0.3348$ (two-sided)
Bayesian methods

In the Bayesian framework, one can avoid the separate calculation of confidence intervals. The posterior distribution of trees shows us how much credence to give different trees (for example, it assigns probabilities to different tree topologies). The unresolved issue is how to summarize this posterior distribution in the best way. In this respect Bayesian methods leave you in a situation analogous to having the cloud of bootstrap-sampled trees without yet having summarized them.
References


more references


more references


more references

Sanderson, M. J. 1995. Objections to bootstrapping phylogenies: a critique. *Systematic Biology* **44**: 299-320. [Good but he accepts a few criticisms I would not have accepted]


How it was done
This projection produced as a PDF, not a PowerPoint file, and viewed using the Full Screen mode (in the View menu of Adobe Acrobat Reader):

- using the prosper style in LaTeX,
- using Latex to make a .dvi file,
- using dvips to turn this into a Postscript file,
- using ps2pdf to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.