An annotated bibliography to frequentist topology testing

Mark T. Holder

July 26, 2011

Notation:

- M characters, sites, or columns;
- N leaves, taxa, or sequences;
- K plausible trees to be tested;
- B bootstrap replicates;
- δ difference in log-likelihood between trees
- $\ell(\theta_i, T_i|X)$ is the likelihood of the tree i and the numerical parameters θ_i given the data (proportional to $\Pr(X|\theta_i, T_i)$
- $\ln L(\theta_i, T_i|X)$ is the log-likelihood of the tree i and the numerical parameters θ_i given the data (proportional to $\Pr(X|\theta_i, T_i)$
- $\hat{\sigma}_i$ is the standard deviation of the log-likelihoods for θ_i, T_i.
- $\hat{\sigma}_i^{(j)}$ is the standard deviation of the log-likelihoods for θ_i, T_i computed on bootstrap replicate j.

Cavender (1978): Citation for Cavender part of CFN model. Defines rejection regions based on low probability of observing the data patterns. Focusses on regions of pattern count space (for four taxa) in which the parsimony score of the best tree is at least r steps better than the next best tree. Table 2 (page 277) shows the critical value of r needed to assure various α values. For $\alpha = 0.05$ you have the depressing:

M	3	4	10	15	20	30	40
r	2	3	5	7	9	13	15

Notes that Pr conflicting synapomorphy can be as high as 1/4 (and as high as the probability of the true synapomorphy). He notes that this test is not "locally most powerful" nor is it unbiased.

Points out that you can estimate the tree if you don't know the number of constant characters (p278; relevant to the Mkv model of (Lewis, 2001)).

Felsenstein (1985): Introduces the bootstrap to phylogenetics. p784 - use of columns as justifiable.
p786 - test of monophyly of a clade. Reject the alternatives if they occur is ¡5% of bootstrap estimates.
p786 - does not correct for inconsistency rather "a confidence interval within which is contained not the true phylogeny, but the phylogeny that would be estimated on repeated sampling"
p787 - multiple tests problem and need to correct for it.
p787 - difficulties caused by tree space.
p787 - delete 1/2 jackknife
p788 - use of pattern weights to implement bootstrapping.
p790 - perfect Hennigian data, "rule of three" and discussion of connection to clean data and random support around a trifurcation.

Gaut and Lewis (1995): Simulation study of ML inference and LRT in the four-taxon case. ML tree estimation is found to be robust to modeling errors. The authors compare the distribution of the LRT to the χ^2_1 and find that they do not match (they mention the possibility of using a mixture of χ^2, but also cite Thompson via pers comm to (**?**) that this may not be a problem in this context because the likelihood is continuous over the trifurcation). LRT using χ^2_1 would reject too often if the model is too simple.

Sanderson (1995): Extensive discussion of objections to bootstapping centered around independence of samples and difficulties with defining a universe of characters to sample from.

pp313-315 provide a nice discussion of objections to Neyman-Pearson hypothesis testing in general.

Huelsenbeck et al. (1996): Introduce a likelihood-ratio test for monophyly in which monophyly is the null. The test statistic is δ, the difference in log-likelihoods for the ML subject to the constraint of monophyly and the ML tree found under no constraints. Significance is assessed using the parametric bootstrap. They note that the parametric bootstrap is prone to overconfidence when the model used to generate the data is too simple.

Goldman et al. (2000): Very clear discussion of the selection bias problem in most uses of the KH-test and survey of the tests up to that point. They introduce a cumbersome scheme of mnemonics:

- a priori | posteriori selection of candidates
- NonParametric | Parametric
- full | partial | no optimization of parameters on bootstrap reps.
- centered | uncentered
- compare attained δ to null distribution: directly | using normality assumption | additional normal assumptions ($\hat{\delta}$ estimated from the sitewise $\delta(m)$ values) | stronger assumption that the sitewise $\delta(m)$ are normally distributed.

Pg 660 clear description of the SH test:

1. Calculate the test statistic for each tree: $\delta_i = \ln L(\hat{\theta}, \hat{T}|X) - \ln L(\hat{\theta}_i, T_i|X)$

2. for each bootstrap rep $1 \leq j \leq B$:
   (a) Calculate and store $t_i^{(j)} = \ln L(\hat{\theta}_i, T_i | X^{(j)})$
3. For each tree calculate the centering offset:

$$\bar{t_i} = (1/B) \sum_{j=1}^{B} t_i^{(j)}$$

4. For each tree i and stored bootstrap result j calculate: $c_i^{(j)} = t_i^{(j)} - \bar{t_i}$
5. For each bootstrap rep $1 \leq j \leq B$ find the largest likelihood relative to the mean for the tree: $c^{(j)} = \max \left[c_i^{(j)} \right]$ over all i
6. P-value for $T_i \approx$ the fraction of reps for which $c^{(j)} > \delta_i$

Point out (pp661) that if an incorrect (a posteriori) KH test is not significant, then the correct SH test won't reject either.

Berry et al. (2000): follows the decision-theoretic framework of Berry and Gascuel (1996) but uses a quartet-distance as a measure of distance from the true tree. A fully-resolved tree is constructed and then poorly supported branches are collapsed.

Billera et al. (2001): describe a geometric approach to tree-space that enables a metric distance that can be used to find centroid trees from a collection of trees (e.g. in summarizing bootstrap results).

Aris-Brosou (2003b): Discuss the fact that the plethora of tests and differing P-values often reflect differing nulls. Discusses a "natural" null is that the k trees are equidistance from the unknown, true distribution (μ); thus $\mathbb{E}_\mu(\ell(\theta_1, T_1|X) = \mathbb{E}_\mu\{\ell(\theta_K, T_K|X)\}$ for all trees in the null. Introduces two means of reducing the composite hypothesis over trees to a simple hypothesis. In his frequentist significance test (FST), the mean of the maximized log likelihoods is used as

$$h_{0,\tau}(X) = (1/K) \sum_{i=1}^{K} \ell(\hat{\theta}_i, T_i|X)$$

resulting in a test like the SH test.

In the frequentist hypothesis test (FHT) the ML tree is used to calculate $h_{0,\tau}(X)$; he argues that this make it similar to the bootstrap.

Argues that FST is geared toward identifying a set of "close" trees; while FHT tries to identify the true tree.

Analysis of real data points to underparameterization of the available models, and leads to a recommendation to adopt conservative tests. AU test showed sensitivity to model mis-specification but still found to be "safer" than other hypothesis tests.

In both cases, the algorithm then consists of the following:

1. For each tree: $\hat{v}_i = \ell(\theta_i, T_i|X) - h_{0,\tau}(X)$
2. Calculate the test statistic for each tree: $t_i = \hat{v}_i/\hat{\sigma}_i$
3. for each bootstrap rep j:
   (a) $s_i^{(j)} = \left(\hat{v}_i^{(j)} - \hat{v}_i\right)/\hat{\sigma}_i{}^{(j)}$
   (b) $\hat{s}^{(j)} = \min\left[s_i^{(j)}\right]$ over all $i \leq K$
4. P-value for $T_i \approx$ the fraction of reps for which $\hat{s}^{(j)} < t_i$

This is similar to SH, except the test-statistic rather than the resampling is centered.

Aris-Brosou (2003a): A Bayesian version of the methods introduced in Aris-Brosou (2003b) where the posterior is weight trees in the reduction from complex to simple hypothesis.

Erixon et al. (2003): Simulation study comparing BP to posterior probabilities under the same model. Find posteriors to be sensitive to model misspecification (high Type I error).

BP tended to be lower than posterior probabilities, and both were lower than the probability of reconstructing the tree given more data (but they did not sweep over branch length space).

Galtier (2004): analytical and simulation study of the effect on extreme violation of the independence assumption (including repeating data patterns and mimicking the effect of pairing across an RNA stem). Non-independence decrease accuracy of parsimony, but did not decrease BP. "Should we, therefore, worry about past and future interpretations of bootstrap scores? Probably not too much because the influence of nonindependence on the bootstrap score appears slight. The overestimation was always $< 10\%$ in Figure 1." Consecutive-site *vs* dispersed bootstrap procedures were compared and suggested as a means of testing for the effect of non-independence in real analyses.

Huelsenbeck and Rannala (2004): show that posteriors behave as expected when the true parameters are drawn from the prior. This paper focusses on Bayesian inference, but is relevant to the interpretation of other studies that examine whether or not BP can be used as Pr correct reconstruction.

Anisimova and Gascuel (2006): Approximate LRT for branches introduced. For testing a branch, δ is calculated by looking for best two ML scores among the three trees that resolve the polytomy created by collapsing the branch.

2δ shown to follow $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ when null is known *a priori*. But Bonferroni or "cubic" correction was needed to account for selection bias. For lower P-values, the Bonferroni correction is to conservative, and the cubic correction is:

$$\begin{aligned} P \approx 1 - (1-\alpha)^3 &\approx 3\alpha + \mathcal{O}(\alpha^2) \\ &\approx 1.5 - 1.5F_{\chi_1^2}(2\delta) \end{aligned}$$

where $F_{\chi_1^2}(2\delta)$ is the cumulative probability function for the χ_1^2 evaluated at the LRT (which is 2δ). This implies that if $\delta > 2.25$, then P-value will be < 0.05.

> "When the analysis model describes data well, the type I error rate obtained using Bonferroni corrected mixture distribution is close to the significance level α, so that the standard LRT remains accurate. Moreover, our results suggest that minor (but detectable) deviations from model assumptions do not significantly affect its accuracy. However, when important factors (e.g., transition/transversion ratio, rate variation among sites) are not accounted for, the test can become very inaccurate.(p545)"

They find 1-BP to be excessively liberal for some tree shapes. They attribute the difference to plotting Type-I error rates against α cutoffs conditionally on tree shape rather than Pr true branch vs cutoff binned by BP (which they find to differ based on simulation tree shape).

Svennblad et al. (2006): Detailed discussion of how the Bayesian interpretation of BP given by Efron et al. (1996) does not correspond to a Bayesian phylogenetic analysis in which the prior is placed on tree/model parameters (rather than a uniform prior over pattern frequency space). They also discuss how the set of *separately informative* data patterns is not the same for ML and Bayesian approaches (e.g. an `AACT` pattern would support 12|34 in a Bayesian context, even under JC, because of interactions between longer on which the changes can occur and the branch length priors).

Wrobel (2008): Review of topology testing including BP and Bayesian clade posteriors.

Susko (2009): Examining the star-tree (which should correspond to an lfc), Susko shows that (*contra* Efron et al., 1996) the BP is not first-order correct. Instead, it is conservative. The general approach is to look at the distribution of bootstrap scores from points generated at the star tree, and see how this distribution changes as M increases. If BP were first order correct then, for large M, the distribution should be uniform (such that $1 - BP$ is uniform and gives the correct P-value for any α). In fact, "BP as large as y arises less frequently than $y \times 100\%$ of the time." (p218). Susko attributes this to the relevant lfc being a at a point where three topologies come together. The AU test returns less biased P-values, but they still do not satisfy first-order correctness.

?: survey of branch support. Simulate under K80+covarion and HKY+Γ_4 models.

References

Anisimova, M. and Gascuel, O. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55(4):539–552.

Aris-Brosou, S. (2003a). How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics (Oxford, England)*, 19(5):618–624.

Aris-Brosou, S. (2003b). Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in the presence of conflicting signals under misspecified models. *Systematic Biology*, 52(6):781–793.

Berry, V. and Gascuel, O. (1996). On the Interpretation of Bootstrap Trees: Appropriate Threshold of Clade Selection and Induced Gain. *Molecular Biology and Evolution*, 13(7):999–1011.

Berry, V., Gascuel, O., and Caraux, G. (2000). Choosing the tree which actually best explains the data: another look at the bootstrap in phylogenetic reconstruction. *Computational Statistics & Data Analysis*, 32(3-4):273–283.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27:733–767.

Cavender, J. A. (1978). Taxonomy with confidence. *Mathematical Biosciences*, 40(3-4):271–280.

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, 93:13429–13434.

Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003). Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. *Systematic Biology*, 52(5):665–673.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791.

Galtier, N. (2004). Sampling Properties of the Bootstrap Support in Molecular Phylogeny: Influence of Nonindependence Among Sites. *Systematic Biology*, 53(1):38–46.

Gaut, B. and Lewis, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution*, 12(1):152–162.

Goldman, N., Anderson, J., and Rodrigo, A. (2000). Likelihood-Based Tests of Topologies in Phylogenetics. *Systematic Biology*, 49(4):652–670.

Huelsenbeck, J., Hillis, D., and Nielsen, R. (1996). A Likelihood-Ratio Test of Monophyly. *Systematic Biology*, 45(4):546.

Huelsenbeck, J. P. and Rannala, B. (2004). Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Systematic Biology*, 53(6):904–913.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Sanderson, M. J. (1995). Objections to Bootstrapping Phylogenies: A Critique. *Systematic Biology*, 44(3):299–320.

Susko, E. (2009). Bootstrap Support Is Not First-Order Correct. *Systematic Biology*, 58(2):211–223.

Svennblad, B., Erixon, P., Oxelman, B., and Britton, T. (2006). Fundamental Differences Between the Methods of Maximum Likelihood and Maximum Posterior Probability in Phylogenetics. *Systematic Biology*, 55(1):116–121.

Wrobel, B. (2008). Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. *Journal of Applied Genetics*, 49(1):49–67.