The Bootstrap

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

September 29, 2009

The Bootstrap: A brief history

- The bootstrap was introduced to the world by Brad Efron, chair of the Department of Statistics at Stanford University, in 1979.
- The bootstrap is one of the most widely used new method in statistics that was invented within the past 50 years.
- In a special issue of *Statistical Science* that celebrates the 25th anniversary of the bootstrap, Brad Efron uses its application to phylogenetics as one of a small number of examples to illustrate its use and importance.

The General Bootstrap Framework

- We have a sample $x_1, \ldots, x_n$ drawn from a distribution F from which we wish to estimate a parameter θ using a statistic $\hat{\theta} = T(x_1, \ldots, x_n)$. (We might think of θ as being the median of the distribution, for example, and $\hat{\theta} = T(x_1, \ldots, x_n)$ as the sample median.)
- If we wanted to compute the standard error of the estimate, we would ideally compute the standard deviation of $T(X_1, \ldots, X_n)$ where $X_i \sim \text{iid } F$.
- We could estimate this to any desired degree of accuracy by generating a large enough number (say B) of random samples $X_1, \ldots, X_n$, computing $\hat{\theta}_i = T(X_1, \ldots, X_n)$ for the ith such sample, and then computing the standard deviation of these estimates.

$$\sqrt{\frac{\sum_{i=1}^{B}(\hat{\theta}_i - \theta)^2}{B}}$$

The Key Idea

- Unfortunately, we cannot take multiple samples from F.
- However, our original sample $x_1, \ldots, x_n$ *is an estimate of the distribution* F.
- Instead of taking samples from F, we could sample from the estimated distribution $\hat{F}$ by sampling from our original sample *with replacement*.

The Procedure

- We sample n values $x_1^*, \ldots, x_n^*$ with replacement from $x_1, \ldots, x_n$.
- It is very likely that some of the original x values will be sampled multiple times and others will not be sampled at all.
- For each sample, compute the estimate of θ using the original statistic.
- The ith estimate is $\hat{\theta}_i^* = T(x_1^*, \ldots, x_n^*)$.
- Repeat this B times and compute the standard deviation of the bootstrap estimates around the estimate from the original sample.

$$\sqrt{\frac{\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta})^2}{B}}$$

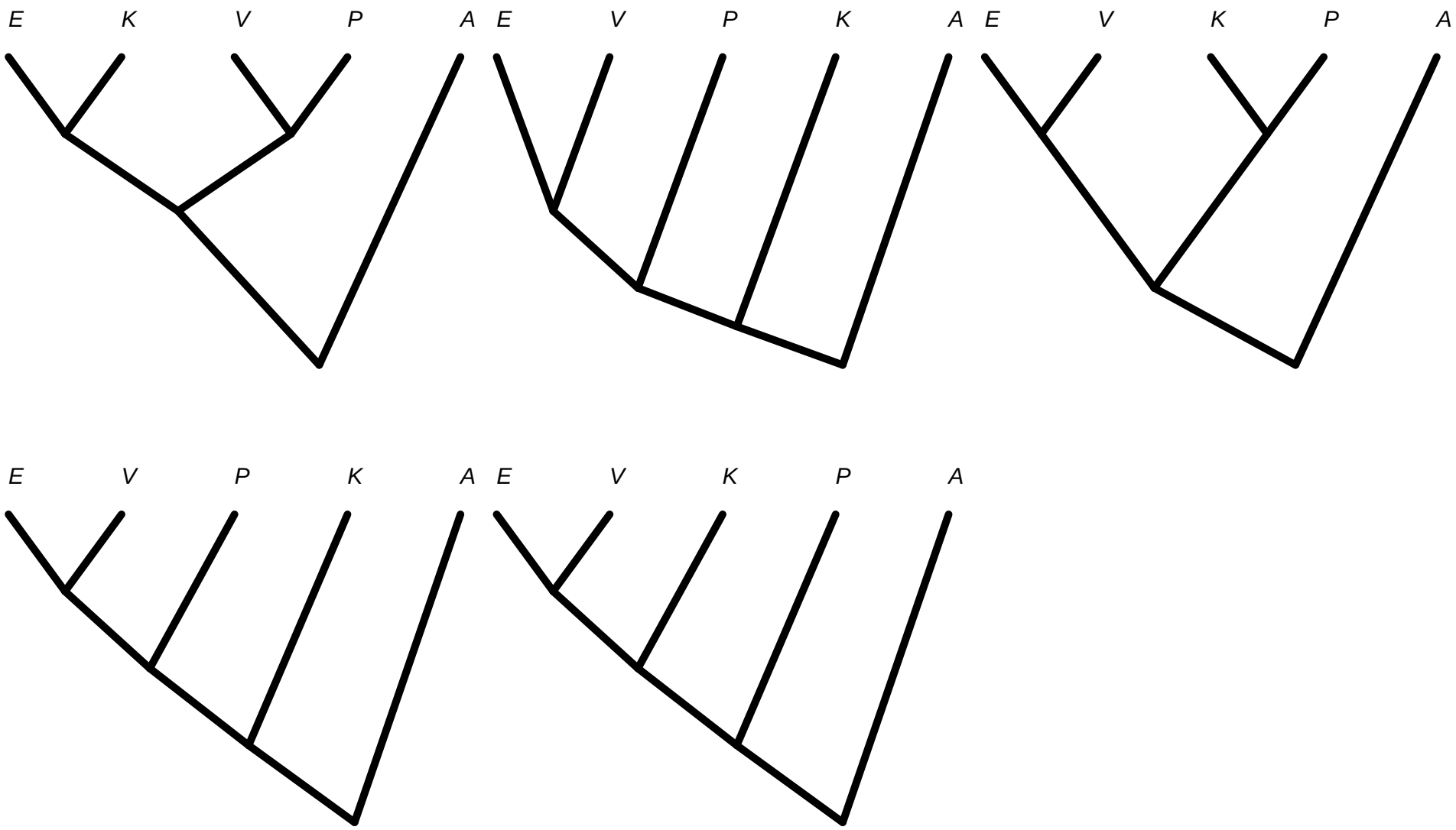
Why it works

- If the sampling distribution of the bootstrap sample estimate $\hat{\theta}^*$ around the estimate $\hat{\theta}$ is similar to the sampling distribution of the estimate $\hat{\theta}$ around the true value θ, then the boostrap standard error will be a good estimate of the real standard error.
- The bootstrap can be used to estimate bias, variance, for confidence intervals, and for hypothesis testing in many situations.
- It does depend critically on the assumption of independence of the original sample.

Consensus Trees

- A *strict consensus tree* shows only those clades that appear in every sampled tree.
- A *majority rule consensus tree* shows all clades that appear in more than half the sample of trees.
- (Notice that two clades that each appear in more than half the sampled trees must appear in at least one tree together, implying that they are compatible with one another.)
- A *priority consensus tree* adds clades to the majority rule consensus tree in order of decreasing frequency in the sample provided that these clades do not conflict with a clade with higher frequency.

Example



Dynamic Exploration of Tree Samples

- Show off Mark Derthick's **Summary Tree Explorer**.
- Software is free and available at `http://cityscape.inf.cs.cmu.edu/phylogeny/` .

Interpretation of Bootstrap Proportions

What does a bootstrap porportion mean? Let me count the ways.

- *Confidence* that the clade is in the true tree.
- Bayesian posterior probability that the clade is in the true tree.
- One minus p-value for a formal hypothesis test that the clade is in the true tree.
- Rough measure of method robustness.
- Measure of repeatability of the inferences for the method at hand.
- Others?