Bootstrapping Dependent Data in Ecology

Mark L Taper
Department of Ecology
310 Lewis Hall
Montana State University/ Bozeman
Bozeman, MT 59717
< taper@ rapid.msu.montana.edu>

Introduction

I have two goals for this report: First, I will discuss the use of bootstrap techniques in the analysis of dependent ecological data. And second, suggest how statisticians could promote the proper and effective use of the bootstrap in ecological analysis. Excellent reviews of the technical details of bootstrapping dependent data can be found in Davison and Hinkley (1997), the May 2003 issue of of Statistical Science (Volume 18, issue 2), which is dedicated to modern bootstrapping, and most recently Lahri (2003)

The concept of the bootstrap is well known in ecology. If there is sufficient data, the unknown distribution from which data is drawn can be approximated by the empirical distribution of the data. Redrawing data sets from the empirical distribution approximates redrawing samples from the underlying distribution. If a parameter can be estimated from the original data set then, in general, it can be estimated on a bootstrapped data set. An ensemble of bootstrap estimates can be developed, and functionals of this distribution used to calculate confidence intervals and other statistics of inferential interest.

The bootstrap has a series of well-known advantages. Most familiar of these is the bootstrap's non-parametric accommodation of unknown observation distributions, but the parametric bootstrap can be very useful when the observation distribution can be assumed. Whether non-parametric or parametric, the bootstrap will automatically estimate difficult to derive distributions of statistics.

Ecologists love the bootstrap. They believe that it is a straightforward and general replacement for classical statistics. In fact, just before I came to the 2004 Joint Statistical Meeting, a colleague asked 'Why talk about the bootstrap, isn't everything automatic?' My response was 'Pretty much if your data are independent, but not at all if your data has even moderate dependencies.' The bootstrap estimate of the variance of your statistic will be wrong (Lahiri 2003). Of course, even with independent data, if data distribution is highly skewed, then accuracy of confidence intervals can be improved by the use of calibration or bias

corrected and accelerated intervals (Effron and Tibshirani 1993).

Parametric versus non-parametric dependent bootstrap.

In the application of the bootstrap to dependent data, there is a great divide in ecology between the parametric and the non-parametric bootstrap. The parametric bootstrap constructs its resamples by randomly drawing data from specified parametric distributions, while the non-parametric bootstrap redraws from the empirical distribution.

Parametric dependent bootstrap

Ecology has a long history of simulation modeling. First in a proof of concept fashion, then in the 1980s a fad for null model simulation developed. This has been largely superseded by more complete statistical inference using parametric bootstrapping.

In 1994 Dennis and Taper used a parametric bootstrap likelihood ratio test for inference on the presence or absence of density dependence in ecological time-series of population sizes. In this case, analytic expressions for the likelihoods under competing hypotheses were derivable, but the distribution of the likelihood ratio test statistic was very poorly approximated by a Chi-square distribution. Fortunately, the distribution could be easily and accurately estimated with a parametric bootstrap, which repeatedly simulated time series under the null hypothesis and accumulated bootstrapped test statistics.

An example of the parametric bootstrap being involved both in parameter estimation and testing can be seen in Kelt et al. (1995). Here the task is to assess the impact of competition on the assembly of Chilean small mammal communities along a transect from the Patagonian stepe to the Chilean temperate rain forests. This problem is rife with dependencies. There are dependencies due to spatial location along the transect, due to local habitat type, and due to the process of faunal build up itself.

Entry into a community is considered to be a two-step process of first invasion and then establishment. Competition is assumed to occur within functional groups, and the presence in the community of species from the same functionall group as the invader reduces its chance of establishment. A functional group is a group of species that make their living in similar fashions. In this case the functional groups were herbivores, granivores, insectivores and omnivores. The probability of an invader being

from a given functional group is taken as the ratio of the number of species in the group available but not yet in the community to the total number of available species not yet in the community. Thus the probability of the j^{th} species to enter a community being from the i^{th} group can be written after some simplification as:

$$P\left(y_j = i \mid \mathbf{X}_{j\text{-}1}\right) = \frac{\left(n_i - \mathbf{X}_{i,j\text{-}1}\right)\left(1-\theta\right)^{\mathbf{X}_{i,j\text{-}1}}}{\sum_{i=1}^{k}\left[\left(n_i - \mathbf{X}_{i,j\text{-}1}\right)\left(1-\theta\right)^{\mathbf{X}_{i,j\text{-}1}}\right]}$$

where $\boldsymbol{n}$ is the species pool vector whose i^{th} element is the number of species available in the i^{th} group, $\mathbf{X_{i,j}}$ is the number of species in the community from the i^{th} group after the j^{th} successful invasion, and θ is the intra group competition parameter, which indicates how much the probability of success is reduced by each already established group member. A θ of 0 indicates that competition is not involved in community assembly. Given this expression the probability of any faunal buildup sequence can be calculated. Unfortunately, faunal buildup histories are unknown. All that is known is the extant species list for each community. However, there is residual information about faunal buildup in the existing functional group structure. Communities can be classified as favorable or unfavorable depending on whether species in the community are spread as evenly as possible among groups or not. Theoretically, the probability of observing a specific number of local communities in favorable states given a θ could be calculated exactly. Unfortunately, given the number of pool species, the number guilds, the number of communities, and the size of each community, combinatorics leads to a computational explosion that prohibits explicit evaluation. However, a parametric bootstrap can be used to effectively estimate the probability of specific numbers of favorable states occurring given specific values of θ. From this foundation, hypothesis testing, power analysis, and approximate maximum likelihood estimation can be developed.

Nonparametric dependent bootstrap

To get a better feel for the use in ecology of nonparametric bootstrap techniques for dependent data, I searched the Institute of Scientific Information's Web of Science with a key-word search the journal Ecology, one of the field of ecology's principal communication organs. I found no mention in title, abstract or key words of any paper of the key phrases "sieve

bootstrap", "block bootstrap", "moving block bootstrap" or "Markov bootstrap".

Startled by this, I expanded my search by removing constraints and searching only for papers mentioning "bootstrap" or "bootstrapping". I read deeply enough in all captured articles published in Ecology between 1985 and July 2004 to ascertain data structure and analysis methods. Only two articles (Elkinton et al. 1996, and Buonaccorsi et al. 2001) were located by this deeper survey that were missed in the previous.

Both papers describe a "sieve bootstrap" approach where the residuals from a model of interest are fit by an ARMA model. Entire sets of errors are built by sequentially drawing from the distributions specified by the ARMA and the points already drawn. Bootstrap resamples are constructed by adding these constructed error sets to the original fitted model.

One could argue that the "sieve bootstrap" is a parametric bootstrap because errors are drawn parametrically from the conditional distributions specified by the ARMA model. Nonetheless, there is a very different feel to the sieve bootstrap compared with the parametric bootstrap examples given above. The ARMA models are selected after the original fit so that they approximate the empirical dependent error distribution to a desired degree of accuracy.

Curiously, despite its common use in phylogenetic analysis (see Felsenstien 2003 for a review) a "block bootstrap" was not employed in any Ecology paper located by my search. The block bootstrap constructs a resampled data set by joining segments randomly drawn from the original data set. The size of these segments is chosen so as to capture local dependencies in the data.

Although the use of the block bootstrap *per se* was not found in my search of Ecology, article applying block jackknife techniques can be found. Lele et al (1998) use prediction error sum of squares from a block for model selection. They jackknifed year to year transitions in their spatio-temporal analysis of gypsy moth dynamics on Michigan's lower peninsula. After fitting the models to the jackknifed data, predictions were made for log population densities at all spatial locations for the eliminated year and squared error sums accumulated.

Kramer et al. (2001) construct confidence intervals for parameter estimates in a logistic regression of windthrow in Alaskan coastal temperate rainforest using a block jackknife. Because high spatial autocorrelation

was found up to 1500 m E-W and 3000 m N-S, a 3000x6000 meter rectangle of data centered on each prediction cell was deleted. The remaining data were used to estimate model coefficients and compute a probability of windthrow occurrence for that individual cell. The block was then centered on each forested cell that would result in non-overlapping blocks on the forested landscape in an iterative fashion. Ninety-five percent prediction and coefficient confidence intervals for both windthrown and non-windthrown forests were then calculated based on these results.

Conclusions

Although the basic bootstrap is quite popular, ecologists are not utilizing the suite of dependent data bootstrap techniques available to them, despite ecological problems rife with complex data dependencies. At least, that is the implication of this survey of a single but important journal. Bootstrapping dependent data appears to be *terra incognita* for ecologists and they are not going there without a guide. The two papers identified in report were both written by a single group of researchers, which included a professional statistician. Both papers rely heavily on Davison and Hinkley's 1997 book. I think this is indicative, ecologists, by and large, will not be reading the primary statistics literature, but will receive their statistical education from secondary sources such as books, textbooks and review articles. One of the main sources of information on the bootstrap for ecologists is Brian Manly's introductory book, Randomization, Bootstrap and Monte Carlo Methods in Biology (Manly 1997). One might even speculate that the dearth of application of the bootstrap to dependent ecological data stems from Manly's treatment of dependent data through means other than the bootstrap.

There is a need for a summary text, either a new volume directed towards ecologists or a review article in a major ecological journal. This "*Primer for the Bootstrapping of Dependent Ecological Data*" should: 1) Clearly define and differentiate the major methods available. 2) Delineate their domains of applicability. 3) Recognize the perils of each technique. 4) Describe practical algorithms. 4) List diagnostics that warn of problems. 5) And finally, fully integrate model selection/identification into the bootstrap based inference.

References

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK.

Dennis, B.; Taper, M.L. (1994) Density dependence in time series observations of natural populations: Estimation and testing. Ecological Monographs 64(2) 205-224.

Efron, B., and R. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman and Hall, London, UK.

Felsenstein, J. 2003. *Inferring Phylogenies*. Sinauer Associates, Sunderland

Kelt, D.A.; Taper, M.L.; and Meserve, P.L. 1995 Assessing the impact of competition on the assembly of communities: The biogeography of Chilean small mammals. Ecology 76:1283-1296.

Kramer, M.G., Hansen, A. , Kissinger, E. and Taper, M.L. 2001. Abiotic controls on windthrow and forest dynamics in a coastal temperate rainforest, Kuiu island, southeast Alaska. Ecology.82(10):2749-2768.

Lahiri, S. N. 2003. Resampling Methods for Dependent Data. Springer-Verlag, New York.

Lele, S; Taper, M.L. ; and Gage, S. 1998. Statistical analysis of population dynamics in space and time using estimating functions. Ecology 79:1489-1502.

Manly, B. F. J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology 2nd edition. Chapman and Hall. London, U.K.