The CAIC package: comparative methods using independent contrasts

David Orme

October 4, 2008

This document illustrates the use of the CAIC package for R (R Development Core Team, 2008) in carrying out a range of comparative analysis methods for phylogenetic data. The CAIC package reimplements the methods originally implemented in the Mac Classic programs CAIC (Purvis and Rambaut, 1995) and MacroCAIC Agapow and Isaac (2002) and this vignette draw heavily on the original manuals to those programs. The CAIC package, and the code in this vignette, requires the 'ape' package (Paradis et al., 2004).

1 Background

Comparing the traits of species (or of groups of species of higher taxonomic rank) can produce deep insights into evolutionary processes. However, all such analyses should take into account the degree to which species are related and hence do not provide independent data on a hypothesis. This can lead both to situations in which apparently strong relationships rest on relatively few truly independent events (Fig. 1a) or where strong relationships within groups are masked by phylogenetic differences between groups (Fig. 1b). One way around this problem, originally described by Felsenstein (1985), is to recognize that the differences between taxa on either side of a bifurcating node represent independent evolutionary trajectories and that these differences ('independent contrasts') can be used to test hypotheses in a way that accounts for the phylogenetic autocorrelation between the taxa. Pagel (1992) extended this method to permit contrasts to be calculated at polytomies.

2 Datasets

The example in this document will make use of the following artificially generated datasets. In addition to being used in these examples, these datasets were also analysed using the original programs and so provide a benchmark test for the re-implemenation

SmallTree A dataset containing two phylogenies and two data frames. One phylogeny (diTree) is a 15 tip, fully bifurcating tree; three nodes in this phylogeny have been collapsed to give a second topology (polyTree) containing polytomies (Fig. 2). The data frames (SmallTreeDat and SmallTreeDatNA) contain data for each of the tips in the tree. They contain

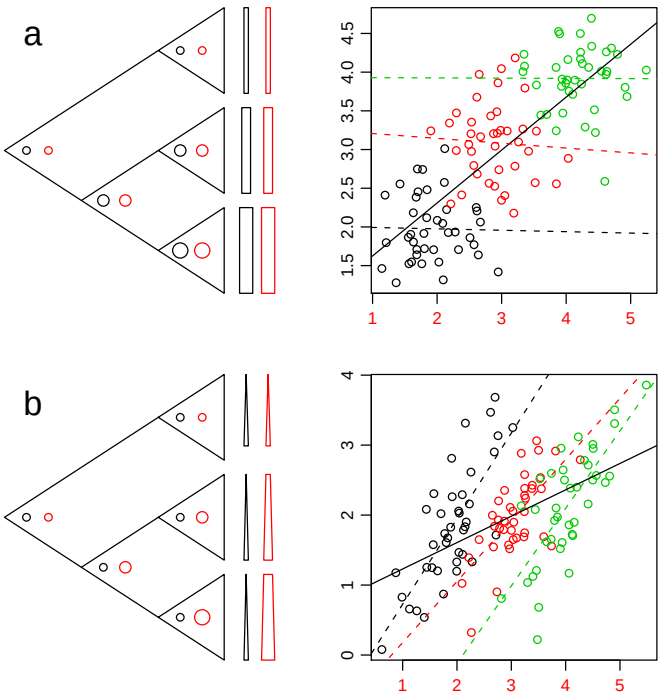

Figure 1: Phylogenetic autocorrelation in action. a) Simple regression (solid line) suggests a strong relationship between two variables; the phylogeny shows that within the three main groups, there is no consistent relationship (dashed lines) and the apparent relationship stems from only two linked shifts in the means of the traits early in the evolution of the clade. b) Simple regression (solid line) suggests a weak positive relationship between the two variables; the phylogeny shows that there are strong positive relationships between the traits within each of the three main groups but this is masked by early shifts in the mean value of the red trait.

a column of tip names (tip), two continuous variables (weight and othervar), a column of the number of species in the tip group (nSpp) and two categorical variables (catX2 and catX3). SmallTreeDatNA differs only in having two tips for which no data is available.

BigTree A dataset containing a larger phylogeny of 200 tip (BigTree) and a dataframe (BigTreeDat) providing four continuous variables (yv, xc1, xc2, xc3) and one binary categorical variable (xf) for each of the tips in the tree.

```
> library(ape)
> data(SmallTree, package = "CAIC")
> par(mfrow = c(2, 1))
> plot(diTree)
> plot(polyTree)
```

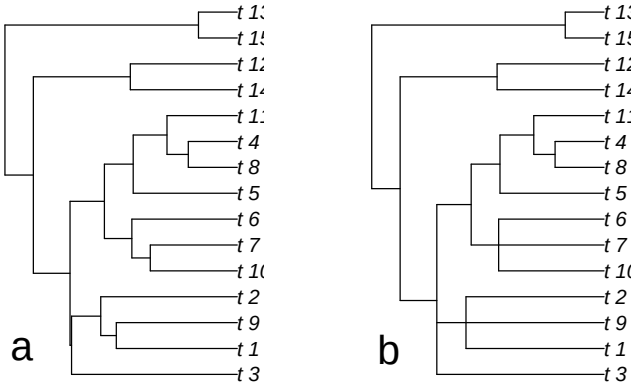

Figure 2: Phylogenies available in the smallTree dataset: a) diTree and b) polyTree.

3 Analyses using the CAIC package.

The CAIC package implements both CAIC and MacroCAIC analyses using a formula interface, as in the standard R `lm()` function. Both the `crunch()` and `macrocaic()` functions need the user to provide a phylogeny (`phy`) and a data frame (`data`) containing trait data on the tips of the phylogeny. One column from the data frame (`names.col`) is used to match data between the tips of the phylogeny and the data frame; the matching process is done automatically by both functions. The phylogeny must be provided in the `phylo` structure provided by the ape package, which provides functions to read trees from newick, nexus and ape format files. In both functions, a reference variable (`ref.var`) may also be provided. This is required in order to standardize the directions in

which contrasts are calculated in multivariate models. If no reference variable is provided, the function defaults to using the first explanatory variable specified in the model formula.

3.1 'Crunch' algorithm

3.2 'Brunch' algorithm

References

Paul-Michael Agapow and Nick J B Isaac. Macrocaic: revealing correlates of species richness by comparative analysis. *Diversity and Distributions*, 8:41–43, 2002.

Joseph Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125:1–15, 1985.

Mark D Pagel. A method for the analysis of comparative data. *Journal of Theoretical Biology*, 156(4):431–442, 1992.

Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

Andy Purvis and Andy Rambaut. Comparative analysis by independent contrasts (caic): an apple macintosh application for analysing comparative data. *Computer Applications In the Biosciences*, 11:247–251, 1995.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.