BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 7 2008, pages 1018–1020
doi:10.1093/bioinformatics/btn065

Phylogenetics

PCCA: a program for phylogenetic canonical correlation analysis

Liam J. Revell* and Alexis S. Harrison

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

Received on November 14, 2007; revised on January 7, 2008; accepted on February 15, 2008

Advance Access publication February 21, 2008

Associate Editor: Keith Crandall

ABSTRACT

Summary: PCCA (phylogenetic canonical correlation analysis) is a new program for canonical correlation analysis of multivariate, continuously valued data from biological species. Canonical correlation analysis is a technique in which derived variables are obtained from two sets of original variables whereby the correlations between corresponding derived variables are maximized. It is a very useful multivariate statistical method for the calculation and analysis of correlations between character sets. The program controls for species non-independence due to phylogenetic history and computes canonical coefficients, correlations and scores; and conducts hypothesis tests on the canonical correlations. It can also compute a multivariate version of Pagel's λ, which can then be used in the phylogenetic transformation.

Availability: PCCA is distributed as DOS/Windows, Mac OS X and Linux/Unix executables with a detailed program manual and is freely available on the World Wide Web at:
http://anolis.oeb.harvard.edu/~liam/programs/.

Contact: lrevell@fas.harvard.edu

1 INTRODUCTION

Multivariate analyses of phenotypic data from biological species are hampered by non-independence among the observations due to shared history (Felsenstein, 1985). Statistical methods, called phylogenetic comparative methods, have been developed specifically to deal with such phylogenetic non-independence. Here we introduce PCCA (phylogenetic canonical correlation analysis), a program for the canonical correlation analysis of multivariate, continuously valued character data obtained from species related by a phylogenetic tree. PCCA first corrects for phylogenetic non-independence using a phylogenetic generalized least squares (PGLS; Grafen, 1989) transformation, and then performs canonical correlation analysis on the transformed variates.

2 DESCRIPTION

2.1 Canonical correlation analysis

Canonical correlation analysis (CCA) is a procedure in which two sets of orthogonal derived variables are computed from two sets of original variables whereby the correlations between corresponding derived variables are maximized (Miles and Ricklefs, 1984). CCA first requires that the variables in the study are naturally divisible into two groups. A good example in an evolutionary or ecological study is the set of morphological variables for species, and the set of ecological or environmental variables for the same species (James and McCulloch, 1990). Linear combinations of each set of variables are then constructed to produce two sets of derived variables, each containing a number of variables equal to the smaller of the two numbers of variables in each of the original variable sets.

In particular, given two data matrices—an $n \times m_1$ matrix, $\mathbf{X}$ (for n species and m_1 variables in set one); and an $n \times m_2$ matrix, $\mathbf{Y}$ (for m_2 variables in set two)—vectors of coefficients, $\mathbf{a}_1$ and $\mathbf{b}_1$, are first computed so as to maximize the correlation, $\rho(\mathbf{u}_1,\mathbf{v}_1)$, between derived variables, $\mathbf{u}_1 = \mathbf{a}_1'\mathbf{X}$ and $\mathbf{v}_1 = \mathbf{b}_1'\mathbf{Y}$. The next pair of derived variables ($\mathbf{u}_2 = \mathbf{a}_2'\mathbf{X}$ and $\mathbf{v}_2 = \mathbf{b}_2'\mathbf{Y}$) are then computed according to the same procedure, but with the constraint that they are orthogonal with respect to the first derived variables. This procedure is repeated, with each new canonical axis orthogonal to all prior axes, $\min(m_1, m_2)$ times. Specific equations for CCA can be obtained from many standard multivariate statistical texts (e.g. Rencher, 2002).

CCA is a useful technique for the situation in which no single variable can serve as a measure of our character of interest. From a biological perspective, we might suspect that the morphologies of the species in our study are related to a set of environmental variables—but not sure how or in what manner. For example, in Harrison *et al.* (in preparation) we used phylogenetic CCA to examine the relationships between sets of environmental, behavioral and morphological variables in *Anolis* lizards.

2.2 Phylogenetic non-independence

In the analysis of evolutionary data, statistical non-independence among the observations for species related by a phylogenetic tree must be considered (Harvey and Pagel, 1991). This statistical dependence among species can be removed by way of several procedures. One such procedure is the PGLS approach of Grafen (1989). PGLS has the convenient property of having as a special case the most commonly used approach (phylogenetically independent contrasts; Felsenstein, 1985), given an assumption of constant rate Brownian motion (BM) as a model for the evolutionary process (Rohlf, 2001).

In the PGLS approach, we first calculate the $n \times n$, for n species, matrix, $\mathbf{C}$. $\mathbf{C}$ is proportional to the expected

*To whom correspondence should be addressed.

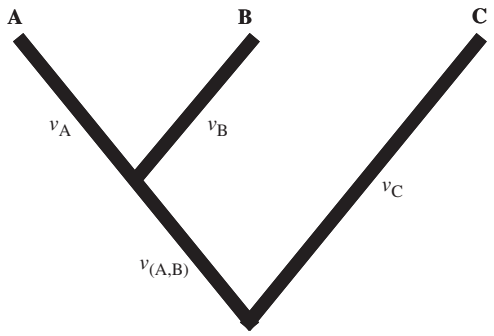

Fig. 1. A simple, three taxon phylogenetic tree with branch lengths.

variance-covariance matrix for the observations at the tips given our tree and evolutionary model (equivalent to $\mathbf{\Sigma}$ in Rohlf, 2001). For constant rate BM, the matrix, **C**, consists of elements C_{ij} equal to the summed branch lengths from the root of the tree to the common ancestor of species i and j. For the phylogeny in Figure 1, the expected variance-covariance matrix, **C**, is thus proportional to:

$$\mathbf{C} = \begin{bmatrix} v_A + v_{(A,B)} & v_{(A,B)} & 0 \\ v_{(A,B)} & v_B + v_{(A,B)} & 0 \\ 0 & 0 & v_C \end{bmatrix}$$

To remove statistical non-independence among the observations due to the phylogeny, we use the $n \times n$ matrix, **D**, obtained via singular value decomposition such that $\mathbf{DCD}' = \mathbf{I}$ (in which **I** is the identity matrix; Garland and Ives, 2000). We then compute data matrices:

$$\mathbf{W} = \mathbf{DX} - \mathbf{D1a}_x{}'$$

and

$$\mathbf{Z} = \mathbf{DY} - \mathbf{D1a}_y{}'$$

in which **W** and **Z** are the transformed variates for **X** and **Y**, respectively; **1** is an $n \times 1$ column vector of 1.0s; and $\mathbf{a}_x$ and $\mathbf{a}_y$ and are vectors of the 'phylogenetic means' for each trait in **X** and **Y**, respectively, estimated as: $\mathbf{a}_x = (\mathbf{1}'\mathbf{C}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\ \mathbf{C}^{-1}\mathbf{X})$ and $\mathbf{a}_y = (\mathbf{1}'\mathbf{C}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\ \mathbf{C}^{-1}\mathbf{Y})$ (Rohlf, 2001). This operation is analogous to transforming **X** and **Y** by the inverse square root of **C** and then recentering each variable on its phylogenetic mean for greater convenience in the CCA calculations. Resultant data matrices, **W** and **Z**, are expected to contain observations unfettered by covariances due to shared history and can hypothetically be subjected many standard statistical analysis, such as correlation or regression.

Pagel (1999) proposed that computing **C** as above may not appropriately describe the expected variances and covariances among the observations at the tips of the tree for a given trait. He suggested a parameter, λ, to be estimated using likelihood, by which the off-diagonal elements (the covariances) of **C** are scaled (Freckleton *et al*., 2002).

For multiple traits, λ is simultaneously optimized for all characters in both variable sets. The maximum likelihood estimate (MLE) for λ is obtained by maximizing the equation for the likelihood, which is based on the multivariate normal, and can be expressed as follows:

$$L = \frac{\exp[-(\mathbf{b}-\mathbf{a})'(\mathbf{R} \otimes \mathbf{C}_\lambda)^{-1}(\mathbf{b}-\mathbf{a})]}{\sqrt{(2\pi)^{n\cdot(m_1+m_2)} \cdot |\mathbf{R} \otimes \mathbf{C}_\lambda|}}$$

In this equation, **b** is an $n(m_1+m_2) \times 1$ columnarized data vector from **X** and **Y**; **a** is an $n(m_1+m_2) \times 1$ column vector of phylogenetic means (see above); **R** is an $(m_1+m_2) \times (m_1+m_2)$ matrix equal to the MLE of the evolutionary variance-covariance matrix for all m_1+m_2 traits, which can be computed analytically for given **X**, **Y**, and $\mathbf{C}_\lambda$ (Freckleton *et al*., 2002; Revell and Harmon, 2008); $\mathbf{C}_\lambda$ is the phylogenetic covariance matrix, described above, in which off-diagonal elements have been scaled by λ; and ⊗ indicates that a Kronecker product is calculated. The likelihood, L, is maximized by numerical methods. It is inevitable that a single, simultaneously optimized λ be used regardless of the MLEs for λ for each character calculated separately. This is because transforming each character by a separate λ would render multivariate analyses of the data invalid.

2.3 Program input and options

PCCA reads a Newick format, fully bifurcating phylogeny with branch lengths, and an input file with continuously varying phenotypic characters in two sets. Multifurcating nodes in the phylogenetic tree should be arbitrarily resolved with branches of zero length before analysis (Rohlf, 2001).

Prior to CCA, data can be transformed setting $\lambda = 1.0$ (i.e. assuming constant rate BM); setting $\lambda = 0.0$ (i.e. assuming no phylogenetic dependence) and finally, setting λ to its multivariate MLE. If the third option is selected, the program also evaluates the hypotheses that the MLE for λ is significantly more likely than $\lambda = 1.0$ and $\lambda = 0.0$. It should be noted that, given an ultrametric phylogeny, $\lambda = 0.0$ is equivalent to CCA without any phylogenetic transformation; and $\lambda = 1.0$ is equivalent to CCA performed on the phylogenetically independent contrasts (e.g. Losos, 1990).

2.4 Analysis and output

Following phylogenetic transformation, canonical weights, scores and correlations are calculated by standard means. The program returns an output containing: parameter estimates and statistics for λ; phylogenetically transformed variates; raw and standardized canonical coefficients for all canonical variates; canonical scores on each canonical axis; canonical correlations; Wilk's λ, χ^2, and the significance, $P(\chi^2)$, of each canonical correlation and, optionally, canonical structure coefficients (loadings). Canonical scores, of course, are in terms of the evolutionary differences among species, rather than in terms of the original species, and would have to be back-transformed for interpretation (Rohlf, 2001).

3 CONCLUSION

PCCA performs canonical correlation analysis after first controlling for the phylogenetic non-independence among the observations for biological species. Canonical correlation analysis is a useful multivariate method which has heretofore not been implemented in a phylogenetic comparative context.

This program should facilitate the collection and analysis of new multivariate, continuously varying morphological, behavioral and ecological data in the context of a phylogenetic tree.

ACKNOWLEDGEMENTS

L. Harmon kindly commented on this manuscript.

Conflict of Interest: none declared.

REFERENCES

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat*., **125**, 1–15.

Freckleton,R.P. *et al*. (2002) Phylogenetic analysis and comparative data: A test and review of evidence. *Am. Nat*., **160**, 712–726.

Garland,T. Jr. and Ives,A.R. (2000) Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat*., **155**, 346–364.

Grafen,A. (1989) The phylogenetic regression. *Philos. T. Roy. Soc. B*., **326**, 119–157.

Harvey,P.H. and Pagel,M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

James,F.C. and McCulloch,C.E. (1990) Multivariate analysis in ecology and systematics: Panacea or Pandora's box. *Annu. Rev. Ecol. Syst*., **21**, 129–166.

Losos,J.B. (1990) Ecomorphology, performance capability, and scaling of West Indian *Anolis* lizards: An evolutionary analysis. *Ecol. Monogr*., **60**, 369–388.

Miles,D.B. and Ricklefs,R.E. (1984) The correlation between ecology and morphology in deciduous forest passerine birds. *Ecology*, **65**, 1629–1640.

Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Rencher,A.C. (2002) *Methods of Multivariate Analysis*. 2nd edn. John Wiley & Sons, New York.

Revell,L.J. and Harmon,L.J. (2008) Testing quantitative genetic hypotheses about the evolutionary rate matrix for continuous characters. *Evol. Ecol. Res*., **10**, 311–331.

Rohlf,F.J. (2001) Comparative methods for the analysis of continuous variables: Geometric interpretations. *Evolution*, **55**, 2143–2160.