The adephylo package

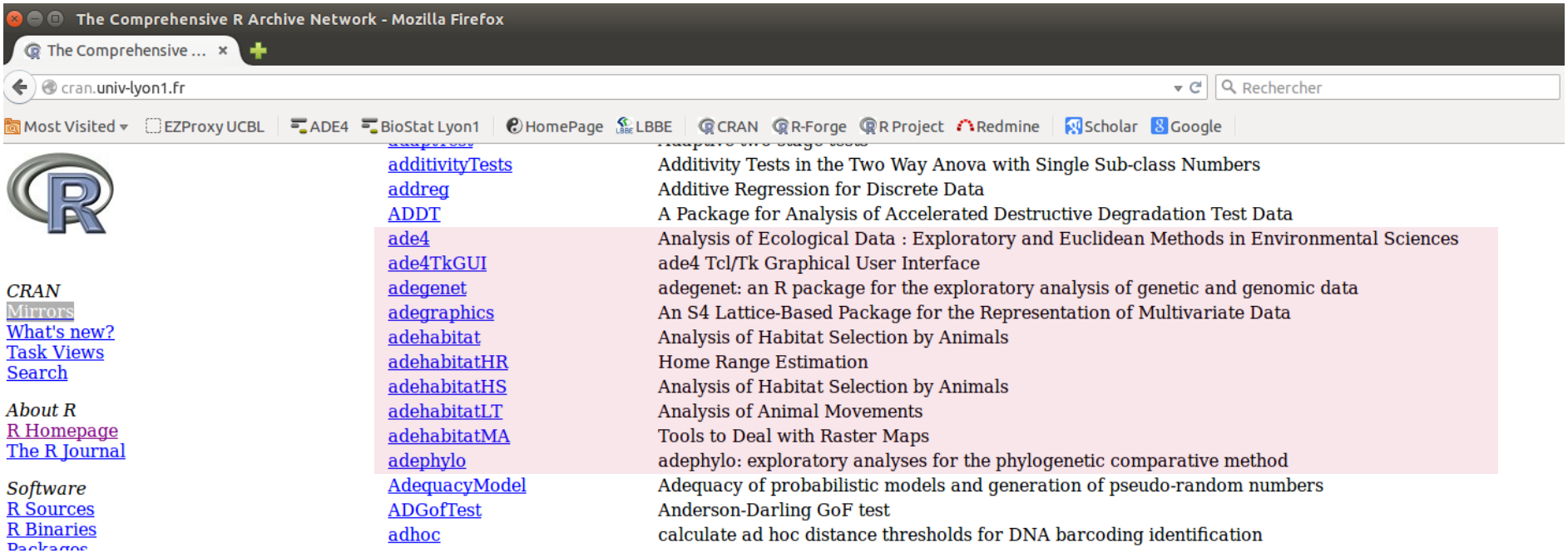
S. Dray

Univ. Lyon 1

2015, Lausanne

The ade family

analyse de **d**onnées **é**cologiques



an R package to analyse phylogenetic signal in traits data

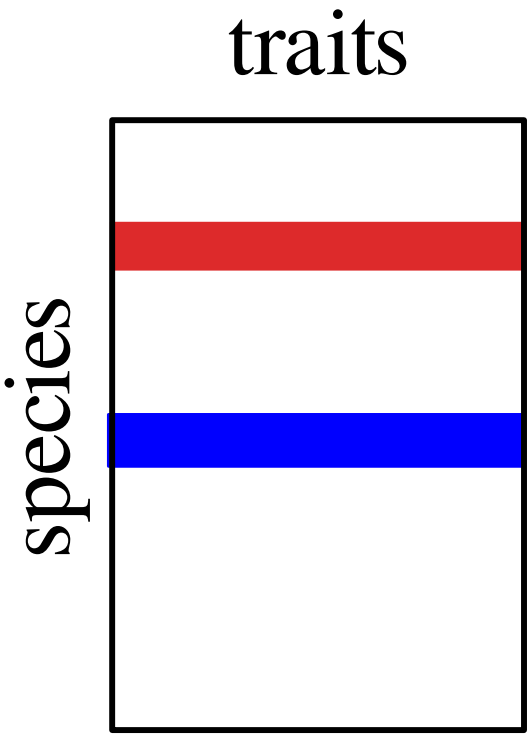
Reimplementation and development of `ade4` functionalities

- use of `phylo` (`ape`), `phylo4d` (`phylobase`) classes instead of `phylog`
- new methods (e.g., `ppca`) and functions

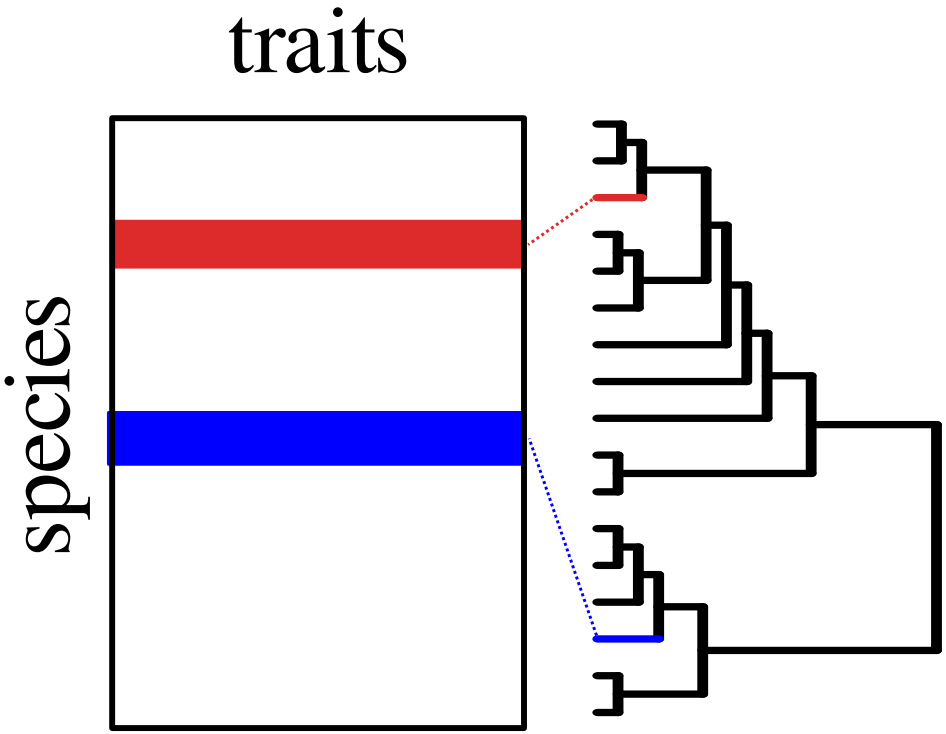
`ade4` $\rightarrow$ `adephylo` $\leftarrow$ `ape, phylobase`

multivariate phylogeny

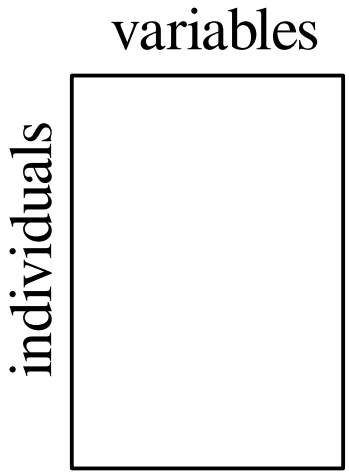
Two ingredients



Two ingredients

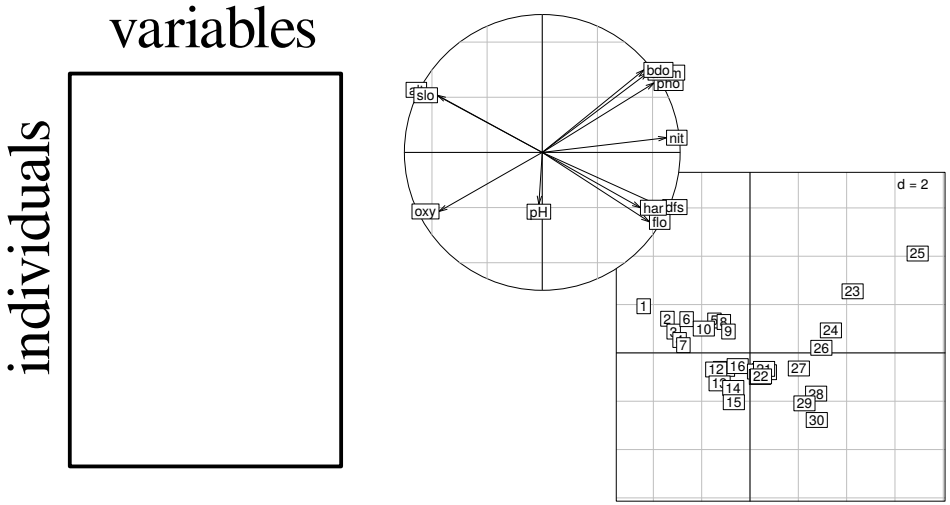


Summarizing data



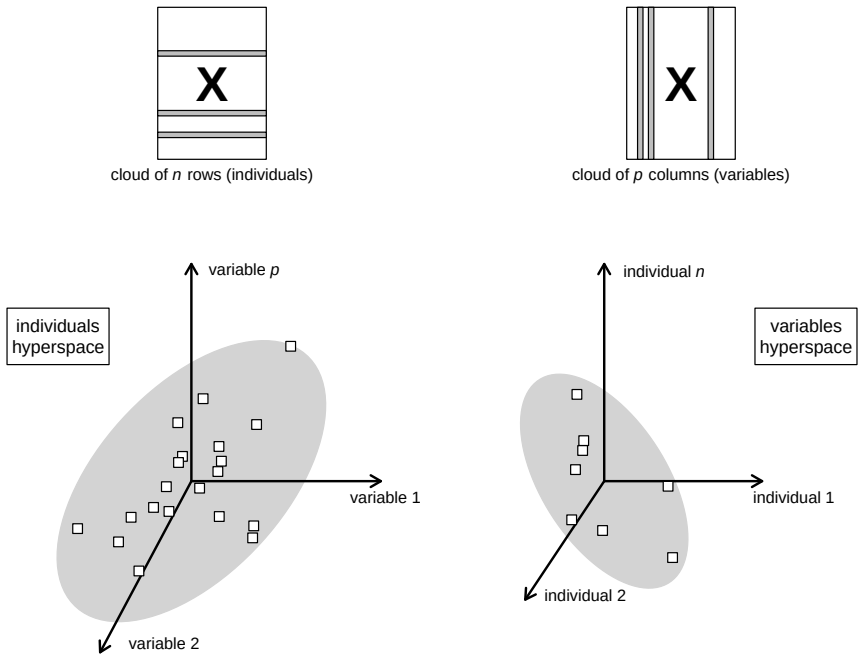
- what are the relationships between the variables ?
- what are the resemblances/differences between the individuals ?

Summarizing data with multivariate methods



- what are the relationships between the variables ?
- what are the resemblances/differences between the individuals ?

One table, two geometric viewpoints



Multivariate methods aim to answer these two questions and seek for small dimension hyperspaces (few axes) where the representations of individuals and variables are as close as possible to the original ones.

Principal Component Analysis

In `ade4` : `dudi.pca(df)`

- $\mathbf{X} = \left[\frac{x_{ij} - \bar{\mathbf{x}_j}}{s(\mathbf{x}_j)}\right]$
- $\mathbf{Q} = \mathbf{I}_p$
- $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$

Maximization of :

$$Q(\mathbf{a}) = \mathbf{a}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a} = \| \mathbf{XQa} \|^2_\mathbf{D} = var(\mathbf{XQa})$$

$$S(\mathbf{k}) = \mathbf{k}^\mathsf{T}\mathbf{D}^\mathsf{T}\mathbf{X}\mathbf{Q}\mathbf{X}^\mathsf{T}\mathbf{D}\mathbf{k} = \| \mathbf{X}^\mathsf{T}\mathbf{Dk} \|^2_\mathbf{Q} = \sum_{j=1}^{p} cor^2(\mathbf{k}, \mathbf{x}^j)$$

Lizards data

18 species, 8 traits :

- `mean.L` : mean adult female length (mm)
- `matur.L` : female length at maturity (mm)
- `max.L` : maximum length of adult female (mm)
- `hatch.L` : hatchling length (mm)
- `hatch.m` : hatchling mass (g)
- `clutch.S` : clutch size (n. eggs)
- `age.mat` : age at maturity (months)
- `clutch.F` : clutch frequency (n. per year)

Demo

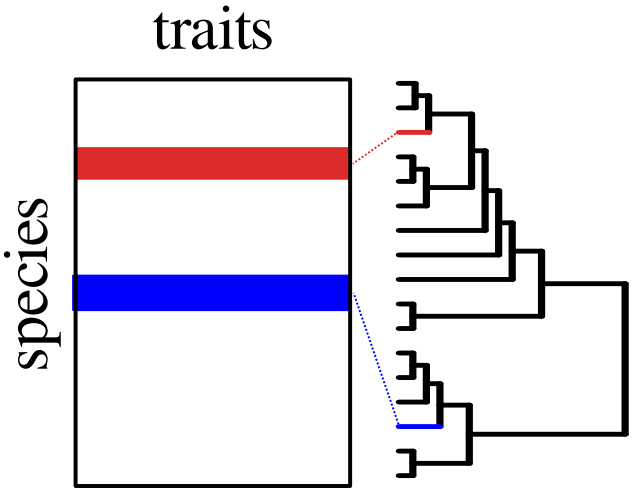
Bauwens, D. et R. Díaz-Uriarte. 1997. Covariation of life-history traits in Lacertid lizards : a comparative study. American Naturalist. 149 :91-111.

Management in R

Packages ape and `phylobase` provides functions, methods and classes to deal with phylogenetic data

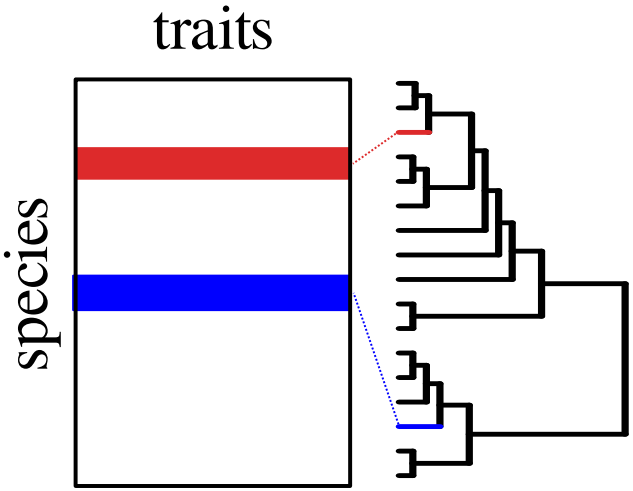
- Import : `read.tree`
- Classes for a tree : `phylo` (`ape`), `phylo4` (`phylobase`)
- Class for a tree + data : `phylo4d` (`phylobase`)
- Graphic : `plot`

Demo 



Phylogenetic structures (i.e. phylogenetic *autocorrelation* or signal) : the values of biological traits observed in a set of taxa are not independent from their position in the phylogenetic tree.

- *positive* : closely related taxa tend to share similar trait values
- *negative* : strong contrasts between sister taxa



Phylogenetic structures (i.e. phylogenetic *autocorrelation* or signal) : the values of biological traits observed in a set of taxa are not independent from their position in the phylogenetic tree.

- *positive* : closely related taxa tend to share similar trait values
- *negative* : strong contrasts between sister taxa

Need for mathematical representations of the phylogenetic relatedness

Phylogeny as a distance/similarity matrix

Function `distTips` computes distances. The argument `method` can take different values :

- `patristic` : patristic distance, i.e. sum of branch lengths on the shortest path between two tips
- `nNodes` : number of nodes on the shortest path between two tips
- `Abouheif` : Abouheif's distance
- `sumDD` : sum of the number of direct descendants of all nodes on the shortest path between two tips

Function `proxTips` returns phylogenetic proximities w_{ij} based on a phylogenetic distance d_{ij} using $w_{ij} = \frac{1}{d_{ij}^a}$

Moran's index

The n-by-1 vector $\mathbf{x} = [x_1 \cdots x_n]^\mathsf{T}$ contains the measurements of a quantitative trait for n species and $\mathbf{W} = [w_{ij}]$ is the the n-by-n phylogenetic proximity matrix.

$$MC(\mathbf{x}) = \frac{n \sum_{(i,j)} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{(i,j)} w_{ij} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

see `moran.idx`, `abouheif.moran`

Moran's index and Abouheif's C_{mean}

Abouheif's test of phylogenetic signal is exactly a test of Moran's index with phylogenetic proximities defined as :

$$w_{ij} = \frac{a_{ij}}{\sum_{j,i \neq j} a_{ij}}$$

with

$$a_{ij} = \left(\prod_{p \in P_{ij}} f(p) \right)^{-1}$$

where P_{ij} is the set of nodes on the shortest path from tip i to tip j and $f(p)$ is the number of direct descendents from node p.

Demo

Pavoine, S., Ollier, S., Pontier, D. and Chessel, D. 2008. Testing for phylogenetic signal in phenotypic traits : new matrices of phylogenetic proximities. Theoretical Population Biology, 73, 79–91.

- Moran's index allows to test the phylogenetic autocorrelation
- Phylogenetic structure is summarized by a single number
- Different stories can lead to the same value

- Moran's index allows to test the phylogenetic autocorrelation
- Phylogenetic structure is summarized by a single number
- Different stories can lead to the same value

Measuring $\rightarrow$ Describing

How the variance of a quantitative trait is decomposed along the phylogenetic tree ?

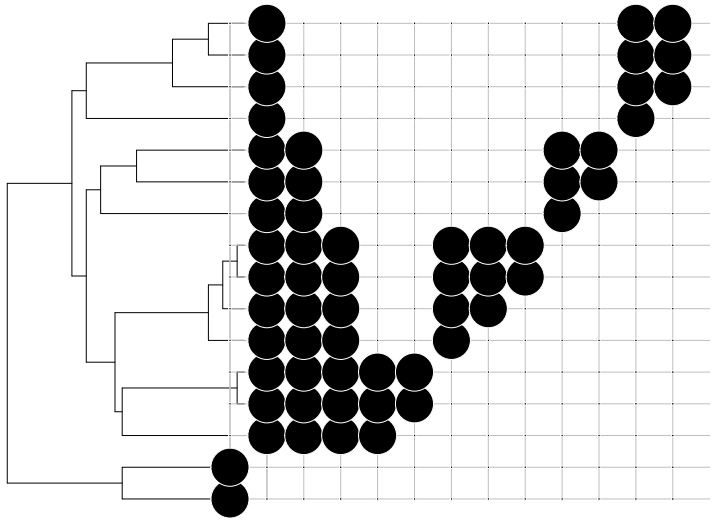
Phylogeny as an orthonormal basis

Tools to represent the structure of a tree. Orthonormal basis allows a simple and unique decomposition of the variance.

- Dummy variables
- Moran's eigenvectors

Dummy variables

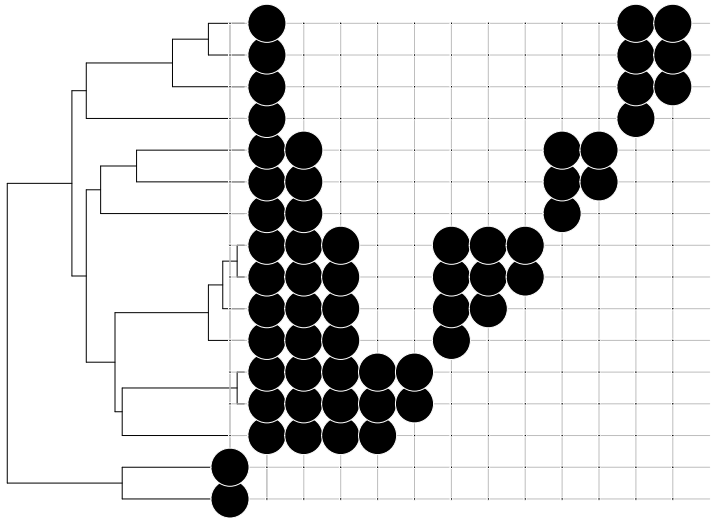
It defines partitions of tips reflecting the topology of the tree : each node (except the root) is translated into a dummy variable having one value for each tip (1 if the tip descends from this node and 0 otherwise).



Ollier, S., Chessel, D. and Couteron, P. 2005 Orthonormal Transform to Decompose the Variance of a Life-History Trait across a Phylogenetic Tree. Biometrics, 62, 471–477.

Dummy variables

It defines partitions of tips reflecting the topology of the tree : each node (except the root) is translated into a dummy variable having one value for each tip (1 if the tip descends from this node and 0 otherwise).



- Not an orthonormal basis
- Only based on the topology

Ollier, S., Chessel, D. and Couteron, P. 2005 Orthonormal Transform to Decompose the Variance of a Life-History Trait across a Phylogenetic Tree. Biometrics, 62, 471–477.

Moran's eigenvectors

The eigenvectors ($\mathbf{B}$) of a doubly centred matrix of phylogenetic proximities :

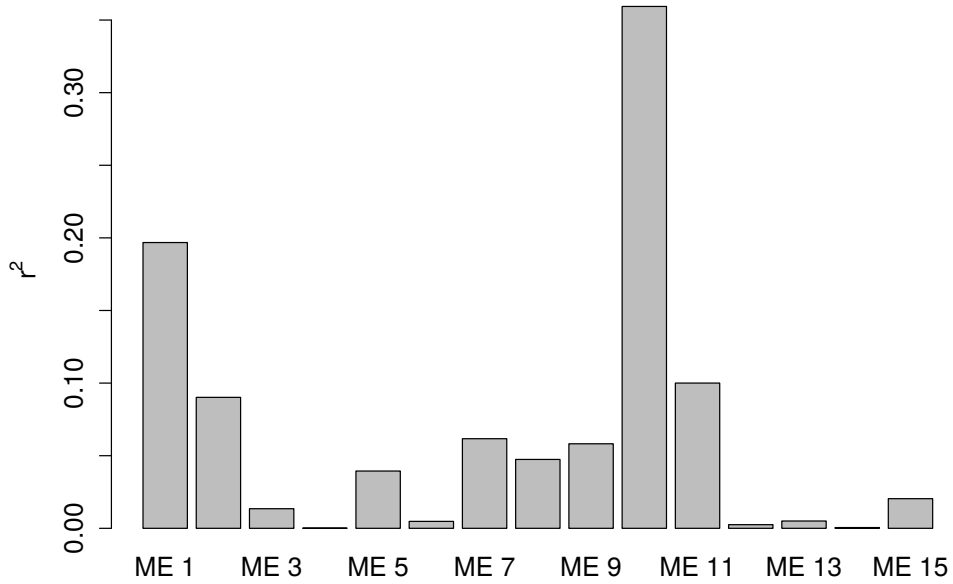
$$\mathbf{H}(\frac{1}{2}(\mathbf{W}^{\mathsf{T}} + \mathbf{W}))\mathbf{H}$$

where $\mathbf{H} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathsf{T}}/n$

The $n-1$ column-vectors of $\mathbf{B}$ (sorted by decreasing eigenvalue) are orthonormal variables ranging from the largest to the lowest possible phylogenetic autocorrelation as measured by Moran's index.

Decomposition of a trait on an orthonormal basis

The vector of squared correlation $[cor^2(\mathbf{x}, \mathbf{b}_1), \ldots, cor^2(\mathbf{x}, \mathbf{b}_{n-1})]$ provides a decomposition of a quantitative trait on the phylogeny.



Demo

Associated tests

The function `orthogram` provides different statistics for detecting phylogenetic signal :

- The maximum squared correlation :

$$\text{R2Max}(\mathbf{x}) = \max(r_1^2, \ldots, r_{n-1}^2)$$

- The deviation from an ordered uniform distribution (KS) :

$$D\text{max}(\mathbf{x}) = \max_{1 \leq m \leq n-1} \left(\sum_{i=1}^{m} r_i^2 - \frac{m}{n-1}\right)$$

- The skewness (to the root or to the tips) of the variance decomposition :

$$\text{SkR2k}(\mathbf{x}) = \sum_{i=1}^{n-1} i r_i^2$$

- The average local variation :

$$\text{SCE}(\mathbf{x}) = \sum_{i=2}^{n-1} (r_i^2 - r_{i-1}^2)^2$$

From univariate to multivariate data

Phylogenetic tools are mainly adapted to univariate data $\rightarrow$ indirect approach :

- summarize multivariate data by PCA
- apply phylogenetic analysis on PCA scores

Demo 

From univariate to multivariate data

Phylogenetic tools are mainly adapted to univariate data $\rightarrow$ indirect approach :

- summarize multivariate data by PCA
- apply phylogenetic analysis on PCA scores

Not optimal as PCA identifies the main resemblances/differences between the individuals but these differences are not constrained by the phylogenetic relatedness

Demo 

From PCA to phylogenetic PCA

pPCA is an extension of PCA that includes the matrix of phylogenetic proximities in the algorithm. It modifies the criteria maximized by the analysis

- PCA maximizes

$$Q(\mathbf{a}) = \mathbf{a}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a} = var(\mathbf{X}\mathbf{Q}\mathbf{a})$$

- pPCA maximizes

$$Q(\mathbf{a}) = \mathbf{a}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{X}^\mathsf{T}\frac{1}{2}(\mathbf{W}^\mathsf{T}\mathbf{D}^\mathsf{T} + \mathbf{D}\mathbf{W})\mathbf{X}\mathbf{Q}\mathbf{a} = var(\mathbf{X}\mathbf{Q}\mathbf{a}) \cdot MC(\mathbf{X}\mathbf{Q}\mathbf{a})$$

Jombart, T., Pavoine, S., Devillard, S., and Pontier, D. 2010. Putting phylogeny into the analysis of biological traits : A methodological approach. Journal of Theoretical Biology, 264(3), 693–701.

Demo

```
vignette("adephylo")
```

adephylo: exploratory analyses for the phylogenetic comparative method

Thibaut Jombart and Stéphane Dray

August 10, 2015

Contents

1 Introduction 1

2 First steps 2
2.1 Data reproduction: why we are not interesting the tree 2
2.2 Installing the packages 3
2.3 Loading data 3
2.4 Putting data into shape 4
2.4.1 Making a phylo object 4
2.4.2 Making a phylo4 object 5

3 Exploratory data analysis 6
3.1 Graphical methods exploring phylogenetic signal 6
3.1.1 Moran's I 7
3.1.2 Geary's C test 9
3.1.3 Phylogenetic decomposition of trait variation 9
3.2 Multivariate methods 11
3.2.1 Using orthonormal bases 11
3.2.2 Autocorrelation models 17
3.3 Using multivariate analyses 17

1 Introduction

This document describes the adephylo package for the R software. adephylo provides exploratory methods for the analysis of phylogenetic comparative data, i.e. biological traits measured for taxa whose phylogeny is also provided. The package extends and explores implementations of phylogenetic methods in the ade4 package (http://pbil.univ-lyon1.fr/ADE-4/home.php?lang=en).

Procedures implemented in adephylo rely on exploratory data analysis. They include data visualization and manipulation. Users for phylogenetic

1

R-sig-phylo -- R SIG on phylogenetic and comparative methods and analyses