Package ‘Rphylopars’

June 1, 2015

Version 0.1.0

Title Phylogenetic Comparative Tools for Missing Data and Within-Species Variation

Author Eric W. Goolsby, Jorn Bruggeman, Cecile Ane

Maintainer Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

Description Tools for performing phylogenetic comparative methods for datasets with with multiple observations per species (intraspecific variation or measurement error) and/or missing data. Available functions allow ancestral state reconstruction, missing data imputation, phylogenetic regression (phylogenetic generalized least squares), and phylogenetic principal components analysis (pPCA).

Date 2015-06-01

Depends ape

License GPL (>= 2)

Imports Rcpp, doBy, geiger, phylolm, phytools, mvnmle

LinkingTo Rcpp, RcppArmadillo

NeedsCompilation yes

Archs i386, x64

R topics documented:

`Rphylopars-package` *Phylogenetic Comparative Tools for Missing Data and Within-Species Variation*

Description

Tools for performing phylogenetic comparative methods for datasets with with multiple observations per species (intraspecific variation or measurement error) and/or missing data. Available functions allow ancestral state reconstruction, missing data imputation, phylogenetic regression (phylogenetic generalized least squares), and phylogenetic principal components analysis (pPCA).

Details

Package:	Rphylopars
Type:	Package
Version:	0.1.0
Date:	2015-06-01
License:	GPL (>= 2)

Author(s)

Eric W. Goolsby, Jorn Bruggeman, Cecile Ane

Maintainer: Eric W. Goolsby <`eric.goolsby.evolution@gmail.com`>

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Ho, L. S. T. and Ane, C. 2014. "A linear-time algorithm for Gaussian and non-Gaussian trait evolution models". Systematic Biology 63(3):397-408.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)
```

```
PPE

# perform phylogenetic PCA
PPE_pca <- phylopars.pca(PPE)
summary(PPE_pca)

# perform phylogenetic regression
p_BM <- phylopars.lm(V3~V1+V2,trait_data = sim_data$trait_data,tree = sim_data$tree,model = "BM")

# perform ancestral state reconstruction / tip prediction
PPP_ace <- phylopars.predict(PPE)
```

`anova.phylopars.lm` *Phylopars regression ANOVA*

Description

Generic S3 method for phylopars

Usage

```
## S3 method for class 'phylopars.lm'
anova(object, ...)
```

Arguments

`object`	Fitted phylopars.lm object
`...`	

`logLik.phylopars` *Extract Log_likelihood*

Description

Generic S3 method for phylopars

Usage

```
## S3 method for class 'phylopars'
logLik(object, ...)
```

Arguments

`object`	Fitted phylopars object
`...`	

`logLik.phylopars.lm` *Extract Log_likelihood*

Description

Generic S3 method for phylopars

Usage

```
## S3 method for class 'phylopars.lm'
logLik(object, ...)
```

Arguments

`object`	Fitted phylopars.lm object
`...`	

`phylopars` *Estimation of phylogenetic and phenotypic covariance parameters*

Description

This function estimates parameters for the phylogenetic and phenotypic variance-covariance matrices for datasets with missing observations and multiple within-species observations. This function is a prerequisite for ancestral state reconstruction and missing tip estimation (`phylopars.predict`) and phylogenetic PCA (`phylopars.pca`).

Usage

```
phylopars(trait_data, tree, model = "BM", pheno_error = TRUE,
phylo_correlated = TRUE, pheno_correlated = FALSE, calc_pheno = FALSE,
calc_pheno_auto_n = 20, calc_pheno_phenocovs, use_means=FALSE,
species_identifier = "species", verbose = FALSE, phylocov_start,
phenocov_start, theta_start, model_start, skip_optim = FALSE, REML = TRUE,
optim_limit = 50, BM_first = TRUE, usezscores = TRUE)
```

Arguments

`trait_data`	A data frame with one column per trait, as well as a column labeled "species" (with species names matching tips on the phylogeny). Each row corresponds to a single observation, and multiple observation for species are allowed. Missing data should be represented with NA.
`tree`	An object of class phylo
`model`	Model of evolution. Default is "BM". Also accepts "lambda", "kappa", "delta", "EB", and "OUrandomRoot", "OUfixedRoot",
`pheno_error`	If TRUE (default), parameters are estimated assuming within-species variation.
`phylo_correlated`	If TRUE (default), parameters are estimated assuming traits are correlated.

`pheno_correlated`	If TRUE, parameters are estimated assuming within-species observations traits are correlated (default is FALSE).
`calc_pheno`	If TRUE, phenotypic error variance is estimated for each species. If the number of observations for a given species is less than calc_pheno_auto_n (default=20), pooled within-species variance is used, unless phenotypic covariance matrices are supplied in the calc_pheno_phenocovs argument.
`calc_pheno_auto_n`	The minimum number of within-species observations for a given species for which within-species variance is calculated individually for a given species rather than pooled within-species variance. Only relevant if calc_pheno=TRUE and calc_pheno_phenocovs is not supplied.
`calc_pheno_phenocovs`	User-supplied phenotypic covariance matrix OR a list of phenotypic covariance matrices for each species.
`use_means`	Whether to use means (as in Ives et al. 2007) or raw data. Default is raw data (use_means=FALSE).
`species_identifier`	Optional argument. The name of the column in `data` where species names are stored. The default value is "species", but can be changed for convenience if `data` has an alternative species column name.
`verbose`	Optional argument. Whether or not to print the log-likelihood during optimization (NOTE: constants in the likelihood calculation are dropped from the value reported during optimization.)
`phylocov_start`	Optional starting value for phylogenetic trait variance-covariance matrix. Must be of dimensions n_traits by n_traits.
`phenocov_start`	Optional starting value for phenotypic trait variance-covariance matrix. Must be of dimensions n_traits by n_traits.
`theta_start`	Optional starting parameters for `theta`.
`model_start`	Optional starting parameters for the evolutionary model.
`skip_optim`	Optional argument to skip optimization (NOTE: this will produce the WRONG answer unless BM=TRUE, there is no missing data, and pheno_error=FALSE; generally leave set to FALSE).
`REML`	If TRUE (default), the algorithm will return REML estimates. If FALSE, maximum likelihood estimates will be returned.
`optim_limit`	Optional: number of extra optimization attempts allowed if convergence is not achieved. Only necessary if convergence failure is suspected.
`BM_first`	Whether to estimate covariance assuming Brownian motion prior to estimating alternative evolutionary model paramters.
`usezscores`	Whether to convert data to z-scores during optimization (recommended).

Value

An object of class `phylopars`.

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`, Cecile Ane, Jorn Bruggeman

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

PPE

# perform phylogenetic PCA
PPE_pca <- phylopars.pca(PPE)
summary(PPE_pca)

# perform phylogenetic regression
p_BM <- phylopars.lm(V3~V1+V2,trait_data = sim_data$trait_data,tree = sim_data$tree,model = "BM")

# perform ancestral state reconstruction / tip prediction
PPP_ace <- phylopars.predict(PPE)
```

`phylopars.crossvalidate` *Phylopars cross-validation*

Description

This function performs cross-validation of the fitted phylopars object using the leave-one-out method. For each species, all observations for a single trait are temporarily deleted, and the value is phylognetically imputed using phylopars.predict(). A regression between the observed and predicted species means is returned. Also estimated is mean bias and error of the phylopars model, as well as two alternative reference models: mean model and nearest neighbor model.

Usage

```
phylopars.crossvalidate(PPE, plot = FALSE, verbose = FALSE)
```

Arguments

`PPE`	Fitted `phylopars` object
`plot`	Whether or not to plot cross-validation statistics (default is TRUE)
`verbose`	Whether or not to print progress of cross-validation (default is FALSE)

Value

Returns a summary of cross-validation statistics.

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

PPE

phylopars.crossvalidate(PPE)
```

`phylopars.lm` *Phylopars regression*

Description

Performs phylogenetic regression.

Usage

```
phylopars.lm(formula, trait_data, tree, model = "BM", pheno_error = TRUE,
phylo_correlated = TRUE, pheno_correlated = FALSE, calc_pheno = FALSE,
calc_pheno_auto_n = 20, calc_pheno_phenocovs, use_means = FALSE,
species_identifier = "species", verbose = FALSE, phylocov_start,
phenocov_start, theta_start, model_start, skip_optim = FALSE,
REML = TRUE, optim_limit = 50, BM_first = TRUE, usezscores = TRUE)
```

Arguments

`formula`	A model formula.
`trait_data`	A data frame with one column per trait, as well as a column labeled "species" (with species names matching tips on the phylogeny). Each row corresponds to a single observation, and multiple observation for species are allowed. Missing data should be represented with NA.

`tree`	An object of class `phylo`
`model`	Model of evolution. Default is "BM". Also accepts "lambda", "kappa", "delta", "EB", and "OUrandomRoot", "OUfixedRoot",
`pheno_error`	If TRUE (default), parameters are estimated assuming within-species variation.
`phylo_correlated`	If TRUE (default), parameters are estimated assuming traits are correlated.
`pheno_correlated`	If TRUE, parameters are estimated assuming within-species observations traits are correlated (default is FALSE).
`calc_pheno`	If TRUE, phenotypic error variance is estimated for each species. If the number of observations for a given species is less than calc_pheno_auto_n (default=20), pooled within-species variance is used, unless phenotypic covariance matrices are supplied in the calc_pheno_phenocovs argument.
`calc_pheno_auto_n`	The minimum number of within-species observations for a given species for which within-species variance is calculated individually for a given species rather than pooled within-species variance. Only relevant if calc_pheno=TRUE and calc_pheno_phenocovs is not supplied.
`calc_pheno_phenocovs`	User-supplied phenotypic covariance matrix OR a list of phenotypic covariance matrices for each species.
`use_means`	Whether to use means (as in Ives et al. 2007) or raw data. Default is raw data (use_means=FALSE).
`species_identifier`	Optional argument. The name of the column in `data` where species names are stored. The default value is "species", but can be changed for convenience if `data` has an alternative species column name.
`verbose`	Optional argument. Whether or not to print the log-likelihood during optimization (NOTE: constants in the likelihood calculation are dropped from the value reported during optimization.)
`phylocov_start`	Optional starting value for phylogenetic trait variance-covariance matrix. Must be of dimensions n_traits by n_traits.
`phenocov_start`	Optional starting value for phenotypic trait variance-covariance matrix. Must be of dimensions n_traits by n_traits.
`theta_start`	Optional starting parameters for `theta`.
`model_start`	Optional starting parameters for the evolutionary model.
`skip_optim`	Optional argument to skip optimization (NOTE: this will produce the WRONG answer unless BM=TRUE, there is no missing data, and pheno_error=FALSE; generally leave set to FALSE).
`REML`	If TRUE (default), the algorithm will return REML estimates. If FALSE, maximum likelihood estimates will be returned.
`optim_limit`	Optional: number of extra optimization attempts allowed if convergence is not achieved. Only necessary if convergence failure is suspected.
`BM_first`	Whether to estimate covariance assuming Brownian motion prior to estimating alternative evolutionary model paramters.
`usezscores`	Whether to convert data to z-scores during optimization (recommended).

Value

A fitted phylopars.lm object.

Examples

```
# no missing data
s <- simtraits(ntaxa = 50,ntraits = 3,nreps = 1,nmissing = 0)
p_BM <- phylopars.lm(V3~V1+V2,trait_data = s$trait_data,tree = s$tree,model = "BM")
p_lambda <- phylopars.lm(V3~V1+V2,trait_data = s$trait_data,tree = s$tree,model = "lambda")
summary(p_BM)
summary(p_lambda)

# with missing data
# notice that logLik, AIC, and BIC are not reported if data is missing
s <- simtraits(ntaxa = 50,ntraits = 3,nreps = 1,nmissing = 15)
p_BM <- phylopars.lm(V3~V1+V2,trait_data = s$trait_data,tree = s$tree,model = "BM")
p_lambda <- phylopars.lm(V3~V1+V2,trait_data = s$trait_data,tree = s$tree,model = "lambda")
summary(p_BM)
summary(p_lambda)
```

`phylopars.pca` *Phylogenetic principal components analysis*

Description

This function, adapted from `phyl.pca` in the **phytools** package, performs phylogenetic principal components analysis on a fitted `phylopars` object. Additionally, fast algorithms are used to avoid large matrix inversions using the `three.point.compute` function in **phylolm**.

Usage

```
phylopars.pca(PPE, mode = "cov", pgls_means = FALSE)
```

Arguments

`PPE`	Fitted phylopars object
`mode`	Either "cov" (default) or "corr"
`pgls_means`	Whether to use raw species trait means of PGLS species trait means in forming the PCA

Value

Returns an object of class `phyl.pca`.

Author(s)

Eric W. Goolsby <`eric.goolsby.evolution@gmail.com`>, Cecile Ane, Jorn Bruggeman

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Revell, L. J. (2009) Size-correction and principal components for interspecific comparative studies. *Evolution*, 63, 3258-3268.

Ho, L. S. T. and Ane, C. (2014) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, 63: 397-408.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

PPE

# perform phylogenetic PCA
PPE_pca <- phylopars.pca(PPE)
summary(PPE_pca)
biplot(PPE_pca)
```

`phylopars.predict` *Ancestral state reconstruction and missing value tip reconstruction*

Description

This function performs ancestral state reconstruction (and missing value tip reconstruction) given a fitted `phylopars` object.

Usage

```
phylopars.predict(PPE, nodes, tips, verbose = FALSE)
```

Arguments

`PPE`	Fitted `phylopars` object
`nodes`	Optional argument: vector of which internal nodes to reconstruct.
`tips`	Optional argument: vector of which species values to predict.
`verbose`	Whether or not to print progress of node reconstruction / tip prediction (default is FALSE).

Value

Returns a list of reconstructed node and tip means, as well as standard deviations for each mean.

Author(s)

Eric W. Goolsby <`eric.goolsby.evolution@gmail.com`>

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

phylopars.predict(PPE)
```

`phylopars.rerun` *Rerun phylopars optimization*

Description

This function reruns parameter opimization for a phylopars object if convergence issues are suspected.

Usage

```
phylopars.rerun(PPE)
```

Arguments

`PPE`	Fitted `phylopars` object

Value

Returns a fitted object of class `phylopars`.

Author(s)

Eric W. Goolsby <`eric.goolsby.evolution@gmail.com`>

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

# suspect convergence failure, so optimize parameters again
PPE <- phylopars.rerun(PPE)
```

`plot.crossvalidate` *Phylopars cross-validation*

Description

Plots cross-validation results

Usage

```
## S3 method for class 'crossvalidate'
plot(x, ...)
```

Arguments

`x`	Crossvalidate object
`...`	

Value

Returns a summary of cross-validation statistics.

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

`print.crossvalidate` *Phylopars cross-validation*

Description

Generic S3 method for crossvalidate objects

Usage

```
## S3 method for class 'crossvalidate'
print(x, ...)
```

Arguments

`x`	Crossvalidate object
`...`	

Value

Prints cross-validation results.

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

`print.phylopars` *Print phylopars*

Description

Generic S3 method for phylopars

Usage

```
## S3 method for class 'phylopars'
print(x, ...)
```

Arguments

`x`	Fitted phylopars object
`...`	

`print.phylopars.lm` *Print phylopars.lm*

Description

Generic S3 method for phylopars.lm

Usage

```
## S3 method for class 'phylopars.lm'
print(x, ...)
```

Arguments

`x`	Fitted phylopars.lm object
`...`	

`simtraits` *Simulate traits for phylopars estimation*

Description

Simulates traits for codephylopars estimation.

Usage

```
simtraits(ntaxa = 15, ntraits = 4, nreps = 3, nmissing = 10, tree, v, anc, intraspecific)
```

Arguments

`ntaxa`	Either number of taxa (`ntaxa`) or a `tree` can be supplied.
`ntraits`	Number of traits to be simulated.
`nreps`	Number of replicates per trait per species to simulate.
`nmissing`	Number of randomly missing trait values.
`tree`	Either number of taxa (`ntaxa`) or a `tree` can be supplied.
`v`	Trait covariance (`v`) can be optionally supplied; otherwise off-diagonal elements are set to `0.8`.
`anc`	Value for ancestral state at root node for `fastBM` from the **phytools** package.
`intraspecific`	Optional value for within-species variance.

Value

`data`	Data for `phylopars()`
`tree`	Tree for `phylopars()`

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

# estimate parameters under Brownian motion
# pheno_error = TRUE assumes intraspecific variation
# pheno_correlated = FALSE assumes intraspecific variation is not correlated
# phylo_correlated = TRUE assumed traits are correlated

PPE <- phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,
  pheno_error = TRUE,phylo_correlated = TRUE,pheno_correlated = FALSE)

PPE
```

`summary.phylopars` *Phylopars summary*

Description

Summarizes phylopars

Usage

```
## S3 method for class 'phylopars'
summary(object, ...)
```

Arguments

`object`	Fitted phylopars object
`...`	

`summary.phylopars.lm` *phylopars.lm summary*

Description

Summarizes phylopars.lm

Usage

```
## S3 method for class 'phylopars.lm'
summary(object, ...)
```

Arguments

`object`	Fitted phylopars.lm object
`...`	

`write.phylopars` *Write data and tree files for Python phylopars compatability.*

Description

Writes data and tree files for Python phylopars compatibility.

Usage

```
write.phylopars(trait_data, tree, data_file, tree_file, species_identifier = "species")
```

Arguments

`trait_data`	A data frame with one column per trait, as well as a column labeled "species" (with species names matching tips on the phylogeny). Each row corresponds to a single observation, and multiple observation for species are allowed. Missing data should be represented with NA.
`tree`	An object of class `phylo`
`data_file`	Desired path to write data file.
`tree_file`	Desired path to write tree file.
`species_identifier`	Title of species column in data file. Defaulted to `"species"`

Author(s)

Eric W. Goolsby `<eric.goolsby.evolution@gmail.com>`

References

Bruggeman J, Heringa J and Brandt BW. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research* 37: W179-W184.

Examples

```
# simulate data
sim_data <- simtraits(ntaxa = 15,ntraits = 4,nreps = 3,nmissing = 10)

write.phylopars(trait_data = sim_data$trait_data,tree = sim_data$tree,data_file = "data_path.txt",
tree_file = "tree_path.new")
```

Index