Package 'VARSEDIG'

April 12, 2018

Version 1.6

Title An Algorithm for Morphometric Characters Selection and Statistical Validation in Morphological Taxonomy

Author Cástor Guisande González

Maintainer Cástor Guisande González <castor@uvigo.es>

Description An algorithm which identifies the morphometric features that significantly discriminate two taxa and validates the morphological distinctness between them via a Monte-Carlo test, polar coordinates and overlap of the area under the density curve.

License GPL (>= 2)

Encoding latin1

Depends R (>= 3.1.1)

Suggests adehabitatHS, kulife, MASS, car, ade4, IDPmisc

Repository CRAN

NeedsCompilation no

Date/Publication 2018-04-12 10:30:35 UTC

R topics documented:

`characiformes` *MORPHOMETRIC VARIABLES OF CHARACIFORMS*

Description

Morphometric data of several species of Characiforms, as the length of the dorsal fin base (M12), body height (M11), etc. For details see Guisande et al. (2010).

Usage

```
data(characiformes)
```

Format

An array (matrix) with 31 columns: taxonomic data (order, family, genus and species) and 27 morphometric variables.

Source

http://www.ipez.es.

References

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

VARSEDIG *Variable selection to discriminate among taxonomic groups*

Description

This function performs an algorithm for morphometric characters selection and statistical validation in morphological taxonomy.

Usage

```
VARSEDIG(data, variables, group, group1, group2, method="overlap", stepwise=TRUE,
VARSEDIG=TRUE, minimum=TRUE, kernel="gaussian", cor=TRUE, ellipse=TRUE, convex=FALSE,
DPLOT=NULL, SCATTERPLOT=NULL, BIVTEST12=NULL, BIVTEST21=NULL, Pcol="red",
colbiv="lightblue", br=20, sub="", lty=1, lwd=2.5, ResetPAR=TRUE, PAR=NULL, XLABd=NULL,
YLABd=NULL, XLIMd=NULL, YLIMd=NULL, COLORd=NULL, COLORB=NULL, LEGENDd=NULL, AXISd=NULL,
MTEXTd= NULL, TEXTd=NULL, XLABs=NULL, YLABs=NULL, XLIMs=NULL, YLIMs=NULL,
PCHs=NULL, COLORs=NULL, LEGENDs=NULL, MTEXTs=NULL, TEXTs=NULL, LEGENDr=NULL,
MTEXTr= NULL, TEXTr=NULL, arrows=TRUE, larrow=1, ARROWS=NULL, TEXTa=NULL,
model="Model.rda", file1="Overlap.csv", file2="Coefficients.csv",
file3="Predictions.csv", file4="Polar coordinates.csv", file="Output.txt",
na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Data file.
variables	Variables to be selected.
group	Variable with the groups to be discriminated.

`group1` First group.

`group2` Second group.

`method` Three different methods for prioritizing the variables according to their capacity for discrimination can be used. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).

`stepwise` If TRUE, the logistic regression is applied by steps, in order to eliminate those variables that are not significant. The Akaike information criterion (*AIC*) is used to define what are the variables that are excluded.

`VARSEDIG` If it is TRUE, the variables are added for the estimation of polar coordinates in the priority order according to the method "overlap", "Monte-Carlo", or "logistic regression" and the variable is selected if it significantly contributes to discriminate between both groups. See details section for further information.

`minimum` If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. It is FALSE, the algorithm selects all significant variables,and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.

`kernel` A character string giving the smoothing kernel to be used for estimating the overlap of the area under the curve between groups. This must be one of "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine". For further details about the estimation of the density curve see the details section of the function density of base stats package.

`cor` If it is TRUE the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one that has a greater positive correlation.

`ellipse` If it is TRUE the ellipses with the levels of significance to the 0.5 (inner ellipse) and 0.95 (outer ellipse) of each category of the variable *group* is depicted. These levels of significance can be modified by entering the function scatterplot using the argument *SCATTERPLOT* and modifying the argument *levels=c(0.5,0.95)*.

`convex` If it is TRUE the convex hull is depicted for each category.

DPLOT	It allows to specify the characteristics of the function plot.default of the density plot.
SCATTERPLOT	It accesses the function scatterplot of the car package, with the graph *biplot* that performs the X an Y polar coordinates.
BIVTEST12	It accesses the function biv.test of the package adehabitatHS, which performs the bivariate plot that displays the results of a bivariate randomisation test. From all values of group 2, it shows the value with higher probability to belong to group 1.
BIVTEST21	As in the argument *BIVTEST12*, but from all values of group 1, it shows the value with higher probability to belong to group 2.
Pcol	Color or name for the observation of group 2 in the BIVTEST12 plot and for the value of group 1 in the BIVTEST21 plot.
colbiv	Color or name of all values of group 1 in the BIVTEST12 plot and all values of group 2 in the BIVTEST21 plot.
br	Numbers of breaks of the histograms in the BIVTEST plots.
sub	Title in the BIVTEST plots.
lty	Type of line of the density curve for each group. If it is a vector, it must be as many as different categories of the variable *group*.
lwd	Line width relative to the default (default=1), so 2 is twice as wide of the density curve.
ResetPAR	If it is FALSE, the default condition of the functio PAR is not placed and maintained those defined by the user in previous graphics.
PAR	It accesses the function PAR that allows to modify many different aspects of the graph.
XLABd	Legend of the X axis in the density plot.
YLABd	Legend of the Y axis in the density plot.
XLIMd	Vector with the limits of the X axis in the density plot.
YLIMd	Vector with the limits of the Y axis in the density plot.
COLORd	Color of the density curves in the density plot. It must be as many as different categories of the variable *group*. As the color has transparency, the plot must be copy as bitmap and not metafile.
COLORB	Color of the lines in the density plot. It must be as many as different categories of the variable *group*.
LEGENDd	It allows to modify the legend of the density plot. If it is FALSE the legend is not shown.
AXISd	It allows to add axes to the density plot.
MTEXTd	It allows to add text on the margins of the density plot.
TEXTd	It allows to add text in any area of the inner part of the density plot.
XLABs	Legend of the X axis in the scatterplot.
YLABs	Legend of the Y axis in the scatterplot.
XLIMs	Vector with the limits of the X axis in the scatterplot.

YLIMs	Vector with the limits of the Y axis in the scatterplot.
PCHs	Vector with the symbols of the scatterplot, that should be as many as different groups the variable *group* has. If NULL, they are automatically calculated starting with the symbol 15.
COLORs	It allows to modify the colors of the scatterplot. It must be as many as different categories of the variable *group*.
LEGENDs	It allows to modify the legend of the scatterplot.
MTEXTs	It allows to add text on the margins of the scatterplot.
TEXTs	It allows to add text in any area of the inner part of the scatterplot.
LEGENDr	It allows to modify the legend of the BIVTEST plot. If it is FALSE the legend is not shown.
MTEXTr	It allows to add text on the margins of the BIVTEST plot.
TEXTr	It allows to add text in any area of the inner part of the BIVTEST plot.
arrows	If it is TRUE the arrows are shown in the scatterplot with the polar coordinates. These arrows show the vector of the variables selected when calculating the polar coordinates.
larrow	It modifies the length of the arrows.
ARROWS	It accesses the function Arrows of the package IDPmisc, which performs the arrows.
TEXTa	It allows to modify the labels at the end of the arrows.
model	Filename with the model of the binomial logistic regression.
file1	CSV FILE. Filename with the overlap of the area under the curve between both categories for all variables.
file2	CSV FILES. Filename with regression coefficients of the binomial logistic regression.
file3	CSV FILES. Filename with the predictions of the binomial logistic regression.
file4	CSV FILES. Filename with the polar coordinates for both categories of the variable *group*.
file	TXT FILE. Name of the output file with the results of the binomial logistic regression, the variables that significantly discriminate between the two groups and Euclidean distance between the two groups considering the polar coordinates.
na	CSV FILE. Text that is used in the cells without data.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
row.names	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

Classification methods such as logistic regression and discriminant analysis are probably the best available methods for the identification of the variables optimally able to predict group membership (Guisande et al. 2011; Guisande & Vaamonde 2012). Classification and Regression Trees (CARTs) are useful for identifying the variables that best discriminate groups, it is impossible using those methods to test the significance of the variables or to predict group membership (Guisande & Vaamonde 2012).

There are three advantages of logistic regression over discriminant analysis (Guisande et al., 2011): 1) the logistic regression is much more relaxed and flexible in its assumptions than the discriminant analysis because, unlike the discriminant analysis, the logistic regression does not have the requirements of the independent variables to be normally distributed, linearly related, nor equal variance within each group; 2) logistic regression may be more powerful and efficient analytic strategy if there are qualitative variables among predictors; 3) it is possible to use a stepwise logistic regression and, therefore, to select only those variables that significantly discriminate between groups. Discriminant analysis, however, does not have a statistical test of the coefficients of individual independent variables comparable to logistic regression, so it is not possible to test significance of variables and, therefore, to select only the variables that significantly predict group membership. Actually, to include variables with low discrimination capacity leads to reduce the identification success of the discriminant analysis.

The disadvantages of logistic regression are mainly also three: 1) the lack of a graphical representation of the results; 2) to evaluate the predictability of the final model chosen from the analysis it is not enough with the information about the percentage of cases correctly identified; 3) when the assumptions mentioned above regarding the distribution of predictors are met, discriminant function analysis may be more powerful and efficient analytic strategy than logistic regression (Tabachnick & Fidell, 1996)

This function performs an algorithm for: 1) prioritizing the variables by their discrimination capacity using three different methods, 2) selecting only those variables that significantly discriminate between two groups, 3) evaluating the predictability of the final model chosen with a Monte-Carlo test and 4) the results are graphically depicted in four different plots.

1. Prioritizing the variables by their discrimination capacity

Three different methods for prioritizing the variables according to their capacity for discrimination can be used.

1. If the argument *method="overlap"*, a density curve is obtained for each variable and the overlap of the area under the curve between the two groups is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. This information is saved in *file1="Overlap.csv"*.

2. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity).

3. If the argument *method="logistic regression"*, then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC). The model of the regression is saved in *model="Model.rda"*, the coefficients in *file2="Coefficients.csv"* and the predictions of the regression in *file3="Predictions.csv"*.

2. Polar coordinates

All variables are transformed to a scale ranged between -1 and 1. For each value the X and Y polar coordinates are estimated using the following equations:

$$X = \sum_{i=1}^{n} |z_j| cos(\alpha) \quad Y = \sum_{i=1}^{n} |z_j| sin(\alpha)$$

where *z* is the value of the variable *j* and *n* the number of variables.

Each variable is assigned an angle (α). The increment value of the angle is always $\frac{360}{n*2}$. If for instance the number of variables is 5, the increment angle is 36. Therefore, for the first variable if the value is ≥ 0 the α value is 36 and if the value is < 0 the value is 36+180, for the second variable if the value is ≥ 0 the α value is 36+36 and if the value is < 0 the value is 36+36+180, etc. Conversion of degrees to radians angle is carried out assuming that 1 degree = 0.0174532925 radians.

The order of the variables is consequently important because a different alpha value is assigned. If the argument *cor=TRUE*, this order is established calculating the correlation matrix of the variables and by ordering them such that each variable is followed by the variable to which it is highly correlated. The goal is to favor a larger dispersion of the data in the resulting polar coordinates system.

3. Algorithm for variables selection

The variables are added for the estimation of polar coordinates in the priority order according to *method="overlap"*, *method="Monte-Carlo"* or *method="logistic regression"*.

Mean X and Y polar coordinates are estimated for both groups and via these means the Euclidean distance is calculated between both groups.

In the case of the X and Y polar coordinates, a Monte-Carlo test is used for testing the statistical hypothesis if a value of one group is significantly higher or lower that the values of the other group. The test is performed for both X and Y polar coordinates and compares all values of one group with those of the other group. For instance, when all values of group 1 are compared with group 2, and the mean X polar coordinate of group 1 is higher than the one of group 2, the alternative hypothesis of the Monte-Carlo test is *greater*, and the p-value is estimated as (number of random values equal to or greater than the observed one + 1)/(number of permutations + 1). The null hypothesis is rejected if the p-value is less than the significance level. If the mean X polar coordinate of group 1 is lower than the one of group 2, the alternative hypothesis is *smaller*, a p-value is estimated as (number of random values equal to or less than the observed one + 1)/(number of permutations + 1). Again, the null hypothesis is rejected if the p-value is less than the significance level. The same process is applied when comparing all values of group 2 with those of group 1.

A variable is selected if it both: 1) contributes to increase Euclidean distance between both groups compared with the Euclidean distance obtained with the set of previously selected variables; and 2) the p-values of the Monte-Carlo test for X and Y coordinates when comparing both group 1 with group 2 and group 2 with group 1 are smaller than the p-values obtained with the set of previous selected variables. Therefore, from the pool of all independent variables, only those variables with the highest significant contribution to discriminating between both groups are selected.

The variables selected are saved in the *file="Output.txt"* and the polar coordinates of all values of both groups estimated with the variables selected are depicted in a scatterplot and saved in *file4="Polar coordinates.csv"*.

At the end of the process, it is selected the value with the highest p-value. Therefore, if this p-value is close or lower than the significance level of 0.05, it may be concluded that any of the values of one group may be identified as belonging to the other group.

Two plots are obtained with the value of the group 1 with the highest p-value of belonging to group 2 and the value of the group 2 with the highest p-value of belonging to group 1, respectively. In both plots, the x-axis corresponds to the X polar coordinates and the y-axis corresponds to Y polar coordinates.

If p-value is close or lower than 0.05 for X or Y polar coordinates, but in both cases when comparing group 1 with group 2 and group 2 with 1, it may be concluded that the variables selected are significantly contributing to discriminate between both groups, so with these variables is possible to achieve a 100% of identification success when predicting group membership.

FUNCTIONS

The density plot is performed with the function plot.default of base graphics package. The density curve is estimated with the function density of base stats package. The area under the curve is estimated with the function auc of the package kulife (Ekstrom et al., 2015). The random test was performed with the function as.randtest of the package ade4 (Chessel et al., 2004; Dray et al., 2007; 2015). The bivariate plot that displays the results of a bivariate randomisation test, for which the p-values are computed with the function as.randtest (one-sided tests), was performed with the function biv.test of the package adehabitatHS (Calenge, 2006; 2015). The arrows are depicted with the function Arrows of the package IDPmisc (Locher & Ruckstuhl, 2014). The scatterplot is performed with the function scatterplot of the car package (Fox & Weisberg, 2011; Fox et al., 2014). The convex hull is estimated with the function chull of the package grDevices.

EXAMPLES

For the example, morphometric data of three families of freshwater fishes, as the distance from the origin of the dorsal fin to the origin of the anal fin (M13), the length of the dorsal fin base (M12), body height (M11), etc., are used. For details see Guisande et al. (2010).

Figure shows the plots obtained with VARSEDIG (Guisande et al., 2016), in an example comparing the species *Moenkhausia dichroura* and *Moenkhausia oligolepis*.

The variables that better discriminate between both species are the M26 (interorbital width) and M11 (distance from the dorsal-fin origin to the dorsal limit of the pelvic-fin base). Between these two variables, a density plot is depicted for the quantitative variable with lower overlap between both groups and, thus, the highest discrimination capacity: in this example M26 (Figure 1A).

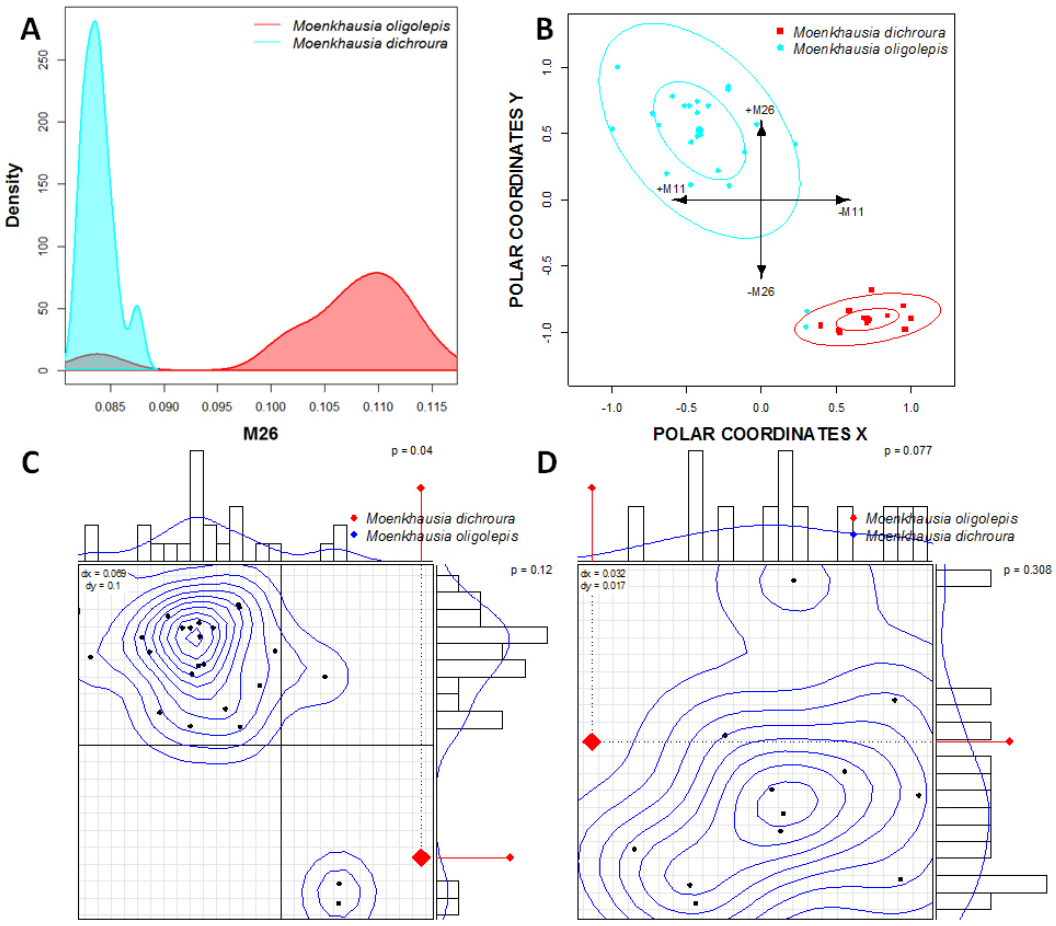
Figure 1B shows the scatterplot of the polar coordinates obtained for both species using variables M26 and M11. The arrows show the vector of the variables with both of these variables higher in *M. oligolepis*.

This example illustrates that the VARSEDIG algorithm is not only useful for identifying the variables that better discriminate between two taxa, but also may be informative when it comes to finding misidentified individuals. In the example, it appears that two individuals identified as *M. oligolepis* are *M. dichroura* (Figure 1B).

Figure 1C displays the results of a bivariate randomisation test. From all individuals of the species *M. dichroura*, the figure shows the individual of *M. dichroura* (red point) with higher probability to be identified as belonging to the M. oligolepis. Kernel density is estimated to indicate the contours of the distribution of randomised values. The two marginal histograms correspond to the univariate tests on each axis, for which the p-values (one-sided tests) are computed. As p-value is lower than 0.05 for X axis ($p = 0.04$), the null hypothesis is rejected. Consequently the X polar coordinates of

all individuals of the of the species *M. dichroura* are significantly different than those of the species *M. oligolepis* and, therefore, none of the individuals designated as *M. dichroura* may be identified as belonging to the species *M. oligolepis*.

Figure 1D also displays the results of a bivariate randomisation test but, in this case, from all individuals of the species *M. oligolepis*, the figure shows the individual (red point) with higher probability to belong to the species *M. dichroura*. Both p-values are higher than 0.05, so null hypothesis is accepted for both X and Y polar coordinates. This that some individuals of the species *M. oligolepis* may be identified as belonging to the species *M. dichroura*.



It is not necessary a p-value lower than 0.05 for both X and Y, but it is just necessary and p-value lower than 0.05 for X or Y when comparing both group 1 with 2 and group 2 with 1. Therefore, if p-value is close or lower than the significance level of 0.05 for X or Y polar coordinates in both cases comparing group 1 with 2 and group 2 with 1, it would mean a 100% of identification success between both groups. In this example, however, with the variables M16 and M11 is not possible to predict group membership with a 100% of accuracy because, although none of the individuals of the species *M. dichroura* may be identified as belonging to the species *M. oligolepis*, some individuals of the species *M. oligolepis* may be identified as belonging to the species *M. dichroura*. The failure to reach 100% may be due to the possible misidentification of two individuals of *M. dichroura* as

M. oligolepis.

Value

It is depicted 4 plots: 1) a density plot with the overlap of the area under de curve between the two groups for the variable that better discriminates between both groups, 2) a scatter plot with the polar coordinates for both groups, 3) a bivariate plot that shows from all values of group 2 the value with higher probability to belong to group 1, and 4) a bivariate plot that shows from all values of group 1 the value with higher probability to belong to group 2. Moreover, 5 files are saved: 1) overlap of the area under the curve between both categories for all variables, 2) regression coefficients of the binomial logistic regression, 3) predictions of the binomial logistic regression, 4) polar coordinates for both categories of the variable *group*, and 5) a TXT file with the results of the binomial logistic regression, the variables that better discriminate between the two groups and the Euclidean distance between groups considering the polar coordinates.

Author(s)

Cástor Guisande González, Universidad de Vigo, Spain.

References

Calenge, C. (2006) The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197, 516-519.

Calenge, C. (2016) Analysis of Habitat Selection by Animals. R package version 0.3.12. Available at: https://CRAN.R-project.org/package=adehabitatHS.

Chessel, D., Dufour, A.B. and Thioulouse, J. (2004) The ade4 package-I- One-table methods. *R News*, 4, 5-10.

Dray, S. & Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1-20.

Dray, S. & Dufour, A.B. and Chessel, D. (2007) The ade4 package-II: Two-table and K-table methods. *R News*, 7(2), 47-52.

Dray, S., Dufour, A-B. & Thioulouse, J. (2015) Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences. R package version 1.7-2. Available at: https://CRAN.R-project.org/package=ade4.

Ekstrom, C., Skovgaard, Ib M. & Martinussen, T.(2015) Datasets and functions from the (now non-existing). R package version 0.1-14. Available at: https://CRAN.R-project.org/package=kulife.

Fox, J. & Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R., Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W. & Zeileis, A. (2014) Companion to Applied Regression. R package version 2.0-20. Available at: https://CRAN.R-project.org/package=car.

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez,

S., Duque, V. & Salmerón, F. (2010) IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, 102, 240-247.

Guisande, C., Vaamonde, A. & Barreiro, A. (2011) *Tratamiento de datos con R, SPSS y STATISTICA*. Ediciones Díaz de Santos, Madrid, 978 pp.

Guisande, C. & Vaamonde, A. (2012) *Gráficos estadísticos y mapas con R*. Ediciones Díaz de Santos, Madrid, 367 pp.

Guisande, C., Vari, R.P., Heine, J., García-Roselló, E., González-Dacosta, J., Pérez-Schofield, B.J., González-Vilas, L. & Pelayo-Villamil, P. (2016) VARSEDIG: an algorithm for morphometric characters selection and statistical validation in morphological taxonomy. *Zootaxa*, 4162 (3), 571-580

Locher, R. & Ruckstuhl, A. (2014) Utilities of Institute of Data Analyses and Process Design. R package version 1.1.17. Available at: https://CRAN.R-project.org/package=IDPmisc.

Tabachnick, B.G. & Fidell, L.S. (1996) *Using Multivariate Statistics*. NY, HarperCollins.

Examples

```
data(characiformes)

VARSEDIG(data = characiformes , variables = c("M2","M3","M4","M5","M6","M7","M8","M9","M10",
"M11","M12","M13","M14","M15","M16","M17","M18","M19","M20","M21","M22","M23",
"M24","M25","M26","M27","M28"), group="Species" , group1= "Moenkhausia oligolepis",
group2="Moenkhausia dichroura", LEGENDd=c("x='topright'", "legend = dati",
"col = COLORB", "lty=lty", "bty='n'", "cex=1.2", "text.font= 3"),
LEGENDs=c("x='topright'", "legend=unique(datosF[,'Group'])", "col = color1",
"pch = pcht", "bty='n'", "cex=1.2", "text.font=3"), LEGENDr=c("x='topright'",
"legend = dati", "col=col", "pch= c(16,16)", "bty='n'", "cex=1.2", "text.font=3"),
XLIMs=c(-1.2,1.2), YLIMs=c(-1.3,1.3), BIVTEST12=c("br=br", "cex=1.1",
"col=colbiv", "sub=sub", "Pcol=Pcol"), BIVTEST21=c("br=br", "cex=1.1",
"col=colbiv", "sub=sub", "Pcol=Pcol"), colbiv="blue")
```

VARSEDIM *Variable selection to discriminate many taxonomic groups*

Description

This function performs an algorithm for morphometric characters selection and statistical validation in morphological taxonomy among many taxonomic groups.

Usage

```
VARSEDIM(data, variables, group, method="overlap", stepwise=TRUE,
VARSEDIG=TRUE, minimum=TRUE, kernel="gaussian", cor=TRUE, file1="Overlap.csv",
file2="Coefficients.csv", file3="Predictions.csv", file4="Polar coordinates.csv",
file="Resuts.txt", na="NA", dec=",", row.names=FALSE)
```

Arguments

data	Data file.
variables	Variables to be selected.
group	Variable with the groups to be discriminated.
method	Three different methods for prioritizing the variables according to their capacity for discrimination can be used. If the method is "overlap", a density curve is obtained for each variable and the overlap of the area under the curve between the two groups of the variable *group* is estimated for all variables. Those variables with lower overlap should have better discrimination capacities and, hence, all variables are ordered from lowest to highest overlap; in other words from the highest to lowest discrimination capacity. If the method is "Monte-Carlo", a Monte-Carlo test is performed comparing all values of group 1 with group 2, and all values of group 2 with 1. The variables are prioritized from the variable with the lowest mean of all p-values (highest discrimination capacity) to the variable with the highest mean of all p-values (lowest discrimination capacity). If the method is "logistic regression", then a binomial logistic regression is calculated and if the argument stepwise=TRUE (default option), then only significant variables are selected for further analyses with the regression performed by steps using the Akaike Information Criterion (AIC).
stepwise	If TRUE, the logistic regression is applied by steps, in order to eliminate those variables that are not significant. The Akaike information criterion (*AIC*) is used to define what are the variables that are excluded.
VARSEDIG	If it is TRUE, the variables are added for the estimation of polar coordinates in the priority order according to the method "overlap", "Monte-Carlo", or "logistic regression" and the variable is selected if it significantly contributes to discriminate between both groups. See details section for further information.
minimum	If it is TRUE, the algorithm is designed to find a significant discrimination between both groups with the minimum possible number of significant variables. Therefore, only the variables with higher discrimination capacity are selected. It is FALSE, the algorithm selects all significant variables,and not only those with higher discrimination capacity. This argument is only valid with the methods "Monte-Carlo" and "overlap" and it is useful in those cases that discrimination between the groups is difficult and requires to include as many as variables as possible.
kernel	A character string giving the smoothing kernel to be used for estimating the overlap of the area under the curve between groups. This must be one of "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine". For further details about the estimation of the density curve see the details section of the function density of base stats package.
cor	If it is TRUE the variables are ordered according to the correlation between them when estimating the polar coordinates. Therefore, the next variable to another variable is the one that has a greater positive correlation.
file1	CSV FILE. Filename with the overlap of the area under the curve between both categories for all variables.
file2	CSV FILES. Filename with regression coefficients of the binomial logistic regression.

file3	CSV FILES. Filename with the predictions of the binomial logistic regression.
file4	CSV FILES. Filename with the polar coordinates for both categories of the variable *group*.
file	TXT FILE. Name of the output file with the results of the binomial logistic regression, the variables that significantly discriminate between the two groups and Euclidean distance between the two groups considering the polar coordinates.
na	CSV FILE. Text that is used in the cells without data.
dec	CSV FILE. It defines if the comma "," is used as decimal separator or the dot ".".
row.names	CSV FILE. Logical value that defines if identifiers are put in rows or a vector with a text for each of the rows.

Details

The difference with the function VARSEDIG is that all the different taxa of the variable group are compared with each other, instead of just comparing two taxa. It uses the same algorithm described in the function VARSEDIG.

Value

It is obtained a TXT file, which is called by default Results.TXT, with the results of the comparison of each taxon with the rest of individuals of the other taxa (named "Others").

For each comparison, it is depicted 4 plots: 1) a density plot with the overlap of the area under de curve between the two groups for the variable that better discriminates between both groups, 2) a scatter plot with the polar coordinates for both groups, 3) a bivariate plot that shows from all values of group 2 the value with higher probability to belong to group 1, and 4) a bivariate plot that shows from all values of group 1 the value with higher probability to belong to group 2.

Finally, fot the last comparison 5 files are saved: 1) overlap of the area under the curve between both categories for all variables, 2) regression coefficients of the binomial logistic regression, 3) predictions of the binomial logistic regression, 4) polar coordinates for both categories of the variable *group*, and 5) a TXT file with the results.

Author(s)

Cástor Guisande González, Universidad de Vigo, Spain.

Examples

```
## Not run:
data(characiformes)
VARSEDIM(data=characiformes, variables= c("M2", "M3", "M4",  "M5", "M6",
"M7", "M8", "M9", "M10", "M11", "M12", "M13", "M14", "M15", "M16", "M17",
"M18", "M19", "M20", "M21", "M22", "M23", "M24", "M25", "M26", "M27", "M28"),
group="Genus")

## End(Not run)
```

Index