Package ‘isopam’

February 20, 2015

Type Package
Title Isopam (Clustering)
Version 0.9-13
Date 2014-12-08
Author Sebastian Schmidtlein
Maintainer Sebastian Schmidtlein <schmidtlein@kit.edu>
Depends vegan, cluster
Suggests proxy
Description Isopam clustering algorithm and utilities.
Isopam optimizes clusters and optionally cluster numbers in a brute force style and aims at an optimum separation by all or some descriptors (typically species).
License GPL-2
LazyLoad yes
NeedsCompilation no
Repository CRAN
Date/Publication 2014-12-09 13:40:46

R topics documented:

andechs .................................................. 2
isopam .................................................. 2
isotab .................................................. 5

Index 8
### andechs

**Description**

This data set gives the average cover of vascular plant species in subplots nested within 17 whole-plots from mown fen meadows. This is a subset of the data used in Schmidtlein & Sassin (2004).

**Usage**

```r
data(andechs)
```

**Format**

A matrix containing 17 observations and 110 species.

**Source**


---

### isopam

**Isopam (Clustering)**

**Description**

Isopam classification is performed either as a hierarchical, divisive method, or as non-hierarchical partitioning. Optimizes clusters and optionally cluster numbers for maximum performance of group indicators. Developed for matrices representing species abundances in plots.

**Usage**

```r
isopam (dat, c.fix = FALSE, c.opt = TRUE, c.max = 6,
1.max = FALSE, stopat = c(1,7), sieve = TRUE,
Gs = 3.5, ind = NULL, centers = NULL, distance = 'bray',
k.max = 100, d.max = 7, ..., juice = FALSE)
```

```r
## S3 method for class 'isopam'
identify(x, ...)
## S3 method for class 'isopam'
plot(x, ...)
```
Arguments

dat  data matrix: each row corresponds to an object (typically a plot), each column corresponds to a descriptor (typically a species). All variables must be numeric. Missing values (NAs) are not allowed. At least 3 rows (plots) are required.

c.fix  number of clusters (defaults to FALSE). If a number is given, non-hierarchical partitioning is performed. c.opt and c.max are ignored and l.max is set to one.

c.opt  if TRUE (the default) cluster numbers are optimized in the range between 2 and c.max (slow and thorough). If FALSE groups are divided into two subgroups (quick and dirty).

c.max  maximum number of clusters per partition. Applies to all partitioning steps if c.opt = TRUE.

l.max  maximum number of hierarchy levels. Defaults to FALSE (no maximum number). Note that divisions may stop well before this number is reached (see stopat). Use l.max = 1 for non-hierarchical partitioning (or use c.fix).

stopat  vector with stopping rules for hierarchical clustering. Two values define if a partition should be retained in hierarchical clustering: the first determines how many indicators must be present per cluster, the second defines the standardized G-value that must be reached by these indicators. stopat is not effective at the first hierarchy level or in non-hierarchical partitioning.

sieve  logical. If TRUE (the default), only descriptors (species) exceeding a threshold defined by Gs are used in the search for a good clustering solution. Their number is multiplied with their mean standardized G-value. The product is used as optimality criterion. If FALSE all descriptors are used for optimization.

Gs  threshold (standardized G value) for descriptors (species) to be considered in the search for a good clustering solution. Effective with sieve = TRUE.

ind  optional vector of column names from dat defining descriptors (species) used as indicators. This turns Isopam in an expert system. Replaces the automated selection of indicators with sieve = TRUE (ind overrules sieve).

centers  optional vector with observations used as cluster cores (supervised classification).

distance  distance measure for the distance matrix used as a starting point for Isomap. Any distance measure implemented in packages vegan or proxy can be used (see details).

k.max  maximum Isomap k.

d.max  maximum number of Isomap dimensions.

...  other arguments to S3 functions plot and identify corresponding to hclust.

juice  logical. If TRUE input files for Juice are generated.

x  an isopam result object.

Details

Isopam is described in Schmidtlein et al. (2010). It consists of dimensionality reduction (Isomap: Tenenbaum et al. 2000; isomap in vegan) and partitioning of the resulting ordination space (PAM:
Kaufman & Rousseeuw 1990; pam in cluster). The classification is performed either as a hierarchical, divisive method, or as non-hierarchical partitioning. Compared to other clustering methods, it has the following features: (a) it optimizes partitions for the performance of group indicators (typically species) or for maximum average ‘fidelity’ of descriptors to groups; (b) it optionally selects the number of clusters per division; (c) the shapes of groups in feature space are not limited to spherical or other regular geometric shapes (thanks to the underlying Isomap algorithm) and (d) the distance measure used for the initial distance matrix can be freely defined.

Currently, the plot and identify methods for class isopam simply link to the hclust object $dendro resulting from isopam in case of hierarchical partitioning. The methods work just like plot.hclust and identify.hclust.

The preset distance measure is Bray-Curtis (Odum 1950). Distance measures are passed to vegdist in vegan. If vegan does not know the given measure it is passed to dist in proxy. Measures available in vegan are listed in vegdist. Measures registered in proxy can be listed with summary(pr_DB) once proxy is loaded. New measures can be defined and registered as described in _pr_db. Isopam can’t deal with distance matrices as a replacement for the original data matrix because it operates on individual descriptors (species).

Value

call generating call
distance distance measure used by Isomap
flat observations (plots) with group affiliation. Running group numbers for each level of the hierarchy.
hier observations (plots) with group affiliation. Group identifiers reflect the cluster hierarchy. Not present with only one level of partitioning.
medoids observations (plots) representing the medoids of the resulting groups.
analytics table summarizing parameter settings for the final partitioning steps. Name: name of the respective parent cluster (0 in case of the first partition); Subgroups: number of subgroups; Isomap. dim: Isomap dimensions used; Isomap. k.min: minimum possible Isomap k; Isomap. k: Isomap k used; Isomap. k.max: maximum possible Isomap k; Ind.N: number of indicators reaching or exceeding Gs; Ind. Gs: the average standardized G value of these indicators; and Global.Gs: the average standardized G value of all descriptors.
dendro an object of class hclust representing the clustering. Not present with only one level of partitioning.
dat data used

Note

For large datasets, Isopam may need too much memory or too much computation time. The optimization procedure (selection of Isomap dimensions and -k, optionally selection of cluster numbers) is based on a brute force approach that takes its time with large data sets. Low speed is inherent to the method, so don’t complain. If used with data not representing species in plots make sure that the indicator approach is appropriate.

With very small datasets, the indicator based optimization may fail. In such cases consider using filtered = FALSE instead of the default method.
isotab

Author(s)
Sebastian Schmidtlein with contributions from Jason Collison and Lubomir Tichý

References

See Also
isotab for a table of descriptor (species) frequency in clusters.

Examples
```r
## load data to the current environment
data(andechs)

## call isopam with the standard options
ip<-isopam(andechs)

## examine cluster hierarchy
plot(ip)

## examine grouping
ip$flat

## examine frequency table (second hierarchy level)
isotab(ip, 2)

## non-hierarchical partitioning
ip<-isopam(andechs, c.flix=3)
ip$flat
```

---

isotab  

*Ordered frequency table for Isopam clusters*

Description
Computes an ordered frequency table based on Isopam clustering results. The upper part of the table lists typical descriptors (usually species) with a significant binding to single clusters (according to customisable thresholds). The lower part of the table is ordered by descending overall frequency.
Usage

isotab(ip, level = 1, phi.min = 'auto', p.max = .05)

Arguments

ip object of class isopam.
level level in cluster hierarchy starting with 1 = first division.
phi.min threshold of \( \phi \) determining which descriptors (species) are listed in the upper part of the table. Applies only to descriptors passing the criterion defined by \( p.N.max \). If \( \phi.min = \text{auto} \) (the default) isotab suggests a suitable value based on the numbers of clusters, observations, and descriptors.
p.max threshold of Fisher’s \( p \) determining which descriptors (species) are listed in the upper part of the table. Applies only to descriptors passing the criterion defined by \( \phi.min \).

Details

\( \phi.min \) is based on the standardized \( \phi \) value according to Chitrý et al. 2002.

Value

tab dataframe with ordered frequencies and their significance. The latter is derived from Fisher’s exact test (\( p \leq 0.05: *, p \leq 0.01: **, p \leq 0.001: *** \)).
n matrix with cluster sizes.
thresholds \( \phi.min \) and \( p.max \) used.
typical dataframe with items (often species) typically found in clusters (according to thresholds).

Author(s)

Sebastian Schmidtlein

References


See Also

isopam
Examples

```r
## load data to the current environment
data(andechs)

## call isopam with the standard options
ip<-isopam(andechs)

## build table (uppermost hierarchy level)
isotab(ip)

## build table (lower hierarchy level)
isotab(ip,2)
```
Index

*Topic cluster
  isopam, 2
*Topic datasets
  andechs, 2

andechs, 2
dist, 4
hclust, 3, 4
identify.isopam(isopam), 2
isomap, 3
isopam, 2, 6
isotab, 5, 5

pam, 4
plot.isopam(isopam), 2

vegdist, 4