Package 'ICGE'

July 21, 2025

Version 0.4.2

Date 2022-10-14

Title Estimation of Number of Clusters and Identification of Atypical Units

Maintainer Itziar Irigoien <itziar.irigoien@ehu.eus>

Depends R (>= 2.0.1), utils, stats, MASS, cluster, fastcluster

Description It is a package that helps to estimate the number of real clusters in data as well as to identify atypical units. The underlying methods are based on distances rather than on unit x variables.

License GPL (>= 2)

LazyData yes

Author Itziar Irigoien [aut, cre], Concepcion Arenas [aut]

NeedsCompilation no

Repository CRAN

Date/Publication 2022-10-17 09:25:23 UTC

Contents

`chowdary` *Chowdary Database*

Description

The original authors compared pairs of snap-frozen and RNAlater preservative-suspended tissue from lymph node-negative breast tumors (B) and Dukes' B colon tumors (C). The actual data set, by de Souto et. al (2008), is build with purpose of separating B from C.

Usage

```
data(chowdary)
```

Format

Data frame with 183 rows and 104 columns.

Source

Original source from 'National Center for Biotechnology Information' from the United States of America, query GSE3726.

References

de Souto MCP, Costa IG, de Araujo DSA, Ludermir TB, and Schliep A (2008). Clustering Cancer Gene Expression Data: a Comparative Study. *BMC Bioinformatics*, **8**, 497–511.

Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, Yu J, Wang X, and Mazumder A (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *Journal Molecular Diagnosis*, **8**, 31–39.

Examples

```
data(chowdary)

tumor <- as.factor(as.matrix(chowdary[1,]))
x <- as.matrix(chowdary[-1,])
mode(x) <- "numeric"

s <- sample(row.names(x),1)
boxplot( x[s,] ~ tumor , ylab=s)
```

dbhatta *Bhattacharyya Distance*

Description

dbhatta computes and returns the Bhattacharyya distance matrix between the rows of a data matrix. This distance is defined between two units $i = (p_{i1}, ..., p_{im})$ and $j = (p_{j1}, ..., p_{jm})$ being p_{kl} frequencies with $p_{kl} >= 0$ and $p_{k1} + ... + p_{km} = 1$.

Usage

```
dbhatta(x)
```

Arguments

x	a matrix containing, in its rows, the frequencies for each unit. Note: check that each row adds up to 1

Value

A dist object with distance information.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV-EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Bhattacharyya, A. (1946). On a measure of divergence of two multinomial populations. *Sankhya: The Indian Journal of Statistics, Series A*. **14**, 177-136.

See Also

dist, dmahal, dgower, dcor, dproc2

Examples

```
#5 individuals represented by their relative frequencies of 4 characteristics (M1-M4):
f <- matrix(c(0.36, 0.21, 0.23, 0.20,
              0.66, 0.18, 0.11, 0.05,
              0.01, 0.24, 0.62, 0.13,
              0.43, 0.38, 0.08, 0.11,
              0.16, 0.07, 0.09, 0.68),
              byrow=TRUE, nrow=5, dimnames=list(1:5, paste("M", 1:4, sep="")))
```

```
# Bhattacharyya distances between pairs
d <- dbhatta(f)
```

---

dcor *Correlation Distance*

Description

dcor computes and returns the Correlation distance matrix between the rows of a data matrix. This distance is defined by $d = \sqrt{1 - r}$.

Usage

```
dcor(x)
```

Arguments

x	a numeric matrix.

Value

A dist object with distance information.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV-EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Gower, J.C. (1985). Measures of similarity, dissimilarity and distance. In: *Encyclopedia of Statistical Sciences*, volume **5**, 397–405. J. Wiley and Sons.

See Also

dist, dmahal, dgower, dbhatta, dproc2

Examples

```
#Generate 10 objects in dimension 8
n <- 10
mu <- sample(1:10, 8, replace=TRUE)
x <- matrix(rnorm(n*8, mean=mu, sd=1), nrow=n, byrow=TRUE)


# Correlation distances between pairs
d <- dcor(x)
```

deltas *Distance Between Groups*

Description

Assume that n units are divided into k groups C1,...,Ck . Function `deltas` computes and returns the distance between each pair of groups. It uses the distances between pairs of units.

Usage

```
deltas(d, pert = "onegroup")
```

Arguments

`d`	a distance matrix or a `dist` object with distance information between units.
`pert`	an n-vector that indicates which group each unit belongs to. Note that the expected values of `pert` are numbers greater than or equal to 1 (for instance 1,2,3,4..., k). The default value indicates there is only one group in data.

Value

A matrix containing the distances between each pair of groups.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

Cuadras, C.M., Fortiana, J. and Oliva, F. (1997). The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification*, **14**, 117–136.

See Also

`vgeo`, `proxi`

Examples

```
data(iris)
d <- dist(iris[,1:4])
deltas(d,iris[,5])
```

`dermatology` *Dermatology Database*

Description

Data from a dermatology study provided by H.A. Guvenir (Dpt. Computer Engineering and Information Science, Bilkent University, Turkey).The data set contains 366 instances presenting 34 different clinical attributes (12 clinical features as age or family history and 22 histopathological features obtained from a biopsy), and a class variable indicating the disease. There are 8 missing values. This data set has been used extensively for classification tasks.

Usage

```
data(dermatology)
```

Format

Matrix with 366 rows.

Details

Attribute information obtained from the UCI KDD data repository:

Clinical Attributes: (they take values 0, 1, 2, 3, unless otherwise indicated)

1: erythema; 2: scaling; 3: definite borders; 4: itching; 5: koebner phenomenon; 6: polygonal papules; 7: follicular papules; 8: oral mucosal involvement; 9: knee and elbow involvement; 10: scalp involvement; 11: family history, (0 or 1); 34: Age.

Histopathological Attributes: (they take values 0, 1, 2, 3)

12: melanin incontinence; 13: eosinophils in the infiltrate; 14: PNL infiltrate; 15: fibrosis of the papillary dermis; 16: exocytosis; 17: acanthosis; 18: hyperkeratosis; 19: parakeratosis; 20: clubbing of the rete ridges; 21: elongation of the rete ridges; 22: thinning of the suprapapillary epidermis; 23: spongiform pustule; 24: munro microabcess; 25: focal hypergranulosis; 26: disappearance of the granular layer; 27: vacuolisation and damage of basal layer; 28: spongiosis; 29: saw-tooth appearance of retes; 30: follicular horn plug; 31: perifollicular parakeratosis; 32: inflammatory monoluclear inflitrate; 33: band-like infiltrate.

The considered diseases are: 1 - psoriasis, 2 - seboreic dermatitis, 3- lichen planus, 4 - pityriasis rosea, 5 - chronic dermatitis, 6 - pityriasis rubra pilaris.

Source

The UCI KDD Archive.

References

Guvenir H, Demiroz G, Ilter N (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, **13**, 147–165.

Irigoien I, Arenas C (2008). INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**, 2948–2973.

Examples

```
data(dermatology)
x <- dermatology[, 1:34]
group <- as.factor(dermatology[,35])

plot(group)
```

---

dgower *Gower Distance for Mixed Variables*

Description

dgower computes and returns the Gower distance matrix for mixed variables.

Usage

```
dgower(x, type = list())
```

Arguments

x	data matrix.
type	it is a list with components cuant, bin, nom. Each component indicates the column position of the quantitative, binary or nominal variables, respectively.

Details

The distance between two pairs of objects i and j is obtained as $\sqrt{2(1-s_{ij})}$ where s_{ij} is the Gower's similarity coefficient for mixed data. This function allows to include missing values (as NA) and therefore calculates distances based on Gower's weighted similarity coefficient.

Value

A dist object with distance information.

Note

There is the function daisy() in cluster package which can perform the Gower distance for mixed variables. The difference is that in daisy() the distance is calculated as $d(i,j) = 1 - s_{ij}$ and in dgower() it is calculated as $dij = sqrt(1 - s_{ij})$.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

See Also

dist, dmahal, dbhatta, dcor, dproc2

Examples

```
#Generate 10 objects in dimension 6
# Quantitative variables
mu <- sample(1:10, 2, replace=TRUE)
xc <- matrix(rnorm(10*2, mean = mu, sd = 1), ncol=2, byrow=TRUE)

# Binary variables
xb <- cbind(rbinom(10, 1, 0.1), rbinom(10, 1, 0.5), rbinom(10, 1, 0.9))

# Nominal variables
xn <- matrix(sample(1:3, 10, replace=TRUE), ncol=1)

x <- cbind(xc, xb, xn)

# Distances
d <- dgower(x, type=list(cuant=1:2, bin=3:5, nom=6))
```

---

dmahal *Mahalanobis Distance*

Description

dmahal computes and returns the Mahalanobis distance matrix between the rows of a data matrix.

Usage

```
dmahal(datos, S)
```

Arguments

datos	data matrix.
S	covariance matrix.

Value

A dist object with distance information.

Note

There is a function `mahalanobis()` in stats package which can perform the Mahalanobis distance. While `mahalanobis()` calculates the Mahalanobis distance with respect to given a center, function `dmahal()` is designed to calculate the distance between each pair of units given a data matrix.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Everitt B. S. and Dunn G. (2001) *Applied Multivariate Data Analysis*. 2 edition, Edward Arnold, London.

See Also

`dist`, `dbhatta`, `dgower`, `dcor`, `dproc2`

Examples

```
#Generate 10 objects in dimension 2
mu <- rep(0, 2)
Sigma <- matrix(c(10,3,3,2),2,2)

x <- mvrnorm(n=10, rep(0, 2), Sigma)

d <- dmahal(x, Sigma)
```

---

dproc2 *Modified Procrustes distance*

Description

`dproc2` computes and returns all the pairwise procrustes distances between genes in a time course experiment, using their expression profile.

Usage

```
dproc2(x, timepoints = NULL)
```

Arguments

`x`	a matrix containing, in its rows, the gene expression values at the T considered time points.
`timepoints`	a T-vector with the T observed time points. If `timepoints=NULL` (default), then timepoints=1:T.

Details

Each row i of matrix x is arranged in a two column matrix Xi. In Xi, the first column contains the time points and the second column the observed gene expression values (xi1...).

Value

A dist object with distance information.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Irigoien, I. , Vives, S. and Arenas, C. (2011). Microarray Time Course Experiments: Finding Profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**(2), 464–475.

Gower, J. C. and Dijksterhuis, G. B. (2004) *Procrustes Problems*. Oxford University Press.

Sibson, R. (1978). Studies in the Robustness of Multidimensional Scaling: Procrustes statistic. *Journal of the Royal Statistical Society, Series B*, **40**, 234–238.

See Also

dist, dmahal, dgower, dcor dbhatta

Examples

```
# Given  10  hypothetical time course profiles
# over 6 time points at 1, 2, ..., 6 hours.
x <- matrix(c(0.38, 0.39, 0.38, 0.37, 0.385, 0.375,
              0.99, 1.19, 1.50, 1.83, 2.140, 2.770,
              0.38, 0.50, 0.71, 0.72, 0.980, 1.010,
              0.20, 0.40, 0.70, 1.06, 2.000, 2.500,
              0.90, 0.95, 0.97, 1.50, 2.500, 2.990,
              0.64, 2.61, 1.51, 1.34, 1.330 ,1.140,
              0.71, 1.82, 2.28, 1.72, 1.490, 1.060,
              0.71, 1.82, 2.28, 1.99, 1.975, 1.965,
              0.49, 0.78, 1.00, 1.27, 0.590, 0.340,
              0.71,1.00, 1.50, 1.75, 2.090, 1.380), nrow=10, byrow=TRUE)

# Graphical representation
matplot(t(x), type="b")

# Distance matrix between them
d <- dproc2(x)
```

estW *INCA Statistic*

Description

Assume that n units are divided into k clusters C1,...,Ck, and consider a fixed unit x0. Function estW calculates the INCA statistic $W(x0)$ and the related U_i statistics.

Usage

```
estW(d, dx0, pert = "onegroup")
```

Arguments

d	a distance matrix or a dist object with distance information between units.
dx0	an n-vector containing the distances d0j between x0 and unit j.
pert	an n-vector that indicates which group each unit belongs to. Note that the expected values of pert are consecutive integers bigger or equal than 1 (for instance 1,2,3,4..., k). The default value indicates the presence of only one group in data.

Value

The function returns an object of class incaest which is a list containing the following components:

Wvalue	is the INCA statistic $W(x_0)$.
Uvalue	is a vector containing the statistics U_i.

Note

For a correct geometrical interpretation it is convenient to verify whether the distance matrix d is Euclidean.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

Irigoien, I. and Arenas, C. (2008). INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**(15), 2948–2973.

See Also

vgeo, proxi , deltas

Examples

```
data(iris)
d <- dist(iris[,1:4])

# characteristics of a specific flower (likely group 1)
x0 <- c(5.3, 3.6, 1.1, 0.1)
# distances between  flower x0 and the rest of flowers in iris
dx0 <- rep(0,150)
for (i in 1:150){
dif <-x0-iris[i,1:4]
dx0[i] <- sqrt(sum(dif*dif))
}
estW(d, dx0, iris[,5])
```

INCAindex *INCA index*

Description

INCAindex helps to estimate the number of clusters in a dataset.

Usage

```
INCAindex(d, pert_clus)
```

Arguments

d	a distance matrix or a dist object with distance information between units.
pert_clus	an n-vector that indicates which group each unit belongs to. Note that the expected values of pert are numbers greater than or equal to 1 (for instance 1,2,3,4..., k). The default value indicates the presence of only one group in data.

Value

Returns an object of class incaix which is a list containing the following components:

well_class	a vector indicating the number of well classified units.
Ni_cluster	a vector indicating each cluster size.
Total	percentage of objects well classified in the partition defined by pert_clus.

Note

For a correct geometrical interpretation it is convenient to verify whether the distance matrix d is Euclidean. It admits the associated methods summary and plot. The first simply returns the percentage of well-classified units and the second offers a barchart with the percentages of well classified units for each group in the given partition.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

Irigoien, I. and Arenas, C. (2008). INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**(15), 2948–2973.

See Also

estW, INCAtest

Examples

```
#generate 3 clusters, each of them with 20 objects in dimension 5.
mu1 <- sample(1:10, 5, replace=TRUE)
x1 <- matrix(rnorm(20*5, mean = mu1, sd = 1),ncol=5, byrow=TRUE)
mu2 <- sample(1:10, 5, replace=TRUE)
x2 <- matrix(rnorm(20*5, mean = mu2, sd = 1),ncol=5, byrow=TRUE)
mu3 <- sample(1:10, 5, replace=TRUE)
x3 <- matrix(rnorm(20*5, mean = mu3, sd = 1),ncol=5, byrow=TRUE)
x <- rbind(x1,x2,x3)

# Euclidean distance between units.
d <- dist(x)

# given the right partition, calculate the percentage of well classified objects.
partition <- c(rep(1,20), rep(2,20), rep(3,20))
INCAindex(d, partition)


# In order to estimate the number of cluster in data, try several
#  partitions and compare the results
library(cluster)
T <- rep(NA, 5)
for (l in 2:5){
part <- pam(d,l)$clustering
T[l] <- INCAindex(d,part)$Total
```

```
}

plot(T, type="b",xlab="Number of clusters", ylab="INCA", xlim=c(1.5, 5.5))
```

INCAnumclu *Estimation of Number of Clusters in Data*

Description

INCAnumclu helps to estimate the number of clusters in a dataset. The INCA index associated to different partitions with different number of clusters is calculated.

Usage

```
INCAnumclu(d, K, method = "pam", pert, L= NULL, noise=NULL)
```

Arguments

d	a distance matrix or a dist object with distance information between units.
K	the maximum number of cluster to be considered. For each k value (k=2,..,K) a partition with k clusters is calculated.
method	character string defining the clustering method in order to obtain the partitions. The hierarchical aglomerative clustering methods are perfomed via hclust function in package **fastcluster**. Other clustering methods are performed via the functions in package **cluster**, such as: pam, diana and fanny. The available clustering methods are pam (default method), average (UPGMA), single (single linkage), complete (complete linkage), ward.D2 (Ward's method), ward.D, centroid, median, diana (hierarchical divisive) and fanny (fuzzy clustering). Nevertheless, the user can introduce particular or custom partitions indicating method="partition" and specifying the partitions in argument pert.
pert	only useful when parameter method="partition"; it is a matrix and each column contains a partition of the units. That means that each column is an n-vector that indicates which group each unit belongs to. Note that the expected values of each column of pert are numbers greater than or equal to 1 (for instance 1,2,3,4..., k).
L	default value NULL, but when some units are considered by the user as noise units, L must be specified as follows: (a) L is greater than or equal to 1 and all units in clusters with a cardinal <= L are considered noise units; (b) L="custom" when the user wants to specify which units are considered noise units. These units must be specified in argument noise.
noise	when L="custom", it is a logical vector indicating the units considered by the user as noise units.

Value

Returns an object of class `incanc` which is a numeric vector containing the INCA index associated to each of the k (k=2,...,K) partitions. When `noise` is no null, the function returns a list with the INCA index for each partition, which is calculated without noise units as well as with noise units. The associated `plot` returns INCA index plot, both, with and without noise.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Irigoien, I. and Arenas, C. (2008). INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**(15), 2948–2973.

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

See Also

`INCAindex`, `estW`

Examples

```
#------- Example 1 -------------------------------------
#generate 3 clusters, each of them with 20 objects in dimension 5.
mu1 <- sample(1:10, 5, replace=TRUE)
x1 <- matrix(rnorm(20*5, mean = mu1, sd = 1),ncol=5, byrow=TRUE)
mu2 <- sample(1:10, 5, replace=TRUE)
x2 <- matrix(rnorm(20*5, mean = mu2, sd = 1),ncol=5, byrow=TRUE)
mu3 <- sample(1:10, 5, replace=TRUE)
x3 <- matrix(rnorm(20*5, mean = mu3, sd = 1),ncol=5, byrow=TRUE)
x <- rbind(x1,x2,x3)

# calculte euclidean distance between them
d <- dist(x)

# calculate the INCA index associated to partitions with k=2, ..., k=5 clusters.
INCAnumclu(d, K=5)
out <- INCAnumclu(d, K=5)
plot(out)

#------- Example 1 cont. -------------------------------
# With hypothetical noise elements
noiseunits <- rep(FALSE, 60)
noiseunits[sample(1:60, 20)] <- TRUE
out <- INCAnumclu(d, K=5, L="custom", noise=noiseunits)
plot(out)
```

INCAtest *INCA Test*

Description

Assume that n units are divided into k groups C1,...,Ck. Function `INCAtest` performs the typicality INCA test. Therein, the null hypothesis that a new unit x0 is a typical unit with respect to a previously fixed partition is tested versus the alternative hypothesis that the unit is atypical.

Usage

```
INCAtest(d, pert, d_test, np = 1000, alpha = 0.05, P = 1)
```

Arguments

`d`	a distance matrix or a `dist` object with distance information between units.
`pert`	an n-vector that indicates which group each unit belongs to. Note that the expected values of `pert` are numbers greater than or equal to 1 (for instance 1,2,3,4..., k). The default value indicates there is only one group in data.
`d_test`	an n-vector containing the distances from x0 to the other units.
`np`	sample size for the bootstrap sample for the bootstrap procedure.
`alpha`	fixed level for the test.
`P`	Number of times the bootstrap procedure is repeated.

Value

A list with class "incat" containing the following components:

`StatisticW0`	value of the INCA statistic.
`ProjectionsU`	values of statistics measuring the projection from the specific object to each considered group.
`pvalues`	p-values obtained in the `P` times repeated bootstrap procedure. Note: If `P`>1, it is printed the number of times the p-values were smaller than `alpha`.
`alpha`	specified value of the level of the test.

Note

To obtain the INCA statistic distribution, under the null hypothesis, the program can consume long time. For a correct geometrical interpretation it is convenient to verify whether the distance matrix d is Euclidean.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.es>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV-EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Irigoien, I. and Arenas, C. (2008). INCA: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, **27**(15), 2948–2973.

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

See Also

estW, INCAindex

Examples

```
#generate 3 clusters, each of them with 20 objects in dimension 5.
mu1 <- sample(1:10, 5, replace=TRUE)
x1 <- matrix(rnorm(20*5, mean = mu1, sd = 1),ncol=5, byrow=TRUE)
mu2 <- sample(1:10, 5, replace=TRUE)
x2 <- matrix(rnorm(20*5, mean = mu2, sd = 1),ncol=5, byrow=TRUE)
mu3 <- sample(1:10, 5, replace=TRUE)
x3 <- matrix(rnorm(20*5, mean = mu3, sd = 1),ncol=5, byrow=TRUE)
x <- rbind(x1,x2,x3)

# Euclidean distance between units in matrix x.
d <- dist(x)
# given the right partition
partition <- c(rep(1,20), rep(2,20), rep(3,20))

# x0 contains a unit from one group, as for example group 1.
x0 <-  matrix(rnorm(1*5, mean = mu1, sd = 1),ncol=5, byrow=TRUE)

# distances between x0 and the other units.
dx0 <- rep(0,60)
for (i in 1:60){
dif <-x0-x[i,]
dx0[i] <- sqrt(sum(dif*dif))
}

INCAtest(d, partition, dx0, np=10)


# x0 contains a unit from a new group.
x0 <-  matrix(rnorm(1*5, mean = sample(1:10, 5, replace=TRUE),
        sd = 1), ncol=5, byrow=TRUE)

# distances between x0 and the other units in matrix x.
dx0 <- rep(0,60)
for (i in 1:60){
dif <-x0-x[i,]
dx0[i] <- sqrt(sum(dif*dif))
}

INCAtest(d, partition, dx0, np=10)
```

`lympha` *Limphatic Database*

Description

This lymphography domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. The data are available at the UCI KDD data repository (Hettich S and Bay SD, 1999).

The data set consists of 148 instances presenting 18 different mixed attributes (1 cuantitative, 9 binaries and 9 nominals), and a class variable indicating the diagnostic. There are not missing values.

Usage

```
data(lympha)
```

Format

Data frame with 148 instances and 19 features.

Details

Attribute information:

— NOTE: All attribute values in the database have been entered as numeric values corresponding to their index in the list of attribute values for that attribute domain as given below.

1. class: normal find, metastases, malign lymph, fibrosis
2. lymphatics: normal, arched, deformed, displaced
3. block of affere: no, yes
4. bl. of lymph. c: no, yes
5. bl. of lymph. s: no, yes
6. by pass: no, yes
7. extravasates: no, yes
8. regeneration of: no, yes
9. early uptake in: no, yes
10. lym.nodes dimin: 0-3
11. lym.nodes enlar: 1-4
12. changes in lym.: bean, oval, round
13. defect in node: no, lacunar, lac. marginal, lac. central
14. changes in node: no, lacunar, lac. margin, lac. central
15. changes in stru: no, grainy, drop-like, coarse, diluted, reticular, stripped, faint

16. special forms: no, chalices, vesicles
17. dislocation of: no, yes
18. exclusion of no: no, yes
19. no. of nodes in: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70

Source

The UCI KDD Archive.

References

Hettich S and Bay SD (1999). The UCI KDD Archive. Department of Information and Computer Science. University of California at Irvine, Irvine, USA.

Examples

```
data(lympha)
aux <- table(lympha[,1])
barplot(aux, names.arg=c("normal", "metastases", "malign lymph", "fibrosis"))
```

`proxi` *Proximity Function*

Description

Assume that n units are divided into k groups C1,...,Ck. The function calculates the proximity function from a specific unit x0 to the groups Cj.

Usage

```
proxi(d, dx0, pert = "onegroup")
```

Arguments

`d`	a distance matrix or a `dist` object with distance information between units.
`dx0`	an n-vector containing the distances from x0 to the other units.
`pert`	an n-vector that indicates which group each unit belongs to. Note that the expected values of `pert` are numbers greater than or equal to 1 (for instance 1,2,3,4..., k). The default value indicates there is only one group in data.

Value

k-vector containing the proximity function value from x0 to each group.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

Cuadras, C.M., Fortiana, J. and Oliva, F. (1997). The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification*, **14**, 117–136.

See Also

vgeo, deltas

Examples

```
data(iris)
d <- dist(iris[,1:4])

# xo contains a unit from one group, as for example group 1.
x0 <- c(5.3, 3.6, 1.1, 0.1)
# distances between x0 and the other units.
dx0 <- rep(0,150)
for (i in 1:150){
dif <-x0-iris[i,1:4]
dx0[i] <- sqrt(sum(dif*dif))
}

proxi(d, dx0, iris[,5])


# xo contains a unit from one group, as for example group 2.
x0 <- c(6.4, 3.0, 4.8, 1.3)
# distances between x0 and the other units.
dx0 <- rep(0,150)
for (i in 1:150){
dif <-x0-iris[i,1:4]
dx0[i] <- sqrt(sum(dif*dif))
}

proxi(d, dx0, iris[,5])
```

SyntheticTimeCourse *Synthetic Time Course data*

Description

Sythetic time course data where 210 genes profiles along 6 time points are reported and where the genes are drawn from 8 different populations.

Usage

```
data(SyntheticTimeCourse)
```

Format

Data frame with 120 rows and 7 columns.

Details

Attribute information: Column cl: the class that the gen belongs to. Columns t1 - t6: gene's expression along the t1, ..., t6 time points considered.

Examples

```
data(SyntheticTimeCourse)
x <- SyntheticTimeCourse[, 2:7]
cl <- SyntheticTimeCourse[, 1]
par(mfrow=c(3,3))
for (g in 1:8){
   xx <- t(x[cl==g,] )
   yy <- matrix(c(1:6 ), nrow=6, ncol=15, byrow=FALSE)
   matplot(yy,xx,  pch=21, type="b", axes=FALSE,
       ylim=c(0,3.5), xlim=c(0.5,6.5), xlab="", ylab="", col="black", main=paste("G",g))
   abline(h=0)
   abline(v=0.5)
   mtext("Time", side=1)
   mtext("Expression", side=2)
}
```

vgeo *Geometric Variability*

Description

Assume that n units are divided into k groups C1,...,Ck. The function calculates the geometrical variability for each group in data.

Usage

```
vgeo(d, pert = "onegroup")
```

Arguments

d	a distance matrix or a dist object with distance information between units.
pert	an n-vector that indicates which group each unit belongs to. Note that the expected values of pert are numbers greater than or equal to 1 (for instance 1,2,3,4..., k). The default value indicates there is only one group in data.

Value

It is a matrix containing the geometric variability for each group.

Author(s)

Itziar Irigoien <itziar.irigoien@ehu.eus>; Konputazio Zientziak eta Adimen Artifiziala, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia, Spain.

Conchita Arenas <carenas@ub.edu>; Departament d'Estadistica, Universitat de Barcelona, Barcelona, Spain.

References

Arenas, C. and Cuadras, C.M. (2002). Some recent statistical methods based on distances. *Contributions to Science*, **2**, 183–191.

Cuadras, C.M. (1992). Some examples of distance based discrimination. *Biometrical Letters*, **29**(1), 3–20.

See Also

deltas, proxi

Examples

```
data(iris)
d <- dist(iris[,1:4])
vgeo(d,iris[,5])
```

Index