Package 'ACTCD'

December 17, 2016

Type Package

Title Asymptotic Classification Theory for Cognitive Diagnosis

Version 1.1-0

Date 2016-12-16

Author Chia-Yi Chiu (Rutgers, the State University of New Jersey) and Wenchao Ma (Rutgers, the State University of New Jersey)

Maintainer Wenchao Ma <wenchao.ma@rutgers.edu>

Depends R (>= 3.1.0), R.methodsS3

Imports stats, utils

LazyLoad yes

LazyData no

Description Cluster analysis for cognitive diagnosis based on the Asymptotic Classification Theory (Chiu, Douglas & Li, 2009; <doi:10.1007/s11336-009-9125-0>). Given the sample statistic of sum-scores, cluster analysis techniques can be used to classify examinees into latent classes based on their attribute patterns. In addition to the algorithms used to classify data, three labeling approaches are proposed to label clusters so that examinees' attribute profiles can be obtained.

License GPL (>= 2)

NeedsCompilation yes

RoxygenNote 5.0.1

Repository CRAN

Date/Publication 2016-12-17 10:48:08

R topics documented:

ACTCD-package *ACTCD: Asymptotic Classification Theory for Cognitive Diagnosis*

Description

Cluster analysis for cognitive diagnosis based on the Asymptotic Classification Theory (Chiu, Douglas & Li, 2009).

Details

Package:	ACTCD
Type:	Package
Version:	1.0-0
Date:	2013-10-21
License:	GPL (>= 2.0)
Depends:	R (>= 2.15.1), R.methodsS3

Cognitive Diagnosis aims primarily to obtain examinees' mastery or non-mastery on a set of attributes or skills of interest, based on their responses to test items and a pre-specified Q-matrix (Tatsuoka, 1985). The Asymptotic Classification Theory (Chiu, Douglas & Li, 2009) provides mathematical grounds for cognitive diagnosis using cluster analysis.

Briefly speaking, given the responses of N examinees to a test of J items with K attributes, let $\boldsymbol{W} = (W_1, W_2, \ldots, W_K)'$ be a vector of summed scores on the K attributes, where the k_{th} component is defined as

$$W_k = \sum_{j=1}^{J} Y_j q_{jk},$$

where Y_j is the vector of responses of the j^{th} examinee and q_{jk} is the (j, k) entry of the Q-matrix. The sample statistic $\boldsymbol{W}$ is then taken as the input for cluster analysis, such as *K*-means (MacQueen, 1976) and Hierarchical Agglomerative Cluster Analysis (HACA; Hartigan, 1975). This theory indicated that given two different attribute patterns $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, the corresponding conditional expectations, $E[\boldsymbol{W}|\boldsymbol{\alpha}]$ and $E[\boldsymbol{W}|\boldsymbol{\alpha}^*]$, will be distinct, implying that with $\boldsymbol{W}$ as the input, cluster analysis will group subjects correctly as the number of items is sufficiently large. Refer to Chiu, Douglas and Li (2009) for details about this theory. Because cluster analysis does not provide labels for the clusters, various labeling methods (Chiu & Ma, 2013) have been developed to obtain the attribute profiles .

The package ACTCD is an easy-to-use tool. The responses matrix and Q-matrix (Tatsuoka, 1985) are required by the main function of this package, npar.CDM, and the examinees' attribute profiles can be obtained directly using user-specified clustering and labeling methods. It is also possible to

conduct cluster analysis without labeling algorithm using function `cd.cluster` based on HACA or *K*-means. The labeling algorithms can be employed by `labeling` separately.

Author(s)

Chia-Yi Chiu (Rutgers, the State University of New Jersey) and Wenchao Ma (Rutgers, the State University of New Jersey).

Maintainer: Wenchao Ma <wenchao.ma@rutgers.edu>

References

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika, 74*(4), 633-665.

Chiu, C. Y., & Ma, W. (2013). *Assignment of clusters to attribute profiles for cognitive diagnosis*. Manuscript in preparation.

Hatigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

MacQueen, J. (1967). Some methods of classification and analysis of multiariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Bekeley Symposium on Mathematical Statistics and Probability* (pp.281-307).Berkeley: University of California Press.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

`alpha` *All possible attribute patterns*

Description

This function is used to generate all possible attribute patterns given the number of attributes.

Usage

```
alpha(K)
```

Arguments

`K`	The number of attributes.

Value

A $2^K \times K$ binary matrix is returned, 1 representing mastery of the attributes and 0 representing non-mastery of the attributes.

See Also

`eta`

Examples

```
# Generate all possible attribute patterns given the number of attributes, K.
K <- 3
A3 <- alpha(K)
K <- 4
A4 <- alpha(K)
```

`cd.cluster` *Cluster analysis for cognitive diagnosis based on the Asymptotic Classification Theory*

Description

`cd.cluster` is used to classify examinees into unlabeled clusters based on cluster analysis. Available options include *K*-means and Hierarchical Agglomerative Cluster Analysis (HACA) with various links.

Usage

```
cd.cluster (Y, Q, method = c("HACA", "Kmeans"), Kmeans.centers = NULL,
Kmeans.itermax = 10, Kmeans.nstart = 1, HACA.link = c("complete", "ward", "single",
 "average", "mcquitty", "median", "centroid"), HACA.cut = NULL)
```

Arguments

`Y`	A required $N \times J$ response matrix with binary elements (1=correct, 0=incorrect), where N is the number of examinees and J is the number of items.
`Q`	A required $J \times K$ binary item-by-attribute association matrix (Q-matrix), where K is the number of attributes. The j^{th} row of the matrix is an indicator vector, 1 indicating attributes are required and 0 indicating attributes are not required to master item j.
`method`	The clustering algorithm used to classify data. Two options are available, including `"Kmeans"` and `"HACA"`, where `"HACA"` is the default method.
`Kmeans.centers`	The number of clusters when `"Kmeans"` argument is selected. It must be not less than 2 and not greater than 2^K where K is the number of attributes. The default is 2^K.
`Kmeans.itermax`	The maximum number of iterations allowed when `"Kmeans"` argument is selected.
`Kmeans.nstart`	The number of random sets to be chosen when `"Kmeans"` argument is selected.
`HACA.link`	The link to be used with HACA. It must be one of `"ward"`, `"single"`, `"complete"`, `"average"`, `"mcquitty"`, `"median"` or `"centroid"`. The default `"HACA.link"` is `"complete"`.
`HACA.cut`	The number of clusters when `"HACA"` argument is specified. It must be not less than 2 and not greater than 2^K, where K is the number of attributes. The default is 2^K.

Details

Based on the Asymptotic Classification Theory (Chiu, Douglas & Li, 2009), A sample statistic $\boldsymbol{W}$ (See ACTCD) is calculated using the response matrix and Q-matrix provided by the users and then taken as the input for cluster analysis (i.e. K-means and HACA).

The number of latent clusters can be specified by the users in `Kmeans.centers` or `HACA.cut`. It must be not less than 2 and not greater than 2^K, where K is the number of attributes. Note that if the number of latent clusters is less than the default value (2^K), the clusters cannot be labeled in `labeling` using `method="1"` and `method="3"` algorithms. See `labeling` for more information.

Value

`W`	The $N \times K$ sample statistic $\boldsymbol{W}$ for the clustering algorithm. See details for more information.
`size`	A set of integers, indicating the sizes of latent clusters.
`mean.w`	A matrix of cluster centers, representing the average $\boldsymbol{W}$ of the latent clusters.
`wss.w`	The vector of within-cluster sum of squares of $\boldsymbol{W}$.
`sqmwss.w`	The vector of square root of mean of within-cluster sum of squares of $\boldsymbol{W}$.
`mean.y`	The vector of the mean of sum scores of the clusters.
`class`	The vector of estimated memberships for examinees.

References

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika, 74*(4), 633-665.

See Also

`print.cd.cluster`, `labeling`, `npar.CDM`, `ACTCD`

Examples

```
# Classification based on the simulated data and Q matrix
data(sim.dat)
data(sim.Q)
# Information about the dataset
N <- nrow(sim.dat) #number of examinees
J <- nrow(sim.Q) #number of items
K <- ncol(sim.Q) #number of attributes

#the default number of latent clusters is 2^K
cluster.obj <- cd.cluster(sim.dat, sim.Q)
#cluster size
sizeofc <- cluster.obj$size
#W statistics
W <- cluster.obj$W

#User-specified number of latent clusters
M <- 5  # the number of clusters is fixed to 5
```

```
cluster.obj <- cd.cluster(sim.dat, sim.Q, method="HACA", HACA.cut=M)
#cluster size
sizeofc <- cluster.obj$size
#W statistics
W <- cluster.obj$W

M <- 5 # the number of clusters is fixed to 5
cluster.obj <- cd.cluster(sim.dat, sim.Q, method="Kmeans", Kmeans.centers =M)
#cluster size
sizeofc <- cluster.obj$size
#W statistics
W <- cluster.obj$W
```

`eta` *Ideal Response Patterns for all possible attribute profiles*

Description

This function is used to calculate ideal response patterns for all possible attribute profiles based on the DINA model (Junker & Sijtsma, 2001) or conjunctive-type cognitive diagnostic models.

Usage

```
eta(K, J, Q)
```

Arguments

`K`	The number of attributes.
`J`	The number of items.
`Q`	A required $J \times K$ binary item-by-attribute association matrix (Q-matrix), where K is the number of attributes. The j^{th} row of the matrix is an indicator vector, 1 indicating attributes are required and 0 indicating attributes are not required to master item j.

Value

A $2^K \times J$ binary matrix will be returned. Each row of ideal response patterns is corresponding to each of the 2^K possible attribute patterns, which can be obtained from `alpha`.

References

Junker, B., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement, 25*(3), 258-272.

See Also

`alpha`

Examples

```
# Generating ideal response patterns
data(sim.Q)
K <- ncol(sim.Q)
J <- nrow(sim.Q)
IRP <- eta(K, J, sim.Q)
```

`labeling` *labeling for clusters*

Description

This function is used to label the clusters obtained from `cd.cluster`.

Usage

```
labeling(Y, Q, cd.cluster.object, method = c("2b","2a","1","3"),perm=NULL)
```

Arguments

`Y`	A required $N \times J$ response matrix with binary elements (1=correct, 0=incorrect), where N is the number of examinees and J is the number of items.
`Q`	A required $J \times K$ binary item-by-attribute association matrix (Q-matrix), where K is the number of attributes. The j^{th} row of the matrix is an indicator vector, 1 indicating attributes are required and 0 indicating attributes are not required to master item j.
`cd.cluster.object`	An object of `cd.cluster`.
`method`	The algorithm used for labeling. It should be one of "1","2a", "2b" and "3" corresponding to four different labeling methods in Chiu and Ma (2013). The default is "2b". See details for more information.
`perm`	The data matrix of the partial orders of the attribute patterns.

Details

Because cluster analysis such as K-means or HACA can only classify examinees into unlabeled clusters, labeling algorithms are needed to identify the underlying attribute patterns of each latent cluster. Four labeling algorithms proposed in Chiu and Ma (2013) can be implemented using this function.

The first method is the Inconsistency Index method (`method="1"`). The Inconsistency Index, IC, quantifies the amount of deviation of an ordering of clusters due to a specific $\boldsymbol{W}$ (See details in `cd.cluster`) from an arrangement of clusters that is suggested by simple assumptions about the (possible) underlying model. Among all feasible assignments of attribute patterns to clusters, the one that minimizes IC is chosen. Refer to Chiu and Ma (2013) for details. Note that this method appears to be more time-consuming when K is large and thus only the cases of $K = 3$ and $K = 4$

are implemented in the current function. To implement this algorithm, the partial order matrix of the attribute patterns should be provided. See `perm` for details.

For `method="2a"` and `method="2b"`, the label of a latent class is obtained by minimizing the average distance between observed responses and ideal responses. Specifically, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_J)$ be the observed response pattern for a particular examinee and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_J)$ be the ideal response pattern corresponding to a particular attribute pattern $\boldsymbol{\alpha}$. The Weighted Hamming distance d between $\boldsymbol{y}$ and $\boldsymbol{\eta}$ is given by

$$d(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \frac{1}{\bar{p_j}(1-\bar{p_j})} |y_j - \eta_j|.$$

where $\bar{p_j}$ denotes the proportion correction on the j^{th} item. Then the best label or attribute pattern ($\hat{\boldsymbol{\alpha}}$) can be obtained through

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha}_k \in \Omega} D.$$

where D is the average weighted Hamming distance within each cluster and Ω is the set of $\boldsymbol{\alpha}$. In practice, the largest cluster will be labeled first and the smallest cluster will be labeled last.

For `method="2a"`, The selected label $\boldsymbol{\alpha}$ will be eliminated from Ω after each labeling iteration, implying that different clusters will obtain different labels.

For `method="2b"`, The selected label $\boldsymbol{\alpha}$ will not be eliminated from Ω after each labeling iteration, implying that different clusters may obtain the same label.

For `method="3"`, it combines the technique of the partial order and `"2a"` method such that some labels can be eliminated from Ω before each labeling iteration. Refer to Chiu and Ma (2013) for details.

It should be noted that `method` `"1"`, `"2a"` and `"3"` all assume that different latent clusters are distinct in nature, which means different clusters will be given different labels using these methods. But `method` `"2b"` relaxes this assumption and allow the same label for different clusters. In addition, `method` `"1"` and `"3"` may be used when number of clusters is 2^K only. If it is not the case, `method` `"2a"` or `method` `"2b"` should be used.

Value

`att.pattern`	A $N \times K$ binary attribute patterns, where N is the number of examinees and K is the number of attributes.
`att.dist`	A $2^K \times 2$ data frame, where the first column is the attribute pattern, the second column is its frequency.

References

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika, 74*(4), 633-665.

Chiu, C. Y., & Ma, W. (2013). *Assignment of clusters to attribute profiles for cognitive diagnosis*. Manuscript in preparation.

See Also

`print.labeling`, `cd.cluster`, `npar.CDM`

Examples

```
#Labeling based on simulated data and Q matrix
data(sim.dat)
data(sim.Q)

# Information about the dataset
N <- nrow(sim.dat) #number of examinees
J <- nrow(sim.Q) #number of items
K <- ncol(sim.Q) #number of attributes

# Assume 2^K latent clusters
cluster.obj <- cd.cluster(sim.dat, sim.Q)
# Different clusters may have the same attribute patterns
labeled.obj.2b <- labeling(sim.dat, sim.Q, cluster.obj, method="2b")
# Different clusters mhave different attribute patterns
labeled.obj.2a <- labeling(sim.dat, sim.Q, cluster.obj, method="2a")
# labeling using method 1
data(perm3)  #since the number of attributes in this example is 3, perm3 is used here
labeled.obj.1 <- labeling(sim.dat, sim.Q, cluster.obj, method="1",perm=perm3)
remove(perm3) #remove perm3

# Assume 5 attribute patterns exist
M <- 5
cluster.obj <- cd.cluster(sim.dat, sim.Q, method="HACA", HACA.cut=M)
labeled.obj <- labeling(sim.dat, sim.Q, cluster.obj, method="2b")
```

`npar.CDM` *Main function for ACTCD package*

Description

This function is used to classify examinees into labeled classes given responses and the Q-matrix.

Usage

```
npar.CDM(Y, Q, cluster.method = c("HACA", "Kmeans"), Kmeans.centers = NULL,
Kmeans.itermax = 10, Kmeans.nstart = 1, HACA.link = c("complete", "ward", "single",
 "average", "mcquitty", "median", "centroid"), HACA.cut = NULL, label.method =
c("2b", "2a", "1", "3"),perm=NULL)
```

Arguments

`Y`	A required $N \times J$ response matrix with binary elements (1=correct, 0=incorrect), where N is the number of examinees and J is the number of items.
`Q`	A required $J \times K$ binary item-by-attribute association matrix (Q-matrix), where K is the number of attributes. The j^{th} row of the matrix is an indicator vector, 1 indicating attributes are required and 0 indicating attributes are not required to master item j.

`cluster.method` The cluster algorithm used to classify data. Two options are available, including `"Kmeans"` and `"HACA"`, where `"HACA"` is the default method. See `cd.cluster` for details.

`Kmeans.centers` The number of clusters when `"Kmeans"` argument is selected. It must be not less than 2 and not greater than 2^K where K is the number of attributes. The default is 2^K.

`Kmeans.itermax` The maximum number of iterations allowed when `"Kmeans"` argument is selected.

`Kmeans.nstart` The number of random sets to be chosen when `"Kmeans"` argument is selected.

`HACA.link` The link to be used with HACA. It must be one of `"ward"`, `"single"`, `"complete"`, `"average"`, `"mcquitty"`, `"median"` or `"centroid"`. The default is `"complete"`.

`HACA.cut` The number of clusters when `"HACA"` argument is selected. It must be not less than 2 and not greater than 2^K, where K is the number of attributes. The default is 2^K.

`label.method` The algorithm used for labeling. It should be one of "1","2a", "2b" and "3" corresponding to different labeling methods in Chiu and Ma (2013). The default is "2b". See `labeling` for details.

`perm` The data matrix of the partial orders of the attribute patterns.

Value

`att.pattern` A $N \times K$ binary attribute patterns, where N is the number of examinees and K is the number of attributes.

`att.dist` A $2^K \times 2$ data frame, where the first column is the attribute pattern, the second column is its frequency.

`cluster.size` A set of integers, indicating the sizes of latent clusters.

`cluster.class` A vector of estimated memberships for examinees.

See Also

`print.npar.CDM`, `cd.cluster`,`labeling`

Examples

```
# Classification based on the simulated data and Q matrix
data(sim.dat)
data(sim.Q)
# Information about the dataset
N <- nrow(sim.dat) #number of examinees
J <- nrow(sim.Q) #number of items
K <- ncol(sim.Q) #number of attributes

# Compare the difference in results among different labeling methods
# Note that the default cluster method is HACA
labeled.obj.2a <- npar.CDM(sim.dat, sim.Q, label.method="2a")
labeled.obj.2b <- npar.CDM(sim.dat, sim.Q, label.method="2b")
labeled.obj.3 <- npar.CDM(sim.dat, sim.Q, label.method="3")
```

```
data(perm3)
labeled.obj.1 <- npar.CDM(sim.dat, sim.Q, label.method="1",perm=perm3)
remove(perm3)

#User-specified number of latent clusters
M <- 5
labeled.obj.2b <- npar.CDM(sim.dat, sim.Q, cluster.method="HACA",
HACA.cut=M, label.method="2b")
labeled.obj.2a <- npar.CDM(sim.dat, sim.Q, cluster.method="HACA",
HACA.cut=M, label.method="2a")
#The attribute pattern for each examinee
attpatt <- labeled.obj.2b$att.pattern
```

`perm.data` *The partial orders of the attribute patterns for* `labeling`

Description

The data matrix of the partial orders of the attribute patterns used for the first labeling method (`method="1"`; See `labeling` for details). Available matrices are `perm3` and `perm4` for $K = 3$ and $K = 4$, respectively.

Usage

```
data(perm3)
data(perm4)
```

Format

`perm3` is a 48×3 matrix and `perm4` is a 1680384×3 matrix.

`print.output` *The function prints outputs obtained from the functions in the package.*

Description

Print outputs generated from the functions in the package.

Usage

```
## S3 method for class 'cd.cluster'
print(x, ...)
## S3 method for class 'labeling'
print(x, ...)
## S3 method for class 'npar.CDM'
print(x, ...)
```

Arguments

x	The output from the function (The list of all outputs).
...	Other arguments.

Value

cd.cluster	The number of examinees within each cluster, membership based on cluster analysis and $\boldsymbol{W}$ (See cd.cluster for details).
labeling	The estimated attribute profiles.
npar.CDM	The estimated attribute profiles.

See Also

cd.cluster, labeling,npar.CDM

---

sim.dat *Simulated data*

Description

The data are simulated based on the DINA model (Junker & Sijtsma, 2001).

Usage

```
data(sim.dat)
```

Format

A 500×14 binary matrix.

Details

This data set contains responses of 500 examinees to 14 items and is used to demonstrate the functions in this package. The DINA model with a given Q-matrix (sim.Q) was used to generate data and the true values for guessing and slipping parameters are fixed to 0.2.

References

Junker, B., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement, 25*(3), 258-272.

Examples

```
data(sim.dat)
```

sim.Q *A complete Q-matrix used to generate* sim.dat.

Description

The Q-matrix used to generate data (sim.dat) in this package.

Usage

```
data(sim.Q)
```

Format

A 14×3 binary matrix.

Details

It is a binary Q-matrix (Tatsuoka, 1985) associating 14 items and 3 attributes.

References

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

Examples

```
data(sim.Q)
```

Index