## Value

| | |
|---|---|
| ncp | the number of components retained for the FAMD |
| criterion | the criterion (the MSEP) calculated for each number of components |

## Author(s)

Vincent Audigier <audigier@agrocampus-ouest.fr>, Francois Husson <husson@agrocampus-ouest.fr> and Julie Josse <Julie.Josse@agrocampus-ouest.fr>

## References

Audigier, V., Husson, F. & Josse, J. (2014). A principal components method to impute mixed data. Advances in Data Analysis and Classification

## See Also

imputeFAMD

## Examples

```
## Not run:
data(ozone)
result <- estim_ncpFAMD(ozone)

## End(Not run)
```

---

# estim_ncpMCA *Estimate the number of dimensions for the Multiple Correspondence Analysis by cross-validation*

---

## Description

Estimate the number of dimensions for the Multiple Correspondence Analysis by cross-validation

## Usage

```
estim_ncpMCA(don, ncp.min=0, ncp.max=5,  method = c("Regularized","EM"),
     method.cv = c("Kfold","loo"), nbsim=100, pNA=0.05, threshold=1e-4,
 verbose = TRUE)
```

## Arguments

| | |
|---|---|
| don | a data.frame with categorical variables; with missing entries or not |
| ncp.min | integer corresponding to the minimum number of components to test |
| ncp.max | integer corresponding to the maximum number of components to test |
| method | "Regularized" by default or "EM" |

| | |
|---|---|
| method.cv | "Kfold" for cross-validation or "loo" for leave-one-out |
| nbsim | number of simulations, useful only if method.cv="Kfold" |
| pNA | percentage of missing values added in the data set, useful only if method.cv="Kfold" |
| threshold | the threshold for assessing convergence |
| verbose | boolean. TRUE means that a progressbar is writtent |

## Details

For leave-one-out cross-validation (method.cv="loo"), each cell of the data matrix is alternatively removed and predicted with a MCA model using ncp.min to ncp.max dimensions. The number of components which leads to the smallest mean square error of prediction (MSEP) is retained. For the Kfold cross-validation (method.cv="Kfold"), pNA percentage of missing values is inserted at random in the data matrix and predicted with a MCA model using ncp.min to ncp.max dimensions. This process is repeated nbsim times. The number of components which leads to the smallest MSEP is retained. More precisely, for both cross-validation methods, the missing entries are predicted using the imputeMCA function, it means using it means using the regularized iterative MCA algorithm (method="Regularized") or the iterative MCA algorithm (method="EM"). The regularized version is more appropriate to avoid overfitting issues.

## Value

| | |
|---|---|
| ncp | the number of components retained for the MCA |
| criterion | the criterion (the MSEP) calculated for each number of components |

## Author(s)

Francois Husson <husson@agrocampus-ouest.fr> and Julie Josse <Julie.Josse@agrocampus-ouest.fr>

## References

Josse, J., Chavent, M., Liquet, B. and Husson, F. (2010). Handling missing values with Regularized Iterative Multiple Correspondence Analysis, Journal of Clcassification, 29 (1), pp. 91-116.

## See Also

imputeMCA

## Examples

```
## Not run:
data(vnf)
result <- estim_ncpMCA(vnf,ncp.min=0, ncp.max=5)

## End(Not run)
```

`estim_ncpPCA` *Estimate the number of dimensions for the Principal Component Analysis by cross-validation*

## Description

Estimate the number of dimensions for the Principal Component Analysis by cross-validation

## Usage

```
estim_ncpPCA(X, ncp.min = 0, ncp.max = 5, method = c("Regularized","EM"),
       scale = TRUE, method.cv = c("gcv","loo","Kfold"), nbsim = 100,
   pNA = 0.05, threshold=1e-4, verbose = TRUE)
```

## Arguments

| | |
|---|---|
| `X` | a data.frame with continuous variables; with missing entries or not |
| `ncp.min` | integer corresponding to the minimum number of components to test |
| `ncp.max` | integer corresponding to the maximum number of components to test |
| `method` | "Regularized" by default or "EM" |
| `scale` | boolean. TRUE implies a same weight for each variable |
| `method.cv` | string with the values "gcv" for generalised cross-validation, "loo" for leave-one-out or "Kfold" cross-validation |
| `nbsim` | number of simulations, useful only if method.cv="Kfold" |
| `pNA` | percentage of missing values added in the data set, useful only if method.cv="Kfold" |
| `threshold` | the threshold for assessing convergence |
| `verbose` | boolean. TRUE means that a progressbar is writtent |

## Details

For leave-one-out (loo) cross-validation, each cell of the data matrix is alternatively removed and predicted with a PCA model using ncp.min to ncp.max dimensions. The number of components which leads to the smallest mean square error of prediction (MSEP) is retained. For the Kfold cross-validation, pNA percentage of missing values is inserted and predicted with a PCA model using ncp.min to ncp.max dimensions. This process is repeated nbsim times. The number of components which leads to the smallest MSEP is retained.

For both cross-validation methods, missing entries are predicted using the imputePCA function, it means using the regularized iterative PCA algorithm (method="Regularized") or the iterative PCA algorithm (method="EM"). The regularized version is more appropriate when there are already many missing values in the dataset to avoid overfitting issues.

Cross-validation (especially method.cv="loo") is time-consuming. The generalised cross-validation criterion (method.cv="gcv") can be seen as an approximation of the loo cross-validation criterion which provides a straightforward way to estimate the number of dimensions without resorting to a computationally intensive method.