

# Package ‘cstab’

July 22, 2025

**Type** Package

**Title** Selection of Number of Clusters via Normalized Clustering  
Instability

**Version** 0.2-2

**Date** 2018-06-19

**Author** Jonas M. B. Haslbeck, Dirk U. Wulff

**Maintainer** Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com>

**Description** Selection of the number of clusters in cluster analysis using  
stability methods.

**BugReports** <https://github.com/jmbh/cstab/issues>

**License** GPL (>= 2)

**Depends** R (>= 3.1.0), Rcpp (>= 0.11.4)

**Imports** cluster, fastcluster,

**LinkingTo** Rcpp

**RoxygenNote** 6.0.1

**NeedsCompilation** yes

**Repository** CRAN

**Date/Publication** 2018-06-19 16:59:43 UTC

## Contents

|                           |   |
|---------------------------|---|
| cDistance . . . . .       | 2 |
| cluster_example . . . . . | 3 |
| cstab . . . . .           | 3 |
| cStability . . . . .      | 4 |
| lookup . . . . .          | 5 |
| plot.cstab . . . . .      | 6 |
| print.cstab . . . . .     | 6 |

|              |          |
|--------------|----------|
| <b>Index</b> | <b>7</b> |
|--------------|----------|

---

cDistance

---

Selection of number of clusters via distance-based measures

---

### Description

Selection of number of clusters via *gap statistic*, *jump statistic*, and *slope statistic*

### Usage

```
cDistance(data, kseq, method = "kmeans", linkage = "complete",
          kmIter = 10, gapIter = 10)
```

### Arguments

|         |                                                                                                                                                                                                                           |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| data    | a n x p data matrix of type numeric.                                                                                                                                                                                      |
| kseq    | a vector with considered numbers clusters $k > 1$                                                                                                                                                                         |
| method  | character string indicating the clustering algorithm. 'kmeans' for the k-means algorithm, 'hierarchical' for hierarchical clustering.                                                                                     |
| linkage | character specifying the linkage criterion, in case type='hierarchical'. The available options are "single", "complete", "average", "mcquitty", "ward.D", "ward.D2", "centroid" or "median". See <a href="#">hclust</a> . |
| kmIter  | integer specifying the the number of restarts of the k-means algorithm in order to avoid local minima.                                                                                                                    |
| gapIter | integer specifying the number of simulated datasets to compute the <i>gap statistic</i> (see Tibshirani et al., 2001).                                                                                                    |

### Value

a list with the optimal numbers of cluster determined by the *gap statistic* (Tibshirani et al., 2001), the *jump Statistic* (Sugar & James, 2011) and the *slope statistic* (Fujita et al., 2014). Along the function returns the *gap*, *jump* and *slope* for each k in kseq.

### Author(s)

Dirk U. Wulff <dirk.wulff@gmail.com> Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com>

### References

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.

Sugar, C. A., & James, G. M. (2011). Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463), 750-763,

Fujita, A., Takahashi, D. Y., & Patriota, A. G. (2014). A non-parametric method to estimate the number of clusters. *Computational Statistics & Data Analysis*, 73, 27-39.

Examples

```
## Not run:
# Generate Data from Gaussian Mixture
s <- .1
n <- 50
data <- rbind(cbind(rnorm(n, 0, s), rnorm(n, 0, s)),
              cbind(rnorm(n, 1, s), rnorm(n, 1, s)),
              cbind(rnorm(n, 0, s), rnorm(n, 1, s)),
              cbind(rnorm(n, 1, s), rnorm(n, 0, s)))
plot(data)

# Selection of Number of Clusters using Distance-based Measures
cDistance(data, kseq=2:10)

## End(Not run)
```

---

|                 |                        |
|-----------------|------------------------|
| cluster_example | <i>Cluster example</i> |
|-----------------|------------------------|

---

Description

An example, 2-dimensional dataset containing the 100 points for each of five bivariate normal distributions arranged equidistant along the outline of a circle.

Usage

```
cluster_example
```

Format

An object of class `matrix` with 500 rows and 2 columns.

Details

To inspect execute `plot(cluster_example)`.

---

|       |                                                                                     |
|-------|-------------------------------------------------------------------------------------|
| cstab | <i>cstab: Selection of number of clusters via normalized clustering instability</i> |
|-------|-------------------------------------------------------------------------------------|

---

Description

Selection of the number of clusters in cluster analysis using stability methods.

## Details

Package: cstab  
 Type: Package  
 Version: 0.01  
 Date: 2016-07-26  
 License: GPL ( $\geq 2$ )

## Author(s)

Dirk U. Wulff <dirk.wulff@gmail.com> Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com>

---

cStability

*Selection of number of clusters via clustering instability*

---

## Description

Selection of number of clusters via *model-based* or *model-free*, *normalized* or *unnormalized* clustering instability.

## Usage

```
cStability(data, kseq = 2:20, nB = 10, norm = TRUE, predict = TRUE,
  method = "kmeans", linkage = "complete", kmIter = 5, pbar = TRUE)
```

## Arguments

|         |                                                                                                                                                                                                                           |
|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| data    | a $n \times p$ data matrix of type numeric.                                                                                                                                                                               |
| kseq    | a vector with considered numbers clusters $k > 1$                                                                                                                                                                         |
| nB      | an integer specifying the number of bootstrap comparisons.                                                                                                                                                                |
| norm    | logical specifying whether the instability path should be normalized. If TRUE, the instability path is normalized, accounting for a trivial decrease in instability due to a increasing $k$ (see Haslbeck & Wulff, 2016). |
| predict | boolean specifying whether the model-based or the model-free variant should be used (see Haslbeck & Wulff, 2016).                                                                                                         |
| method  | character string specifying the clustering algorithm. 'kmeans' for the k-means algorithm, 'hierarchical' for hierarchical clustering.                                                                                     |
| linkage | character specifying the linkage criterion, in case type='hierarchical'. The available options are "single", "complete", "average", "mcquitty", "ward.D", "ward.D2", "centroid" or "median". See <a href="#">hclust</a> . |
| kmIter  | integer specifying the the number of restarts of the k-means algorithm in order to avoid local minima.                                                                                                                    |
| pbar    | logical                                                                                                                                                                                                                   |

**Value**

a list that contains the optimal k selected by the unnormalized and normalized instability method. It also includes a vector containing the averaged instability path (over bootstrap samples) and a matrix containing the instability path of each bootstrap sample for both the normalized and the unnormalized method.

**Author(s)**

Dirk U. Wulff <dirk.wulff@gmail.com> Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com>

**References**

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2001). A stability based method for discovering structure in clustered data. *Pacific symposium on biocomputing*, 7, 6-17.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3), 511-528.

**Examples**

```
## Not run:
# Generate Data from Gaussian Mixture
s <- .1
n <- 50
data <- rbind(cbind(rnorm(n, 0, s), rnorm(n, 0, s)),
              cbind(rnorm(n, 1, s), rnorm(n, 1, s)),
              cbind(rnorm(n, 0, s), rnorm(n, 1, s)),
              cbind(rnorm(n, 1, s), rnorm(n, 0, s)))
plot(data)

# Selection of Number of Clusters using Instability-based Measures
stab_obj <- cStability(data, kseq=2:10)
print(stab_obj)

## End(Not run)
```

---

lookup

---

Create lookup table

---

**Description**

Create lookup table for faculties

**Usage**

```
lookup(n = 10000L, root = 200)
```

**Arguments**

|      |                                                          |
|------|----------------------------------------------------------|
| n    | integer specifying the number of                         |
| root | numeric specifying the root used to avoid machine limit. |

---

|            |                                      |
|------------|--------------------------------------|
| plot.cstab | <i>Plot method for cstab objects</i> |
|------------|--------------------------------------|

---

**Description**

plot.cstab plots *instability* path.

**Usage**

```
## S3 method for class 'cstab'
plot(x, ...)
```

**Arguments**

|     |                                                  |
|-----|--------------------------------------------------|
| x   | a cstab object (output of functions cStability). |
| ... | additional arguments passed to print.            |

**Author(s)**

Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com> Dirk U. Wulff <dirk.wulff@gmail.com>

---

|             |                                       |
|-------------|---------------------------------------|
| print.cstab | <i>Print method for cstab objects</i> |
|-------------|---------------------------------------|

---

**Description**

print.cstab prints key variables of cstab objects.

**Usage**

```
## S3 method for class 'cstab'
print(x, ...)
```

**Arguments**

|     |                                                  |
|-----|--------------------------------------------------|
| x   | a cstab object (output of functions cStability). |
| ... | additional arguments passed to print.            |

**Author(s)**

Jonas M. B. Haslbeck <jonas.haslbeck@gmail.com> Dirk U. Wulff <dirk.wulff@gmail.com>

# Index

## \* **datasets**

cluster\_example, [3](#)

cDistance, [2](#)

cluster\_example, [3](#)

cstab, [3](#)

cstab-package (cstab), [3](#)

cStability, [4](#)

hclust, [2](#), [4](#)

lookup, [5](#)

plot.cstab, [6](#)

print.cstab, [6](#)