# Package 'kmodR'

May 12, 2022

**Type** Package

**Title** K-Means with Simultaneous Outlier Detection

**Version** 0.2.0

**Date** 2022-04-11

**Maintainer** David Charles Howe <kmodR@edgecondition.com>

**Description** An implementation of the 'k-means--' algorithm proposed by Chawla and Gionis, 2013 in their paper, ``k-means-- : A unified approach to clustering and outlier detection. SIAM International Conference on Data Mining (SDM13)", <doi:10.1137/1.9781611972832.21> and using 'ordering' described by Howe, 2013 in the thesis, Clustering and anomaly detection in tropical cyclones". Useful for creating (potentially) tighter clusters than standard k-means and simultaneously finding outliers inexpensively in multidimensional space.

**License** GPL-3

**Suggests** testthat

**Encoding** UTF-8

**RoxygenNote** 7.1.2

**NeedsCompilation** no

**Author** David Charles Howe [aut, cre] (<https://orcid.org/0000-0003-4942-1300>)

**Repository** CRAN

**Date/Publication** 2022-05-12 11:40:02 UTC

## R topics documented:

kmod *K-Means clustering with simultaneous Outlier Detection*

## Description

An implementation of the 'k-means–' algorithm proposed by Chawla and Gionis, 2013 in their paper, "k-means– : A unified approach to clustering and outlier detection. SIAM International Conference on Data Mining (SDM13)", doi: 10.1137/1.9781611972832.21 and using 'ordering' described by Howe, 2013 in the thesis, "Clustering and anomaly detection in tropical cyclones".

Useful for creating (potentially) tighter clusters than standard k-means and simultaneously finding outliers inexpensively in multidimensional space.

## Usage

```
kmod(
  X,
  k = 5,
  l = 0,
  i_max = 100,
  conv_method = "delta_C",
  conv_error = 0,
  allow_empty_c = FALSE
)
```

## Arguments

| | |
|---|---|
| X | matrix of numeric data or an object that can be coerced to such a matrix (such as a data frame with numeric columns only). |
| k | the number of clusters (default = 5) |
| l | the number of outliers (default = 0) |
| i_max | the maximum number of iterations permissible (default = 100) |
| conv_method | character: the method used to assess if kmod has converged (default = "delta_C") |
| conv_error | numeric: the tolerance permissible when assessing convergence (default = 0) |
| allow_empty_c | logical: set whether empty clusters are permissible (default = FALSE) |

## Value

kmod returns a list comprising the following components

k the number of clusters specified

l the number of outliers specified

C the set of cluster centroids

C_sizes cluster sizes

C_ss the sum of squares for each cluster

L the set of outliers

L_dist_sqr the distance squares for each outlier to C

L_index the index of each outlier in the supplied dataset

XC_dist_sqr_assign the distance square and cluster assignment of each point in the supplied dataset

within_ss the within cluster sum of squares (excludes outliers)

between_ss the between cluster sum of squares

tot_ss the total sum of squares

iterations the number of iterations taken to converge

## Examples

```
# a 2-dimensional example with 2 clusters and 5 outliers
x <- rbind(matrix(rnorm(100, sd = 0.3), ncol = 2),
           matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2))
colnames(x) <- c("x", "y")
(cl <- kmod(x, 2, 5))

# cluster a dataset with 8 clusters and 0 outliers
x <- kmod(x, 8)
```

# Index