Package ‘kpodclustr’

February 20, 2015

Title  Method for Clustering Partially Observed Data

Version  1.0

Description  The kpodclustr package implements the k-POD method for clustering partially observed data.

URL  http://jocelynchi.com/kpodclustr

Depends  R (>= 3.1.0), clues

License  MIT + file LICENSE

LazyData  true

Author  Jocelyn T. Chi [aut, cre],
        Eric C. Chi [ctb]

Maintainer  Jocelyn T. Chi <jocelynchi@alum.berkeley.edu>

NeedsCompilation  no

Repository  CRAN

Date/Publication  2014-11-26 18:33:47

R topics documented:

assign_clustpp ................................................................. 2
findMissing ................................................................. 2
initialImpute ............................................................... 3
kmpp ............................................................... 4
kpod ............................................................... 5
makeData ............................................................... 6

Index  7
assign_clustpp  
*Function for assigning clusters to rows in a matrix*

**Description**

assign_clustpp Function for assigning clusters to rows in a matrix

**Usage**

assign_clustpp(x, init_centers, kmpp_flag = TRUE, max_iter = 20)

**Arguments**

- **x**  
  Data matrix containing missing entries whose rows are observations and columns are features

- **init_centers**  
  Centers for initializing k-means

- **kmpp_flag**  
  (Optional) Indicator for whether or not to initialize with k-means++

- **max_iter**  
  (Optional) Maximum number of iterations

**Author(s)**

Jocelyn T. Chi

**Examples**

```r
p <- 2
n <- 100
k <- 3
sigma <- 0.25
missing <- 0.05
Data <- makeData(p,n,k,sigma,missing)
X <- Data$Missing
Orig <- Data$Orig

clusts <- assign_clustpp(Orig, k)
```

findMissing  
*Function for finding indices of missing data in a matrix*

**Description**

findMissing Function for finding indices of missing data in a matrix

**Usage**

findMissing(X)
**initialImpute**

**Arguments**

- **X**  
  Data matrix containing missing entries whose rows are observations and columns are features

**Value**

A numeric vector containing indices of the missing entries in X

**Author(s)**

Jocelyn T. Chi

**Examples**

```r
p <- 2
n <- 100
k <- 3
sigma <- 0.25
missing <- 0.05
Data <- makeData(p,n,k,sigma,missing)
X <- Data$Missing
missing <- findMissing(X)
```

---

**initialImpute**  
*Function for initial imputation for k-means*

**Description**

initialImpute  
Initial imputation for k-means

**Usage**

initialImpute(X)

**Arguments**

- **X**  
  Data matrix containing missing entries whose rows are observations and columns are features

**Value**

A data matrix containing no missing entries

**Author(s)**

Jocelyn T. Chi
Examples

```r
p <- 2
n <- 100
k <- 3
sigma <- 0.25
missing <- 0.05
Data <- makeData(p, n, k, sigma, missing)
X <- Data$Missing
X_copy <- initialImpute(X)
```

kmpp

---

**Description**

`kmpp` Computes initial centroids via kmeans++

**Usage**

```r
kmpp(X, k)
```

**Arguments**

- `X` Data matrix whose rows are observations and columns are features
- `k` Number of clusters.

**Value**

A data matrix whose rows contain initial centroids for the k clusters

**Examples**

```r
n <- 10
p <- 2
X <- matrix(rnorm(n*p), n, p)
k <- 3
kmpp(X, k)
```
Function for performing k-POD

Description

kpod Function for performing k-POD, a method for k-means clustering on partially observed data

Usage

kpod(x, k, kmpp_flag = TRUE, maxiter = 100)

Arguments

x Data matrix containing missing entries whose rows are observations and columns are features
k Number of clusters
kmpp_flag (Optional) Indicator for whether or not to initialize with k-means++
maxiter (Optional) Maximum number of iterations

Value

cluster: Clustering assignment obtained with k-POD
cluster_list: List containing clustering assignments obtained in each iteration
obj_vals: List containing the k-means objective function in each iteration
fit: Fit of clustering assignment obtained with k-POD (calculated as 1-(total withinss/totss))
fit_list: List containing fit of clustering assignment obtained in each iteration

Author(s)
Jocelyn T. Chi

Examples

p <- 5
n <- 200
k <- 3
sigma <- 0.15
missing <- 0.20
Data <- makeData(p,n,k,sigma,missing)
X <- Data$Missing
Orig <- Data$Orig
truth <- Data$truth

kpod_result <- kpod(X,k)
kpodclusters <- kpod_result$cluster
makeData

Make test data

Description
makeData Function for making test data

Usage
makeData(p, n, k, sigma, missing, seed = 12345)

Arguments
- p: Number of features (or variables)
- n: Number of observations
- k: Number of clusters
- sigma: Variance
- missing: Desired missingness percentage
- seed: (Optional) Seed (default seed is 12345)

Author(s)
Jocelyn T. Chi

Examples
p <- 2
n <- 100
k <- 3
sigma <- 0.25
missing <- 0.05

X <- makeData(p, n, k, sigma, missing)$Orig
Index

assign_clustpp, 2
find_missing, 2
initial_impute, 3
kmpp, 4
kpod, 5
makeData, 6