Package ‘FastHCS’

May 22, 2018

Type Package

Title FastHCS Robust Algorithm for Principal Component Analysis

Version 0.0.6

Date 2018-05-13

Depends R (>= 3.1.1), matrixStats, robustbase

Suggests mvtnorm

Imports methods

LinkingTo Rcpp, RcppEigen

SystemRequirements C++11


License GPL (>= 2)

LazyLoad yes

Author Kaveh Vakili [aut, cre]

Maintainer Kaveh Vakili <vakili.kaveh.email@gmail.com>

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-05-22 16:49:10 UTC

R topics documented:

FastHCS-package .......................... 2
compPcaParams .......................... 3
DnaAlteration .......................... 4
FastHCS ................................. 5
FHCSkernelEVD .......................... 7
FHCSnumStarts .......................... 8
FHCSpsdo ............................... 9
MultipleFeatures ........................ 10
FastHCS-package

Package implementing the FastHCS robust PCA algorithm.

Description

Uses the FastHCS algorithm to compute a robust PCA model.

Details

<table>
<thead>
<tr>
<th>Package:</th>
<th>FastHCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Package</td>
</tr>
<tr>
<td>Version:</td>
<td>0.1</td>
</tr>
<tr>
<td>Date:</td>
<td>2013-01-13</td>
</tr>
<tr>
<td>Suggests:</td>
<td>mvtnorm</td>
</tr>
<tr>
<td>License:</td>
<td>GPL (&gt;= 2)</td>
</tr>
<tr>
<td>LazyLoad:</td>
<td>yes</td>
</tr>
</tbody>
</table>

Index:

- compPcaParams: Internal function used to compute the FastHCS PCA model parameters.
- DnaAlteration: Cytosine methylation beta values for a sample of 198 non-pathological human tissue specimens.
- FastHCS: Function to compute the FastHCS outlyingness index for high-dimensional data-sets.
- FHCSnumStarts: Computes the number of starting subsets for the FastHCS algorithm.
- FHCSkernelEVD: Reduces the data space to the affine subspace spanned by the \( \text{code}(n) \) observations.
- FHCSpsdo: Computes the pseudo Stahel Donoho based PCA estimates.
- MultipleFeatures: Fourier coefficients describing the shape of many hand written replications of the numerals '0' and '1'.
- plot.FastHCS: PCA diagnostic plot for object of class FastHCS.
- quanf: Internal function used to compute the size of the h-subsets used in FastHCS based on the input parameter alpha.
- signFlip: Performs the sign flip operation on a matrix of loadings.
- Tablets: Near-infrared (NIR) spectroscopy of a sample of 310 tablets.

Author(s)

Kaveh Vakili (primary programmer), Eric Schmitt Maintainer: Kaveh Vakili <vakili.kaveh.email@gmail.com>
**compPcaParams**

Computes the center vector, eigenvalues and loading matrix corresponding to a PCA model of a data matrix with respect to a subset of observations in a data set

### Description

This function is used in FastHCS to compute the parameter estimates of the PCA models used at different steps of the algorithm. It is an internal function not intended to be called by the user.

### Usage

```r
compPcaParams(x, fitd, q=NULL, z0=NULL, seed=1)
```

### Arguments

- **x** A data matrix x.
- **fitd** The (internal) result of a call to FastHCS.
- **q** Desired rank of the SVD decomposition.
- **z0** Optional. Result of a call to FHCSkernelEVD.
- **seed** Seed used to initialize the RNG. Defaults to 1.

### Value

A list with the following components:

- **center** The multivariate mean of the observations with indexes in best.
- **loadings** The (rank q) loadings matrix of the observations with indexes in best.
- **eigenvalues** The eigenvalues of the observations with indexes in best multiplied by a consistency factor.
- **scores** The value of the projected on the space of the principal components data (the centred data multiplied by the loadings matrix) is returned. Hence, cov(scores) is the diagonal matrix diag(eigenvalues).

### Author(s)

Kaveh Vakili, Eric Schmitt
**DnaAlteration**  

_Cytosine methylation beta values for a sample of 198 non-pathological human tissue specimens._

### Description

A data frame with the subset of the 'Dna Alteration' data set corresponding to the sample of 'blood' and 'non-blood, non placenta' tissues.

### Usage

```r
DnaAlteration
```

### Format

**Labels**  
Observations with label "0" correspond to the subset of 'blood' tissues.

**Column 2–1414** Cytosine methylation beta values collected at 1413 autosomal CpG loci.

### Source


### Examples

```r
data(DnaAlteration)
alpha<-0.5
Q<-15
p<-ncol(DnaAlteration[,1])
ns<-FHCSnumStarts(q=Q,eps=(1-alpha)*4/5)
RunExample<FALSE
if(RunExample){
  Fit<-FastHCS(x=DnaAlteration[,1],q=Q,nSamp=ns,seed=0)
colvec<-rep("orange",nrow(DnaAlteration))
colvec[DnaAlteration[1,1]==1]<="blue"
plot(Fit,col=colvec,pch=16)
}
```
Performing the FastHCS algorithm for robust PCA.

Description
Computes a robust PCA model with q components for an n by p matrix of multivariate data using the FastHCS algorithm.

Usage
```
FastHCS(x,xSamp=NULL,alpha=0.5,q=10,seed=1)
```

Arguments
- **x**: A numeric n (n>5*q) by p (p>1) matrix or data frame.
- **nsamp**: A positive integer giving the number of resamples required; "nsamp" may not be reached if too many of the q-subsamples, chosen out of the observed vectors, are in a hyperplane. If "nsamp" is omitted, it is calculated to provide a breakdown point of "alpha" with probability 0.99.
- **alpha**: Numeric parameter controlling the size of the active subsets i.e., "h=quanf(alpha,n,q)". Allowed values are between 0.5 and 1 and the default is 0.5.
- **q**: Number of principal components to compute. Note that p>q>1, 1<q<n/2. Default is q=10.
- **seed**: Starting value for random generator. Default is seed = 1.

Value
A list with components:
- **rawBest**: The indexes of the h members of H*, the raw FastHCS optimal subset.
- **obj**: The FastHCS objective function corresponding to H*, the selected subset of h observations.
- **rawDist**: Outlyingness index of the data on the raw q-dimensonal subset that initialized H*.
- **best**: the indexes of the members of the H+, the FastHSC subset after the C-steps.
- **center**: the p-vector of column means of the observations with indexes in best.
- **loadings**: the (rank q) loadings matrix of the observations with indexes in best.
- **eigenvalues**: the first q) eigenvalues of the observations with indexes in best.
- **od**: the orthogonal distances of the centered data wrt to the subspace spanned by the loadings matrix.
- **sd**: the score distances of the data projected on the subspace spanned by the loadings matrix with respect to the estimated center.
- **cutoff.od**: the cutoff for the vector of orthogonal distances.
cutoff.sd: the cutoff for the vector of score distances.
scores The value of the projected on the space of the principal components data (the centred data multiplied by the loadings matrix) is returned. Hence, cov(scores) is the diagonal matrix diag(eigenvalues).

Author(s)
Kaveh Vakili, Eric Schmitt

References

Examples
```r
## testing outlier detection
n<-100
p<-30
Q<-5
set.seed(123)
x0<-matrix(rnorm(n*p),nc=p)
x0[1:30,]<-matrix(rnorm(30*p,4.5,1/100),nc=p)
z<-c(rep(0,30),rep(1,70))
nStarts<-FHCSnumStarts(q=Q,eps=0.4)
Fit<-FastHCS(x=x0,nSamp=nStarts,q=Q)
z[Fit$best]
plot(Fit,col=(!z)+1,pch=16)

## testing outlier detection, different value of alpha
n<-100
p<-30
Q<-5
set.seed(123)
x0<-matrix(rnorm(n*p),nc=p)
x0[1:20,]<-matrix(rnorm(20*p,4.5,1/100),nc=p)
z<-c(rep(0,20),rep(1,80))
nStarts<-FHCSnumStarts(q=Q,eps=0.25)
Fit<-FastHCS(x=x0,nSamp=nStarts,q=Q,alpha=0.75)
z[Fit$best]

#testing exact fit
n<-100
p<-5
Q<-4
set.seed(123)
x0<-matrix(rnorm(n*p),nc=p)
x0[1:30,]<-matrix(rnorm(30*p,4.5,1/100),nc=p)
x0[31:100,4:5]<-x0[31:100,2]
z<-c(rep(0,30),rep(1,70))
nStarts<-FHCSnumStarts(q=Q,eps=0.4)
results<-FastHCS(x=x0,nSamp=nStarts,q=Q)
```
FHCSkernelEVD

```
z[results$best]
results$obj

#testing rotation equivariance
n<-100
p<-10
Q<-3
set.seed(123)
x0<-scale(matrix(rnorm(n*p),nc=p))
A<-diag(rep(1,p))
A[1:2,1:2]<-c(0,1,-1,0)
x1<-x0%*%A
nStart<-FHCSnumStarts(q=Q,eps=0.4)
r0<-FastHCS(x=x0,nSamp=nStart,q=Q.seed=0)
r1<-FastHCS(x=x1,nSamp=nStart,q=Q.seed=0)
max(abs(log(r1$eigenvalues[1:Q])/r0$eigenvalues[1:Q]))
```

 FHCSkernelEVD  

Carries out the kernelEVD algorithm for data reduction

Description

This step reduces the data space to the affine subspace spanned by the n observations.

Usage

```
FHCSkernelEVD(x, best=NULL, q=NULL)
```

Arguments

- `x` A data matrix.
- `best` An optional subset of 1:n.
- `q` Desired rank of the SVD decomposition. Optional.

Value

A reduced data set with full rank.

Author(s)

Small modification of the code from the classPC from rrcov.

References

Examples

\begin{verbatim}
n<-50
p<-200
x<-matrix(rnorm(n*p),nc=p)
W<-FHCSKerneLEVD(x)
\end{verbatim}

**FHCSnumStarts**  Computes the number of starting q-subsets

Description

Computes the number of starting q-subsets to take so that there is a 99 This is an internal function not intended to be called by the user.

Usage

\begin{verbatim}
FHCSnumStarts(q,gamma=0.99,eps=0.5)
\end{verbatim}

Arguments

- **q**: Number of desired components for the PCA model.
- **gamma**: Desired probability of having at least one clean starting q-subset.
- **eps**: suspected contamination rate of the sample.

Value

An integer number of starting q-subsets.

Author(s)

Kaveh Vakili

Examples

\begin{verbatim}
FHCSnumStarts(q=3,gamma=0.99,eps=0.4)
\end{verbatim}
**FHCSpsdo**

*Computes the univariate MCD estimator of scatter*

**Description**

Pseudo Stahel Donoho Outlyingness based estimates of PCA.

**Usage**

```
FHCSpsdo(z0, h=NULL, seed=1, q=NULL, ndir = 1000)
```

**Arguments**

- `z0`  
  Either a data matrix or the result of a call to FHCSkernelEVD.

- `h`  
  Number of observation used to compute the univariate outlyingness. Defaults to \((n+q+1)/2\)+1.

- `seed`  
  Seed used to initialize the RNG. Defaults to 1.

- `q`  
  Number of components. Defaults to `ncol(z0)`.

- `ndir`  
  Number of `componentsprojections`. Defaults to 1000.

**Value**

A list with components:

- `rawdist`  
  Outlyingness index of the data on the raw q-dimensional subset that initialized H*.

- `best`  
  the indexes of the members of the H+, the FastHSC subset after the C-steps.

- `center`  
  the p-vector of column means of the observations with indexes in `best`.

- `loadings`  
  the (rank q) loadings matrix of the observations with indexes in `best`.

- `eigenvalues`  
  the first `min(q)` eigenvalues of the observations with indexes in `best`.

**Author(s)**

Vakili Kaveh.

**References**


**Examples**

```
n<-50
p<-10
x<-matrix(rnorm(n*p),nc=p)
FHCSpsdo(x)
```
MultipleFeatures | Fourier coefficients describing the shape of many hand written replications of the numerals '0' and '1'.

Description

A data frame with the subset of the 'Multiple Features' dataset corresponding to the sample of '0' and '1' numerals.

Usage

MultipleFeatures

Format

Labels Numerals.

Column 2–77 Fourier coefficients describing the shape of each observation.

Source


Examples

data(MultipleFeatures)
alpha<-0.5
Q<-15
p<-ncol(MultipleFeatures[,1])
ns<-FHCSnumStarts(q=Q,eps=(1-alpha)*4/5)
RunExample<-FALSE
if(RunExample){
  Fit<-FastHCS(x=MultipleFeatures[,1],q=Q,nSamp=ns,seed=1)
colvec<-rep("orange",nrow(MultipleFeatures))
colvec[MultipleFeatures[,1]==1]<="blue"
plot(Fit,col=colvec,pch=16)
}
plot.FastHCS

Robust diagnostic plots for FastHCS

Description

Creates a diagnostic plot of the robust SD and OD values from a FastHCS model fit, and their parametric cutoffs.

Usage

```r
## S3 method for class 'FastHCS'
plot(x, col="black", pch=16, ...)
```

Arguments

- `x` For the `plot()` method, a `FastHCS` object, typically resulting as output from `FastHCS`.
- `col` A specification for the default plotting color. Vectors of values are recycled.
- `pch` Either an integer specifying a symbol, or a single character to be used as the default in plotting points. Note that only integers and single-character strings can be set as graphics parameters. Vectors of values are recycled.
- `...` Further arguments passed to the `plot` function.

Details

This function produces the PCA diagnostic plot of Hubert et al. (2005). Score distances are the n-vector of distances of each observation to the robust estimate of location on the robust PCA subspace. Likewise, orthogonal distances are the n-vector of distances of each observations to the robust PCA subspace. The observations whose score distance is larger than cutoff.sd or whose orthogonal distance is larger than cutoff.od are considered outliers and receive a flag equal to zero. The orthogonal distances are displayed along the vertical axis and the score distances along the horizontal axis, with the dotted lines indicating their respective cut-offs.

Author(s)

Kaveh Vakili

References


See Also

`FastHCS`
Examples

data(Tables)
alpha<-0.5
Q<-15
p<-ncol(Tables[,1])
ns<-FHCSnumStarts(q=Q,eps=(1-alpha)*4/5)
RunExample<-FALSE
if(RunExample){
  Fit<-FastHCS(x=Tables[,1],q=Q,nsamp=ns,seed=1,alpha=0.5)
colvec<-rep("orange",nrow(Tables))
colvec[Tables[,1]==1]<="blue"
plot(Fit,col=colvec,pch=16)
}

signFlip

Carries out the signflip adjustment of a loadings matrix

Description

This function solves the sign indeterminacy of the loadings by setting the maximum element in a
singular vector to be positive.

Usage

signFlip(loadings)

Arguments

loadings A matrix of loadings.

Value

An (eventually sign flipped) loadings matrix.

Author(s)

Kaveh Vakili

Examples

x<-diag(10)
x[1,1]<-2
W<-signFlip(x)
W[1,1]
Near-infrared (NIR) spectroscopy of a sample of 310 tablets.

Description
The original data set contains near-infrared (NIR) spectroscopy data for 310 tablets of four different dosages from pilot, laboratory and full scale production settings are included in the study. In this subset, we combine all 80 samples of 80mg tablets with the first 50 samples of 250mg tablets.

Usage
Tablets

Format

Labels The observations with label '1' correspond to the 80mg Tablets samples and the '0' to the 250mg ones.

Column 2–405 Near Infrared Transmittance; 404 variables; 7400 to 10507 cm⁻¹.

Source

Examples
```r
data(Tables)
alpha<-.5
Q<-15
p<-ncol(Tables[,1])
ns<-FHCNumStarts(q=Q,eps=(1-alpha)*4/5)
RunExample<FALSE
if(RunExample){
  Fit<FastHCS(x=Tables[,1],q=Q,nSamp=ns,seed=1,alpha=.5)
colvec<-rep("orange",nrow(Tables))
colvec[Tables[,1]==1]<="blue"
plot(Fit,col=colvec,pch=16)
}
```
Index

*Topic **datasets**
  - DnaAlteration, 4
  - MultipleFeatures, 10
  - Tablets, 13
*Topic **multivariate**
  - compPcaParams, 3
  - FastHCS, 5
  - FHCSKernelEVD, 7
  - FHCSnumStarts, 8
  - FHCSpsdo, 9
  - plot.FastHCS, 11
  - signFlip, 12
*Topic **package**
  - FastHCS-package, 2
*Topic **plot**
  - plot.FastHCS, 11
*Topic **robust**
  - compPcaParams, 3
  - FastHCS, 5
  - FHCSKernelEVD, 7
  - FHCSnumStarts, 8
  - FHCSpsdo, 9
  - plot.FastHCS, 11
  - signFlip, 12

compPcaParams, 3
DnaAlteration, 4
FastHCS, 5, 11
FastHCS-package, 2
FHCSKernelEVD, 7
FHCSnumStarts, 8
FHCSpsdo, 9
MultipleFeatures, 10
plot.FastHCS, 11
signFlip, 12
Tablets, 13