Package ‘FastPCS’

February 19, 2015

Type Package
Title FastPCS Robust Fit of Multivariate Location and Scatter
Version 0.1.2
Date 2014-08-13
Depends R (>= 3.1.1), matrixStats
Suggests mvtnorm
LinkingTo Rcpp, RcppEigen
SystemRequirements C++11
Description The FastPCS algorithm of Vakili and Schmitt (2014) for robust estimation of multivariate location and scatter and multivariate outliers detection.
License GPL (>= 2)
LazyLoad yes
Author Kaveh Vakili [aut, cre]
Maintainer Kaveh Vakili <vakili.kaveh.email@gmail.com>
NeedsCompilation yes
Repository CRAN
Date/Publication 2015-01-07 06:26:19

R topics documented:

FastPCS-package .................................................. 2
FastPCS .......................................................... 2
FPCSnumStarts ................................................... 5
plot.FastPCS ..................................................... 6
quanf .......................................................... 7

Index 8

1
FastPCS-package

FastPCS estimator of location and scatter.

Description

Uses the FastPCS algorithm to compute the robust PCS estimator of location and scatter.

Details

Package: FastPCS
Type: Package
Version: 0.0.2
Date: 2013-01-13
Suggests: mvtnorm
License: GPL (>= 2)
LazyLoad: yes

Index:

- FastPCS: Function to compute the robust FastPCS estimator of location and scatter.
- FPCSnumStarts: Internal function used to compute the number of starting points used by FastPCS.
- quanf: Internal function used to compute h, the minimum number of observations expected to be uncontaminated.
- plot.FastPCS: Plots the robust distances outputted by a FastPCS model.

Author(s)

Kaveh Vakili [aut, cre], Maintainer: Kaveh Vakili <vakili.kaveh.email@gmail.com>

References


FastPCS

Computes the FastPCS multivariate outlyingness index.

Description

Computes a fast and robust multivariate outlyingness index for a n by p matrix of multivariate continuous data.

Usage

`FastPCS(x, nSamp, alpha=0.5, seed=1)`
Arguments

- **x**: A numeric n (n>5*p) 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 p-subsets, chosen out of the observed vectors, lie on a hyperplane. If "nSamp" is omitted, it is calculated so that the probability of getting at least one uncontaminated starting point is always at least 99 percent when there are n/2 outliers.
- **alpha**: Numeric parameter controlling the size of the active subsets, i.e., "h=quanf(alpha, n, p)". Allowed values are between 0.5 and 1 and the default is 0.5.
- **seed**: Starting value for random generator. A positive integer. Default is seed = 1.

Details

The current version of FastPCS includes the use of a C-step procedure to improve efficiency (Rousseeuw and van Driessen (1999)). C-steps are taken after the raw subset (H*) as been chosen (according to the I-index) and before reweighting. In experiments, we found that carrying C-Steps starting from the members of $rawBest$ improves the speed of convergence without increasing the bias of the final estimates. FastPCS is affine equivariant (Schmitt et al. (2014)) and thus consistent at the elliptical model (Maronna et al., (2006) p. 217).

Value

- **alpha**: The value of alpha used.
- **nSamp**: The value of nSamp used.
- **obj**: The value of the FastPCS objective function of the optimal h subset.
- **rawBest**: The index of the h observation with smallest outlyingness indexes.
- **itemBest**: The index of the observations with outlyingness smaller than the rejection threshold after C-steps are taken.
- **center**: The mean vector of the observations with outlyingness smaller than the rejection threshold after C-steps are taken.
- **cov**: Covariance matrix of the observations with outlyingness smaller than the rejection threshold after C-steps are taken.
- **distance**: The statistical distance of each observation wrt the center vector and cov matrix of the observations with outlyingness smaller than the rejection threshold after C-steps are taken.

Author(s)

Kaveh Vakili

References

FastPCS


Examples

```r
## testing outlier detection
set.seed(123)
n<-100
p<-3
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))
nstart<-FPcsNumStarts(p,p,eps=0.4)
results<-FastPCS(x=x0,nSamp=nstart)
z[results$best]

## testing outlier detection, different value of alpha
set.seed(123)
n<-100
p<-3
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))
nstart<-FPcsNumStarts(p,p,eps=0.25)
results<-FastPCS(x=x0,nSamp=nstart,alpha=0.75)
z[results$best]

# testing exact fit
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p,nc=p)
x0[1:30,]<-matrix(rnorm(30*p,5,1/100),nc=p)
x0[31:100,3]<-x0[31:100,2]*2+1
z<-c(rep(0,30),rep(1,70))
nstart<-FPcsNumStarts(p,p,eps=0.4)
results<-FastPCS(x=x0,nSamp=nstart)
z[results$rawBest]
results$obj

# testing affine equivariance
n<-100
p<-3
set.seed(123)
x0<-matrix(rnorm(n*p,nc=p)
nstart<-500
results1<-FastPCS(x=x0,nSamp=nstart,seed=1)
al1<-matrix(0.9,p,p)
```

FPCSnumStarts

Computes the number of starting p-subsets

**Description**

Computes the number of starting p-subsets so that the desired probability of selecting at least one clean one is achieved. This is an internal function not intended to be called by the user.

**Usage**

FPCSnumStarts(p, gamma=0.99, eps=0.5)

**Arguments**

- `p`: number of dimensions of the data matrix X.
- `gamma`: desired probability of having at least one clean starting p-subset.
- `eps`: suspected contamination rate of the sample.

**Value**

An integer number of starting p-subsets.

**Author(s)**

Kaveh Vakili

**Examples**

FPCSnumStarts(p=3, gamma=0.99, eps=0.4)
plot.FastPCS  

**Distance plot for FastPCS**

**Description**

Plots the robust distance values from a FastPCS model fit, and their parametric cut-off.

**Usage**

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

**Arguments**

- `x` For the `plot()` method, a FastPCS object, typically resulting as output from `FastPCS`.
- `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.

**Author(s)**

Kaveh Vakili, Eric Schmitt

**See Also**

`FastPCS`

**Examples**

```r
## generate data
set.seed(123)
n<-100
p<-3
x0<-matrix(rnorm(n*p),nc=p)
x0[1:30,]<-matrix(rnorm(30*p,4.5,1/10),nc=p)
z<-c(rep(0,30),rep(1,70))
nstart<-FPCsnumStarts(p,p,eps=0.4)
results<-FastPCS(x=x0,nSamp=nstart)
colvec<-rep("orange",length(z))
colvec[z==1]<="blue"
plot(results,col=colvec,pch=16)
```
\textit{quaf}

\begin{description}
\item[\texttt{quaf}] \textit{Converts alpha values to h-values}
\end{description}

\textbf{Description}

FastPCS selects the subset of size $h$ that minimizes the I-index criterion. The function \texttt{quaf} determines the size of $h$ based on the rate of contamination the user expects is present in the data. This is an internal function not intended to be called by the user.

\textbf{Usage}

\begin{verbatim}
quaf(n, p, alpha)
\end{verbatim}

\textbf{Arguments}

- \texttt{n} \quad Number of rows of the data matrix.
- \texttt{p} \quad Number of columns of the data matrix.
- \texttt{alpha} \quad Numeric parameter controlling the size of the active subsets, i.e., \texttt{"h=quaf(alpha,n,p)"}. Allowed values are between 0.5 and 1 and the default is 0.5.

\textbf{Value}

An integer number of the size of the starting $p$-subsets.

\textbf{Author(s)}

Kaveh Vakili

\textbf{Examples}

\begin{verbatim}
quaf(p=3, n=500, alpha=0.5)
\end{verbatim}
Index

*Topic **hplot**
  plot.FastPCS, 6

*Topic **multivariate**
  FastPCS, 2
  FPCSnumStarts, 5
  plot.FastPCS, 6
  quanf, 7

*Topic **package**
  FastPCS-package, 2

*Topic **robust**
  FastPCS, 2
  FPCSnumStarts, 5
  plot.FastPCS, 6
  quanf, 7

FastPCS, 2, 6
FastPCS-package, 2
FPCSnumStarts, 5
plot.FastPCS, 6
quanf, 7