Package 'hdpca'

January 13, 2021

Type Package

Title Principal Component Analysis in High-Dimensional Data

Version 1.1.5

Date 2021-01-13

Author Rounak Dey, Seunggeun Lee

Maintainer Rounak Dey <deyrnk@umich.edu>

Description In high-dimensional settings:
Estimate the number of distant spikes based on the Generalized Spiked Population (GSP) model.
Estimate the population eigenvalues, angles between the sample and population eigenvectors, correlations between the sample and population PC scores, and the asymptotic shrinkage factors.
Adjust the shrinkage bias in the predicted PC scores.
Dey, R. and Lee, S. (2019) <doi:10.1016/j.jmva.2019.02.007>.

Depends R (>= 3.0.0)

License GPL (>= 2)

Repository CRAN

Imports lpSolve, boot

NeedsCompilation no

Date/Publication 2021-01-13 18:40:07 UTC

R topics documented:

hapmap *Example dataset - Hapmap Phase III*

Description

The example dataset is from the Hapmap Phase III project (https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/). Our training sample consisted of unrelated individuals from two different populations: a) Utah residents with Northern and Western European ancestry (CEU), and b) Toscans in Italy (TSI). We present the eigenvalues and PC scores obtained from performing PCA on the SNPs on chromosome 7.

Format

This example dataset is a list containing the following elements:

train.eval Sample eigenvalues of the training sample.

trainscore PC scores of the training sample. This has PC1 and PC2 scores for 198 observations.

testscore We obtained the predicted scores by leaving one observation out at a time, applying PCA to the rest of the data and then predicting the PC score of the left out observation. This has PC1 and PC2 scores of 198 observations.

nSamp Number of observations in the training set = 198.

nSNP Number of SNPs on chromosome 7.

hdpc_est *High-dimensional PCA estimation*

Description

Estimates the population eigenvalues, angles between the sample and population eigenvectors, correlations between the sample and population PC scores, and the asymptotic shrinkage factors. Three different estimation methods can be used.

Usage

```
hdpc_est(samp.eval, p, n, method = c("d.gsp", "l.gsp", "osp"),
n.spikes, n.spikes.max, n.spikes.out, nonspikes.out = FALSE, smooth = TRUE)
```

Arguments

samp.eval	Numeric vector containing the sample eigenvalues. The vector must have dimension n or n-1, it may be unordered.
p	The number of features.
n	The number of samples.

method	String specifying the estimation method. Possible values are "d.gsp" (default),"l.gsp" and "osp".
n.spikes	Number of distant spikes in the population (Optional).
n.spikes.max	Upper bound of the number of distant spikes in the population. Optional, but needed if n.spikes is not specified. Ignored if n.spikes is specified.
n.spikes.out	Number of distant spikes to be returned in the output (Optional). If not specified, all the estimated distant spikes are returned.
nonspikes.out	Logical. If TRUE and method="l.gsp", the estimated set of non-spikes are returned. If TRUE and method="osp", the estimated value of the non-spike is returned.
smooth	Logical. If TRUE and method="l.gsp", kernel smoothing will be performed on the estimated population eigenvalue spectrum. Default is TRUE.

Details

The different choices for method are:

- "d.gsp": d-estimation method based on the Generalized Spiked Population (GSP) model.
- "l.gsp": λ-estimation method based on the GSP model.
- "osp": Estimation method based on the Ordinary Spiked Population (OSP) model.

At least one of n.spikes and n.spikes.max must be provided. If n.spikes is provided then n.spikes.max is ignored, else n.spikes.max is used to find out the number of distant spikes using select.nspike.

The argument nonspikes.out is ignored if method="d.gsp".

The argument smooth is useful when the user assumes the population spectral distribution to be continuous.

Value

spikes	An array of estimated distant spikes. If n.spikes.out is specified, only largest n.spikes.out many eigenvalues are returned.
n.spikes	Number of distant spikes. If n.spikes is not provided, then the estimated value is returned.
angles	An array of estimated cosines of angles between the sample and population eigenvectors corresponding to the distant spikes. The k^{th} element of the array is the estimated cosine of the angle between k^{th} sample and population eigenvectors. If n.spikes.out is specified, only first n.spikes.out many $\cos$(angle)-s are returned.
correlations	An array of estimated correlations between the sample and population PC scores corresponding to the distant spikes. The k^{th} element of the array is the estimated correlation between k^{th} sample and population PC scores. If n.spikes.out is specified, only first n.spikes.out many correlations are returned.
shrinkage	An array of estimated asymptotic shrinkage factors corresponding to the distant spikes. If n.spikes.out is specified, only first n.spikes.out many shrinkage factors are returned.

loss	If method="l.gsp", L-infinity loss function for the spectrum estimation is returned.
nonspikes	If nonspikes.out=TRUE, estimated non-spikes are returned. If λ-estimation method is used then this is a numeric vector of length p-n.spikes. If OSP model based method is used then this is a scalar number.

Author(s)

Rounak Dey, <deyrnk@umich.edu>

References

Dey, R. and Lee, S. (2019). *Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model.* Journal of Multivariate Analysis, Vol 173, 145-164.

See Also

select.nspike,pc_adjust

Examples

```
data(hapmap)
#n = 198, p = 75435 for this data

#####################################################
## Not run:
train.eval<-hapmap$train.eval
n<-hapmap$nSamp
p<-hapmap$nSNP

m<-select.nspike(train.eval,p,n,n.spikes.max=10,evals.out=FALSE)$n.spikes
out<-hdpc_est(train.eval, p, n, method = "d.gsp",
n.spikes=m, n.spikes.out=2, nonspikes.out = FALSE) #Output 2 spikes, no non-spike

out<-hdpc_est(train.eval, p, n, method = "l.gsp",
n.spikes=m, nonspikes.out = FALSE) #Output m many spikes, no non-spike

out<-hdpc_est(train.eval, p, n, method = "l.gsp",
n.spikes.max=10, nonspikes.out = TRUE) #Output all eigenvalues

out<-hdpc_est(train.eval, p, n, method = "osp",
n.spikes=m, n.spikes.out=2, nonspikes.out = TRUE) #Output m many spikes, no non-spike

## End(Not run)
```

pc_adjust *Adjusting shrinkage in PC scores*

Description

Adjusts the shrinkage bias in the predicted PC scores based on the estimated shrinkage factors.

Usage

```
pc_adjust(train.eval, p, n, test.scores, method = c("d.gsp", "l.gsp", "osp"),
n.spikes, n.spikes.max, smooth = TRUE)
```

Arguments

train.eval	Numeric vector containing the sample eigenvalues. The vector must have dimension n or n-1, it may be unordered.
p	The number of features.
n	The number of training samples.
test.scores	An $m \times k$ matrix or data frame containing the first k predicted PC scores of m many test samples.
method	String specifying the estimation method. Possible values are "d.gsp" (default),"l.gsp" and "osp".
n.spikes	Number of distant spikes in the population (Optional).
n.spikes.max	Upper bound of the number of distant spikes in the population. Optional, but needed if n.spikes is not specified. Ignored if n.spikes is specified.
smooth	Logical. If TRUE and method="l.gsp", kernel smoothing will be performed on the estimated population eigenvalue spectrum. Default is TRUE.

Details

The different choices for method are:

- "d.gsp": d-estimation method based on the Generalized Spiked Population (GSP) model.
- "l.gsp": λ-estimation method based on the GSP model.
- "osp": Estimation method based on the Ordinary Spiked Population (OSP) model.

The $(i, j)^{th}$ element of test.scores should denote the j^{th} predicted PC score for the i^{th} subject in the test sample.

At least one of n.spikes and n.spikes.max must be provided. If n.spikes is provided then n.spikes.max is ignored, else n.spikes.max is used to find out the number of distant spikes using select.nspike.

The argument nonspikes.out is ignored if method="d.gsp" or "osp".

The argument smooth is useful when the user assumes the population spectral distribution to be continuous.

Value

A matrix containing the bias-adjusted PC scores. The dimension of the matrix is the same as the dimension of `test.scores`.

A printed message shows the number of top PCs that were adjusted for shrinkage bias.

Author(s)

Rounak Dey, <deyrnk@umich.edu>

References

Dey, R. and Lee, S. (2019). *Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model.* Journal of Multivariate Analysis, Vol 173, 145-164.

See Also

`hdpc_est`, `select.nspike`

Examples

```
data(hapmap)
#n = 198, p = 75435 for this data

#####################################################
## Not run:
#First estimate the number of spikes and then adjust test scores based on that
train.eval<-hapmap$train.eval
n<-hapmap$nSamp
p<-hapmap$nSNP
trainscore<-hapmap$trainscore
testscore<-hapmap$testscore

m<-select.nspike(train.eval,p,n,n.spikes.max=10,evals.out=FALSE)$n.spikes
score.adj.o1<-pc_adjust(train.eval,p,n,testscore,method="osp",n.spikes=m)
score.adj.d1<-pc_adjust(train.eval,p,n,testscore,method="d.gsp",n.spikes=m)
score.adj.l1<-pc_adjust(train.eval,p,n,testscore,method="l.gsp",n.spikes=m)

#Or you can provide an upper bound n.spikes.max
score.adj.o2<-pc_adjust(train.eval,p,n,testscore,method="osp",n.spikes.max=10)
score.adj.d2<-pc_adjust(train.eval,p,n,testscore,method="d.gsp",n.spikes.max=10)
score.adj.l2<-pc_adjust(train.eval,p,n,testscore,method="l.gsp",n.spikes.max=10)

#Plot the training score, test score, and adjusted scores
plot(trainscore,pch=19)
points(testscore,col='blue',pch=19)
points(score.adj.o1,col='red',pch=19)
points(score.adj.d2,col='green',pch=19)

## End(Not run)
```

`select.nspike` *Finding Distant Spikes*

Description

Estimates the number of distant spikes in the population based on the Generalized Spiked Population model. A finite upper bound (`n.spikes.max`) of the number of distant spikes must be provided.

Usage

```
select.nspike(samp.eval, p, n, n.spikes.max, evals.out = FALSE, smooth = TRUE)
```

Arguments

`samp.eval`	Numeric vector containing the sample eigenvalues. The vector must have dimension `n` or `n-1`, it may be unordered.
`p`	The number of features.
`n`	The number of samples.
`n.spikes.max`	Upper bound of the number of distant spikes in the population.
`evals.out`	Logical. If `TRUE`, the estimated spikes and non-spikes are returned.
`smooth`	Logical. If `TRUE`, kernel smoothing will be performed on the estimated population eigenvalue spectrum. Default is `TRUE`.

Details

The function searches between 0 and `n.spikes.max` to find out the number of distant spikes in the population. It also estimates both non-spiked and spiked eigenvalues based on the λ-estimation method.

The argument `smooth` is useful when the user assumes the population spectral distribution to be continuous.

Value

`n.spikes`	Estimated number of distant spikes.
`spikes`	If `evals.out=TRUE`, estimated distant spikes are returned.
`nonspikes`	If `evals.out=TRUE`, estimated non-spikes are returned.
`loss`	If `evals.out=TRUE`, L-infinity loss function for the spectrum estimation is returned.

Author(s)

Rounak Dey, <deyrnk@umich.edu>

References

Dey, R. and Lee, S. (2019). *Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model.* Journal of Multivariate Analysis, Vol 173, 145-164.

See Also

hdpc_est, pc_adjust

Examples

```
data(hapmap)
#n = 198, p = 75435 for this data

#####################################################
## Not run:
#If you just want the estimated number of spikes
train.eval<-hapmap$train.eval
n<-hapmap$nSamp
p<-hapmap$nSNP

select.nspike(train.eval,p,n,n.spikes.max=10,evals.out=FALSE)

#If you want the estimated spikes and non-spikes
out<-select.nspike(train.eval,p,n,n.spikes.max=10,evals.out=TRUE)

## End(Not run)
```

Index