Package: choosepc (via r-universe)

September 18, 2024

Type Package

Title Choose the Number of Principal Components via Recistruction Error

Version 1.0

Date 2023-10-23

Author Michail Tsagris [aut, cre]

Maintainer Michail Tsagris <mtsagris@uoc.gr>

Depends R (>= 4.0)

Imports graphics, Rfast2, stats

Description One way to choose the number of principal components is via the reconstruction error. This package is designed mainly for this purpose. Graphical representation is also supported, plus some other principal component analysis related functions. References include: Jolliffe I.T. (2002). Principal Component Analysis. <doi:10.1007/b98835> and Mardia K.V., Kent J.T. and Bibby J.M. (1979). Multivariate Analysis. ISBN: 978-0124712522. London: Academic Press.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2023-10-24 18:10:02 UTC

Contents

choosepc-package *Choose the Number of Principal Components via Recistruction Error*

Description

A new robust principal component analysis algorithm is implemented that relies upon the Cauchy Distribution. The algorithm is suitable for high dimensional data even if the sample size is less than the number of variables.

Details

Package:	choosepc
Type:	Package
Version:	1.0
Date:	2023-10-22
License:	GPL-2

Maintainers

Michail Tsagris <mtsagris@uoc.gr>.

Author(s)

Michail Tsagris <mtsagris@uoc.gr>

References

Jolliffe I.T. (2002). Principal Component Analysis.

Choose the number of principal components via reconstruction error *Choose the number of principal components via reconstruction error*

Description

Choose the number of principal components via reconstruction error.

Usage

```
pc.choose(x, graph = TRUE)
```

Arguments

x	A numerical matrix with more rows than columns.
graph	Should the plot of the PRESS values appear? Default value is TRUE.

Details

SVD stands for Singular Value Decomposition of a rectangular matrix. That is any matrix, not only a square one in contrast to the Spectral Decomposition with eigenvalues and eigenvectors, produced by principal component analysis (PCA). Suppose we have a $n \times p$ matrix $\mathbf{X}$. Then using SVD we can write the matrix as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{U}$ is an orthonormal matrix containing the eigenvectors of $\mathbf{X}\mathbf{X}^T$, the $\mathbf{V}$ is an orthonormal matrix containing the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and D is a $p \times p$ diagonal matrix containing the r non zero singular values $d_1, \ldots, d_r$ (square root of the eigenvalues) of $\mathbf{X}\mathbf{X}^T$ (or $\mathbf{X}^T\mathbf{X}$) and the remaining $p - r$ elements of the diagonal are zero. We remind that the maximum rank of an $n \times p$ matrix is equal to $\min\{n, p\}$. Using the SVD decomposition equaiton above, each column of $\mathbf{X}$ can be written as

$$\mathbf{x}_j = \sum_{k=1}^{r} \mathbf{u}_k d_k \mathbf{v}_{jk}.$$

This means that we can reconstruct the matrix $\mathbf{X}$ using less columns (if $n > p$) than it has.

$$\tilde{\mathbf{x}}_j^m = \sum_{k=1}^{m} \mathbf{u}_k d_k \mathbf{v}_{jk},$$

where $m < r$.

The reconstructed matrix will have some discrepancy of course, but it is the level of discrepancy we are interested in. If we center the matrix $\mathbf{X}$, subtract the column means from every column, and perform the SVD again, we will see that the orthonormal matrix $\mathbf{V}$ contains the eigenvectors of the covariance matrix of the original, the un-centred, matrix $\mathbf{X}$.

Coming back to the a matrix of n observations and p variables, the question was how many principal components to retain. We will give an answer to this using SVD to reconstruct the matrix. We describe the steps of this algorithm below. 1. Center the matrix by subtracting from each variable its mean $\mathbf{Y} = \mathbf{X} - \mathbf{m}$

2. Perform SVD on the centred matrix $\mathbf{Y}$.

3. Choose a number from 1 to r (the rank of the matrix) and reconstruct the matrix. Let us denote by $\widetilde{\mathbf{Y}}^m$ the reconstructed matrix.

4. Calculate the sum of squared differences between the reconstructed and the original values

$$PRESS\,(m) = \sum_{i=1}^{n}\sum_{j=1}^{p} \left(\tilde{y}_{ij}^m - y_{ij}\right)^2, m = 1, .., r.$$

5. Plot $PRESS\,(m)$ for all the values of m and choose graphically the number of principal components.

The graphical way of choosing the number of principal components is not the best and there alternative ways of making a decision (see for example Jolliffe (2002)).

Value

A list including:

`values`	The eigenvalues of the covariance matrix.
`cumprop`	The cumulative proportion of the eigenvalues of the covariance matrix.
`per`	The differences in the cumulative proportion of the eigenvalues of the covariance matrix.
`press`	The reconstruction error $\sqrt{\sum_{ij}(x_{ij} - \hat{x}_{ij})^2}$ for each number of eigenvectors.
`runtime`	The runtime of the algorithm.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

References

Jolliffe I.T. (2002). Principal Component Analysis.

See Also

`eigci`

Examples

```
x <- as.matrix(iris[, 1:4])
a <- pc.choose(x, graph = FALSE)
```

`Confidence interval for the percentage of variance retained by the first k components`

Confidence interval for the percentage of variance retained by the first κ components

---

Description

Confidence interval for the percentage of variance retained by the first κ components.

Usage

```
eigci(x, k, alpha = 0.05, B = 1000, graph = TRUE)
```

Arguments

x	A numerical matrix with more rows than columns.
k	The number of principal components to use.
alpha	This is the significance level. Based on this, an $(1-\alpha)\%$ confidence interval will be computed.
B	The number of bootstrap samples to generate.
graph	Should the plot of the bootstrap replicates appear? Default value is TRUE.

Details

The algorithm is taken by Mardia Kent and Bibby (1979, pg. 233–234). The percentage retained by the fist κ principal components denoted by $\hat{\psi}$ is equal to

$$\hat{\psi} = \frac{\sum_{i=1}^{\kappa} \hat{\lambda}_i}{\sum_{j=1}^{p} \hat{\lambda}_j},$$

where $\hat{\psi}$ is asymptotically normal with mean ψ and variance

$$\tau^2 = \frac{2}{(n-1)\left(tr\mathbf{\Sigma}\right)^2}\left[(1-\psi)^2\left(\lambda_1^2+...+\lambda_k^2\right)+\psi^2\left(\lambda_{\kappa+1}^2+...\lambda_p^2\right)\right],$$

where $a = \left(\lambda_1^2+...+\lambda_k^2\right) / \left(\lambda_1^2+...+\lambda_p^2\right)$ and $\text{tr}\mathbf{\Sigma}^2 = \lambda_1^2+...+\lambda_p^2$.

The bootstrap version provides an estimate of the bias, defined as $\hat{\psi}_{boot} - \hat{\psi}$ and confidence intervals calculated via the percentile method and via the standard (or normal) method Efron and Tibshirani (1993). The funciton gives the option to perform bootstrap.

Value

A list including:

res	If B=1 (no bootstrap) a vector with the esimated percentage of variance due to the first k components, $\hat{\psi}$ and its associated asymptotic $(1-\alpha)\%$ confidence interval. If B>1 (bootstrap) a vector with: the esimated percentage of variance due to the first k components, $\hat{\psi}$, its bootstrap estimate and its bootstrap estimated bias.
ci	This appears if B>1 (bootstrap). The standard bootstrap and the empirical bootstrap $(1-\alpha)\%$ confidence interval for ψ.

Futher, if B>1 and "graph" was set equal to TRUE, a histogram with the bootstrap $\hat{\psi}$ values, the observed $\hat{\psi}$ value and its bootstrap estimate.

Author(s)

Michail Tsagris.

R implementation and documentation: Michail Tsagris <mtsagris@uoc.gr>.

References

Mardia K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis. London: Academic Press.

Efron B. and Tibshirani R. J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

See Also

pc.choose

Examples

```
x <- as.matrix(iris[, 1:4])
eigci(x, k = 2, B = 1)
```

Index