Package: pPCA (via r-universe)

October 22, 2024

Type Package

Title Partial Principal Component Analysis of Partitioned Large Sparse Matrices

Version 1.1

Date 2024-10-22

Maintainer Srika Raja <sri1919@iastate.edu>

Description Performs partial principal component analysis of a large sparse matrix. The matrix may be stored as a list of matrices to be concatenated (implicitly) horizontally. Useful application includes cases where the number of total nonzero entries exceed the capacity of 32 bit integers (e.g., with large Single Nucleotide Polymorphism data).

License GPL-3

Encoding UTF-8

Depends R (>= 3.0.2), methods, RSpectra (>= 0.16-1)

Imports Matrix (>= 1.1-0), Rcpp (>= 0.11.5)

LinkingTo Rcpp

NeedsCompilation yes

Suggests ggbiplot

RoxygenNote 7.3.2

Repository https://srika1919.r-universe.dev

RemoteUrl https://github.com/srika1919/ppca

RemoteRef HEAD

RemoteSha d10b4d5570475852556fd08cef0088e37a182f2b

Contents

pPCA *Performs a principal component analysis on a large sparse matrices or a list of large sparse matrices and returns the results as an object compatible to class prcomp*

Description

Performs a partial principal component analysis on a large sparse matrices or a list of large sparse matrices and returns the results as an object compatible to class prcomp. Uses RSpectra library to compute the largest eigenvalues.

Usage

```
pPCA(x, rank, retX = TRUE, scale. = TRUE, normalize = FALSE, sd.tol = 1e-05)
```

Arguments

x	A matrix, sparse matrix (Matrix::dgCMatrix), or a list of these. When a list is supplied, the entries are concatenated horizontally (implicitly). See description.
rank	An integer specifying the number of principal components to compute.
retX	A logical value indicating whether the rotated variables (PC scores) should be returned.
scale.	A logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place.
normalize	A logical value indicating whether the principal component scores should be normalized.
sd.tol	A positive number, warnings are printed if the standard deviation of any column is less than this threshold.

Details

When the input argument is a matrix (of class "matrix" or "dgCMatrix"), principal component analysis is performed to extract a few largest components. When a list of matrices is passed, the partial PCA is performed on the horizontally concatenated matrix, i.e., if `x = list(X1,X2,X3)` then the partial PCA is done on the matrix [X1 X2 X3], without concatenating the matrices explicitly. This can be useful when the matrix is so high-dimensional that the total number of non-zero entries exceed 2^31-1 (roughly 9.33e10), the capacity of a 32 bit integer. For example, in PCA with very high-dimensional SNP data, the sparse matrices can be stored for each chromosome within the capacity of 32 bit integers.

Value

pPCA returns a list with class "pPCA" (compatible with "prcomp") containing the following components:

sdev	A vector of the singular values (standard deviations of the principal components).

rotation	A matrix whose columns contain the eigenvectors (loadings).
x	A matrix of the principal component scores, returned if retX is true. This is the centred (and scaled if requested) data multiplied by the rotation matrix.
center	column means.
scale	column standard deviations, if scale. is true. Otherwise, FALSE.

Note

The partial SVD is computed through the RSpectra package. All elements in the first row of the rotation matrix are positive.

Author(s)

Srika Raja and Somak Dutta

References

Raja, S. and Dutta, S. (2024). Matrix-free partial PCA of partitioned genetic data. REU project 2024, Iowa State University.

Dai, F., Dutta, S., and, Maitra, R. (2020). A Matrix-Free Likelihood Method for Exploratory Factor Analysis of High-Dimensional Gaussian Data. Journal of Computational and Graphical Statistics, 29(3), 675–680.

See Also

biplot,prcomp

Examples

```
library(Matrix)
set.seed(20190329)
m <- rsparsematrix(50,100,density = 0.35)
results <- pPCA(m, rank = 2)
biplot(results)
data <- list(rsparsematrix(nrow = 50,ncol = 10,density = 0.35),
            rsparsematrix(nrow = 50,ncol = 40,density = 0.35)) # Using a list of matrices
result <- pPCA(data, rank = 3)
print(result)
biplot(result)
```

print.pPCA *Print the Output of Principal Component Analysis (pPCA)*

Description

Prints the output of the pPCA

Usage

```
## S3 method for class 'pPCA'
print(x, digits = 3, ...)
```

Arguments

x	An object of class pPCA that contains the results of a partial principal component analysis.
digits	The number of decimal places to use in printing results such as variance explained and PC scores. Defaults to 3.
...	Further arguments passed to print for additional control over the output.

Value

None.

Index