Package ‘MetaPCA’

February 19, 2015

Type Package

Title MetaPCA: Meta-analysis in the Dimension Reduction of Genomic data

Version 0.1.4

Author Don Kang <donkang75@gmail.com> and George Tseng <ctseng@pitt.edu>

Maintainer Don Kang <donkang75@gmail.com>

Description MetaPCA implements simultaneous dimension reduction using PCA when multiple studies are combined. We propose two basic ideas to find a common PC subspace by eigenvalue maximization approach and angle minimization approach, and we extend the concept to incorporate Robust PCA and Sparse PCA in the meta-analysis realm.

Depends R (>= 2.10.0), foreach

Suggests MASS, GEOquery, pcaPP, affy, hgu133plus2.db, doMC, doSMP, ellipse, impute

License GPL-2

URL https://github.com/donkang75/MetaPCA

LazyLoad yes


Date 2011-06-15

Repository CRAN

Date/Publication 2011-06-15 18:49:55

NeedsCompilation no

R topics documented:

MetaPCA-package .......................................................... 2
DropDupGenes .............................................................. 3
MetaPCA-package

Description

MetaPCA implements simultaneous dimension reduction using PCA when multiple studies are combined. We propose two basic ideas to find a common PC subspace by eigenvalue maximization approach and angle minimization approach, and we extend the concept to incorporate Robust PCA and Sparse PCA in the meta-analysis realm.

Details

Package: MetaPCA
Type: Package
Version: 0.1.4
Date: 2011-06-15
License: GPL-2
LazyLoad: yes

Author(s)

Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

References


Examples

```
# Not run:
#Spellman, 1998 Yeast cell cycle data set
#Consider each synchronization method as a separate data
data(Spellman)
pc <- list(alpha=prcomp(t(Spellman$alpha))$x, cdc15=prcomp(t(Spellman$cdc15))$x,
```
# There are currently 4 meta-pca methods. Run either one of following four.
metaPC <- MetaPCA(Spellman, method="Eigen", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="Angle", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="RobustAngle", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="SparseAngle", doPreprocess=FALSE)
# Comparing between usual pca and meta-pca
# The first lows are four data sets based on usual PCA, and
# the second rows are by MetaPCA
# We're looking for a cyclic pattern.
par(mfrow=c(2,4), cex=1, mar=c(0.2,0.2,0.2,0.2))
for(i in 1:4) {
  plot(pc[,1], pc[,2], type="n", xlab="", ylab="", xaxt="n", yaxt="n")
text(pc[,1], pc[,2], 1:nrow(pc[,1]), cex=1.5)
lines(pc[,1], pc[,2])
}
for(i in 1:4) {
  plot(metapc[,1], metapc[,2], type="n", xlab="", ylab="", xaxt="n", yaxt="n")
text(metapc[,1], metapc[,2], 1:nrow(metapc[,1]), cex=1.5)
lines(metapc[,1], metapc[,2])
}

#4 prostate cancer data which have three classes: normal, primary, metastasis
data(prostate)
# There are currently 4 meta-pca methods. Run either one of following four.
metaPC <- MetaPCA(prostate, method="Eigen", doPreprocess=FALSE, .scale=TRUE)
metaPC <- MetaPCA(prostate, method="Angle", doPreprocess=FALSE)
metaPC <- MetaPCA(prostate, method="RobustAngle", doPreprocess=FALSE)
metaPC <- MetaPCA(prostate, method="SparseAngle", doPreprocess=FALSE)
# Plotting 4 data in the same space!
coord <- foreach(dd=iter(metapc), .combine=bind) %do% dd$coord
PlotPC2D(coord[,1:2], drawEllipse=F, dataset.name="Prostate", .class.order=c("Metastasis","Primary","Normal"), .class.color=c("red","#838383","blue"), .annotation=T, newPlot=T, .class2=rep(names(metapc$x), times=sapply(metapc$x,function(x)nrow(x$coord)) ), .class2.order=names(metapc$x), .points.size=1)

# In the case of "SparseAngle" method, the top contributing genes for all studies can be determined
# For instance, top 20 genes in 1st PC and their coefficients
metapc$x[order(abs(metapc$x[,1]), decreasing=TRUE),1][1:20]

## End(Not run)
Usage

\texttt{DropDupGenes(dat, isParallel=FALSE, nCores=NULL, na.rm=TRUE)}

Arguments

\begin{itemize}
  \item \texttt{dat} \hspace{1cm} A gene expression matrix which has genes in rows and samples in columns.
  \item \texttt{isParallel} \hspace{1cm} Whether to use multiple cores in parallel for fast computing. By default, it is false.
  \item \texttt{nCores} \hspace{1cm} When \texttt{isParallel} is true, the number of cores can be set. By default, all cores in the machine are used in the unix-like machine, and 3 cores are used in windows.
  \item \texttt{na.rm} \hspace{1cm} Whether to remove genes which have no annotation. Default is TRUE.
\end{itemize}

Value

A gene expression matrix which has unique genes in rows and samples in columns.

Author(s)

Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

References


Examples

\begin{verbatim}
## Not run:
# One of example that shows how to generate a expression matrix used in the analysis
requireAll(c('GEOquery', 'affy', 'hgu133plus2.db'))
# It might be needed to download the source files first, and save it to local directory
# such as "./data/Prostate/Varambally" in this example
Varambally <- getGEO('GSE3325', destdir="./data/Prostate/Varambally")
Varambally <- Varambally[[1]]
Varambally.sLabel <- as.character(pData(Varambally)$title)
Varambally.sLabel[grep("Benign",Varambally.sLabel)] <- "Normal"
Varambally.sLabel[grep("primary",Varambally.sLabel)] <- "Primary"
Varambally.sLabel[grep("Metastatic",Varambally.sLabel)] <- "Metastasis"
Varambally <- exprs(Varambally)
colnames(Varambally) <- Varambally.sLabel
rownames(Varambally) <- unlist(mget(rownames(Varambally), hgu133plus2SYMBOL))
Varambally <- DropDupGenes(Varambally, na.rm=TRUE)
Varambally <- log2(Varambally)
## End(Not run)
\end{verbatim}
MetaPCA

**Description**

MetaPCA implements simultaneous dimension reduction using PCA when multiple studies are combined. We propose two basic ideas to find a common PC subspace by eigenvalue maximization approach and angle minimization approach, and we extend the concept to incorporate Robust PCA and Sparse PCA in the meta-analysis realm.

**Usage**

```r
MetaPCA(DList, method=c("Angle","Eigen","RobustAngle","SparseAngle"), robust.var=c("qn","mad"), nPC=1/length(DList), sparse.maxFeatures=NULL, sparse.lambda=NULL, sparse.max.iter=100, sparse.eps=1e-3, .scale=FALSE, .scaleAdjust=TRUE, doPreprocess=TRUE, cutRatioByMean=.4, cutRatioByVar=.4, doImpute=TRUE, na.rm.pct=.1, na.rm.pct.each=.5, verbose=FALSE)
```

**Arguments**

- **DList** A list of all data matrices; Each data name should be set as the name of each list element. Each data should be a numeric matrix that has genes in the rows and samples in the columns. Row names should be official gene symbols and column names be sample labels.

- **method** A vector of four meta PCA methods. The first two methods are basic approaches; the last two are extended approaches of robust PCA and sparse PCA but may be rather slower than the basic methods. Default is "Angle", which is angle minimization method. See the details in the reference.

- **robust.var** Robust measure of variance when "RobustAngle" method was selected in the method.

- **nPC** The number of returned PC’s, i.e. the number of dimension reduced by PCA.

- **.weight** Weight for each data if information is available. Default is equal weight.

- **sparse.maxFeatures** The number of genes left for the Sparse PCA approach. If NULL (default), it is determined based on the default lambda.

- **sparse.lambda** The parameter lambda which determines the sparsity of loading vectors. The default is calculated as the number of data divided by square root of the number of overall genes.

- **sparse.max.iter** The number of maximum iteration for achieving convergence of sparse loading vectors. Default is 100.

- **sparse.eps** The convergence decision precision level. Default is 1e-3.
.scale Whether to apply gene based normalization. Default is FALSE. But for the "Eigen" method, gene scaling is recommended for the comparability reason of covariance matrix.

.scaleAdjust Whether to apply scaling adjustment for a comparable visualization. Default is TRUE.

doPreprocess Whether to apply gene filtering. Default is TRUE. However "SparseAngle" method do not use gene filtering.

cutRatioByMean Proportion of genes filtered by study-wise mean. Default is 40%.

cutRatioByVar Proportion of genes filtered by study-wise variance. Default is 40%.

doImpute Whether to impute missing genes. Default is TRUE, and default imputation method is knn.

na.rm.pct Proportion of genes filtered by study-wise missing proportion. Default is 10%.

na.rm.pct.each Proportion of genes filtered by each study’s missing proportion. Default is 50%.

verbose Whether to print logs. Default is FALSE.

Value
list object having the specified number of PC’s of all data sets and loading matrix of meta subspace.

Author(s)
Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

References

Examples
```r
## Not run:
#Spellman, 1998 Yeast cell cycle data set
#Consider each synchronization method as a separate data
data(Spellman)
pc <- list(alpha=prcomp(t(Spellman$alpha))$x, cdc15=prcomp(t(Spellman$cdc15))$x, 
cdc28=prcomp(t(Spellman$cdc28))$x, elu=prcomp(t(Spellman$elu))$x)
#There are currently 4 meta-pca methods. Run either one of following four.
metaPC <- MetaPCA(Spellman, method="Eigen", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="Angle", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="RobustAngle", doPreprocess=FALSE)
metaPC <- MetaPCA(Spellman, method="SparseAngle", doPreprocess=FALSE)
#Comparing between usual pca and meta-pca
#The first lows are four data sets based on usual PCA, and
#the second rows are by MetaPCA
#We’re looking for a cyclic pattern.
par(mfrow=c(2,4), cex=1, mar=c(0.2,0.2,0.2,0.2))
for(i in 1:4) {
  plot(pc[i,1], pc[i,2], type="n", xlab="", ylab="", xaxt="n", yaxt="n")
}
text(pc[[i]][,1], pc[[i]][,2], 1:nrow(pc[[i]]), cex=1.5)
lines(pc[[i]][,1], pc[[i]][,2])
}
for(i in 1:4) {
plot(metapc$x[[i]]$coord[,1], metapc$x[[i]]$coord[,2], type="n", xlab="", ylab="", xaxt="n", yaxt="n")
text(metapc$x[[i]]$coord[,1], metapc$x[[i]]$coord[,2], 1:nrow(metapc$x[[i]]$coord), cex=1.5)
lines(metapc$x[[i]]$coord[,1], metapc$x[[i]]$coord[,2])
}

#4 prostate cancer data which have three classes: normal, primary, metastasis
data(prostate)
#There are currently 4 meta-pca methods. Run either one of following four.
metaPC <- MetaPCA(prostate, method="Eigen", doPreprocess=FALSE, .scale=TRUE)
metaPC <- MetaPCA(prostate, method="Angle", doPreprocess=FALSE)
metaPC <- MetaPCA(prostate, method="RobustAngle", doPreprocess=FALSE)
metaPC <- MetaPCA(prostate, method="SparseAngle", doPreprocess=FALSE)

#Plotting 4 data in the same space!
coord <- foreach(dd=iter(metapc$x), .combine=bind) %do% dd$coord
plotPC2D(coord[,1:2], drawEllipse=F, dataset.name="Prostate", .class.order=c("Metastasis","Primary","Normal"),
.class.color=c("red","#838383","blue"), .annotation=T, newPlot=T,
.class2=rep(names(metaPC$x), times=sapply(metaPC$x,function(x)nrow(x$coord))),
.class2.order=names(metaPC$x), .points.size=1)

#In the case of "SparseAngle" method, the top contributing genes for all studies can be determined
#For instance, top 20 genes in 1st PC and their coefficients
metaPC$v[order(abs(metaPC$v[,1]), decreasing=TRUE),1][1:20]

## End(Not run)
drawObjects     Whether to draw objects as points.
drawEllipse     Whether to draw ellipses estimated from objects 2D distribution.
dataset.name    Name to be displayed as a part of title.
pctInfo         Explained percentage of variance by each PC.
main            Main title.
sub             Sub title.
xlab            Label for x-axis.
ylab            Label for y-axis.
newPlot         Whether to draw a plot in the new frame.
.points.size    Size of objects’ points.
.class          Object’s class label such as disease classification.
.class.order    The order of class representation.
.class.color    The color of class representation.
.class2         The second class label of each object such as study name.
.class2.order   The order of 2nd class representation.
.class2.shape   The shape of 2nd class representation.
.annotation     Whether to present annotation such as x,y axis labels, legend, or titles.
.legend         Location of legend in a plot.

Value

NA. A PCA plot is drawn.

Author(s)

Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

References


Examples

```r
CC not run:
## Not run:
#4 prostate cancer data which have three classes: normal, primary, metastasis
data(prostate)
metaPC <- MetaPCA(prostate, method="Angle", doPreprocess=FALSE)
#Plotting 4 data in the same space with ellipses overlayed!
coord <- foreach(dd=iter(metaPC$x), .combine=bind) %do% dd$coord
PlotPC2D(coord[,1:2], drawEllipse=T, dataset.name="Prostate", .class.order=c("Metastasis","Primary","Normal"),
         .class.color=c('red','#838383','blue'), .annotation=T, newPlot=T,
         .class2=rep(names(metaPC$x), times=sapply(metaPC$x,function(x)nrow(x$coord))),
         .class2.order=names(metaPC$x), .points.size=1)

## End(Not run)
```
**PreprocessMetaAnalysis**

*MetaPCA: Meta-analysis in the Dimension Reduction of Genomic data*

**Description**

Preprocessing for microarray meta-analysis. It is about gene filtering and missing value imputation.

**Usage**

```r
PreprocessMetaAnalysis(DList, cutRatioByMean=0.4, cutRatioByVar=0.4, doImpute=FALSE, na.rm.pct=0.1, na.rm.pct.each=0.5, verbose=FALSE)
```

**Arguments**

- **DList**: A list of all data matrices; Each data name should be set as the name of each list element. Each data should be a numeric matrix that has genes in the rows and samples in the columns. Row names should be official gene symbols and column names be sample labels.
- **cutRatioByMean**: Proportion of genes filtered by study-wise mean. Default is 40%.
- **cutRatioByVar**: Proportion of genes filtered by study-wise variance. Default is 40%.
- **doImpute**: Whether to impute missing genes. Default is TRUE, and default imputation method is knn.
- **na.rm.pct**: Proportion of genes filtered by study-wise missing proportion. Default is 10%.
- **na.rm.pct.each**: Proportion of genes filtered by each study’s missing proportion. Default is 50%.
- **verbose**: Whether to print logs. Default is FALSE.

**Value**

list object of all data matrices after filtering and imputation.

**Author(s)**

Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

**References**


**Examples**

```r
## Not run:
DList <- PreprocessMetaAnalysis(list(Yu=Yu, Lapointe=Lapointe, Tomlins=Tomlins, Varambally=Varambally),
cutRatioByMean=0.1, cutRatioByVar=0.1, doImpute=T, na.rm.pct=0.2)
str(DList)
## End(Not run)
```
4 prostate cancer studies comparing three classes: normal, primary, metastasis.

<table>
<thead>
<tr>
<th>Data Name</th>
<th>Published Year</th>
<th>Array Platform</th>
<th>Sample Size</th>
<th>GEO Accession ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lapointe</td>
<td>2004</td>
<td>cDNA</td>
<td>112</td>
<td>GSE3933</td>
</tr>
<tr>
<td>Yu</td>
<td>2004</td>
<td>HG-U95Av2</td>
<td>108</td>
<td>GSE6919</td>
</tr>
<tr>
<td>Varambally</td>
<td>2005</td>
<td>HG-U133 Plus 2</td>
<td>19</td>
<td>GSE3325</td>
</tr>
<tr>
<td>Tomlins</td>
<td>2007</td>
<td>cDNA</td>
<td>76</td>
<td>GSE6099</td>
</tr>
</tbody>
</table>

Usage

prostate

Format

A list containing 4 matrices. Each matrix is gene expression data after gene filtering.

Source

Gene Expression Omnibus (GEO)

References


### requireAll

**MetaPCA: Meta-analysis in the Dimension Reduction of Genomic data**

**Description**

requireAll description

**Usage**

```r
requireAll(packages)
```

**Arguments**

- `packages` A character vector of required packages. Unavailable packages are going to be installed.

**Value**

None

**Author(s)**

Don Kang (donkang75@gmail.com) and George Tseng (ctseng@pitt.edu)

**Examples**

```r
# Not run:
libs <- c("proto", "foreach", ifelse(.Platform$OS.type == "unix", "doMC", "doSMP"))
requireAll(libs)

# End(Not run)
```

### Spellman

**4 Spellman cancer studies**

**Description**

Yeast cell-cycle data set was divided into four subsets which correspond to the four different synchronization methods: alpha arrest (alpha), arrest of cdc15 or cdc28 temperature-sensitive mutant (cdc15 and cdc28), and elutriation (elu). We filtered out genes which have overall missing values $\geq 10\%$ or log2 transformed standard deviation $\geq .45$. 1025 genes were left, and the number of time points in the experiments were 18, 24, 17, and 14 for alpha, cdc15, cdc28, and elu, respectively. Additionally, we have imputed missing values using knn.
Usage

Spellman

Format

A list containing 4 matrices. Each matrix is gene expression data after gene filtering.

Source

Gene Expression Omnibus (GEO)

References

Index

*Topic **MetaAnalysis**
  - DropDupGenes, 3
  - MetaPCA, 5
  - MetaPCA-package, 2
  - PlotPC2D, 7
  - PreprocessMetaAnalysis, 9

*Topic **MetaPCA**
  - prostate, 10
  - Spellman, 11

*Topic **Microarray**
  - MetaPCA, 5
  - MetaPCA-package, 2
  - PlotPC2D, 7
  - PreprocessMetaAnalysis, 9

*Topic **PCA**
  - MetaPCA, 5
  - MetaPCA-package, 2
  - PlotPC2D, 7
  - PreprocessMetaAnalysis, 9

*Topic **require**
  - requireAll, 11

DropDupGenes, 3

MetaPCA, 5
MetaPCA-package, 2

PlotPC2D, 7
PreprocessMetaAnalysis, 9
prostate, 10

requireAll, 11

Spellman, 11