

# Package ‘gPCA’

July 2, 2014

**Type** Package

**Title** Batch Effect Detection via Guided Principal Components Analysis

**Version** 1.0

**Date** 2013-07-25

**Author** Sarah Reese

**Maintainer** Sarah Reese <reesese@vcu.edu>

**Description** This package implements guided principal components analysis for the detection of batch effects in high-throughput data.

**License** GPL (>= 2)

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2013-07-31 17:55:22

## R topics documented:

gPCA-package . . . . .	2
caseDat . . . . .	2
CumulativeVarPlot . . . . .	3
gDist . . . . .	4
gPCA.batchdetect . . . . .	5
PCplot . . . . .	7
<b>Index</b>	<b>8</b>

---

gPCA-package

*Batch Effect Detection via Guided Principal Components Analysis*

---

### Description

This package implements guided principal components analysis for the detection of batch effects in high-throughput data.

### Details

Package: gPCA  
Type: Package  
Version: 1.0  
Date: 2013-07-25  
License: GPL (>=2)

The function `gPCA.batchdetect()` is used to perform the batch detection test and outputs the resulting  $\delta$  statistic and corresponding  $p$ -value, along with other useful measures for visualization.

### Author(s)

Sarah Reese

Maintainer: Sarah Reese <reesese@vcu.edu>

### References

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

### See Also

[gPCA.batchdetect](#)

---

caseDat

*Case study copy number variation data*

---

### Description

This is a data set of copy number variation data with  $n = 500$  observations and  $p = 1000$  features. The length  $n$  batch vector (first column of `caseDat`) indicates the batch for each sample.

**Usage**

```
data(caseDat)
```

**Format**

A list with two objects:

batch A numeric vector indicating batch for the  $n = 500$  samples.

data A matrix of  $n = 500$  samples and  $p = 1000$  features.

**References**

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

**Examples**

```
data(caseDat)
```

---

CumulativeVarPlot	<i>Plot of the Cumulative Variance due to the Principal Components</i>
-------------------	--

---

**Description**

The function plots the cumulative variance of the principal components.

**Usage**

```
CumulativeVarPlot(out, ug = "unguided", ...)
```

**Arguments**

out	object resulting from <code>gPCA.batchdetect()</code> call.
ug	"guided" or "unguided". Do you want the cumulative variance from guided or unguided PCA plotted.
...	any other plot calls.

**Details**

This function plots the cumulative variance of the principal components from guided or unguided PCA calculated as (for the unguided case)

$$Var_l = \frac{\sum_{i=1}^l (PC_u)_i}{\sum_{i=1}^n (PC_u)_i}$$

for the  $l$ th principal component ( $l = 1, \dots, n$ ).

**Author(s)**

Sarah Reese <reesese@vcu.edu>

**References**

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

**See Also**

[gPCA.batchdetect](#), [gDist](#), [PCplot](#)

**Examples**

```
# CumulativeVarPlot(out,ug="unguided",col="blue")
```

---

gDist

*Density/Distribution Plot for gPCA*

---

**Description**

This function produces a density plot of the permutation  $\delta_p$  values.

**Usage**

```
gDist(out)
```

**Arguments**

out                    object resulting from `gPCA.batchdetect()` call.

**Author(s)**

Sarah Reese <reesese@vcu.edu>

**References**

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

**See Also**

[gPCA.batchdetect](#), [PCplot](#), [CumulativeVarPlot](#)

**Examples**

```
# gDist(out)
```

---

gPCA.batchdetect      *Guided Principal Components Analysis*


---

**Description**

Tests for batch effects an  $n \times p$  data set with batch vector given by batch using the  $\delta$  statistic resulting from guided principal components analysis (gPCA).

**Usage**

```
gPCA.batchdetect(x, batch, filt = NULL, nperm = 1000, center = FALSE, scaleY=FALSE,
seed = NULL)
```

**Arguments**

x	an $n \times p$ matrix of data where $n$ denotes observations and $p$ denotes the number of features (e.g. probe, gene, SNP, etc.).
batch	a length $n$ vector that indicates batch (group or class) for each observation.
filt	(optional) the number of features to retain after applying a variance filter. If NULL, no filter is applied. Filtering can significantly reduce the processing time in the case of very large data sets.
nperm	the number of permutations to perform for the permutation test, default is 1000.
center	(logical) Is your data x centered? If not, then center=FALSE and gPCA.batchdetect will center it for you.
scaleY	(logical) Do you want to scale the Y matrix by the number of samples in each batch? If not, then center=FALSE (default), otherwise, center=TRUE.
seed	the seed number for set.seed(). Default is NULL.

**Details**

Guided principal components analysis (gPCA) is an extension of principal components analysis (PCA) that guides the singular value decomposition (SVD) of PCA by applying SVD to  $\mathbf{Y}'\mathbf{X}$  where  $\mathbf{Y}$  is a  $n \times b$  batch indicator matrix of ones when an observation  $i$  ( $i = 1, \dots, n$ ) is in batch  $b$  and zeros otherwise.

The test statistic  $\delta$  along with a one-sided  $p$ -value results from a gPCA.batchdetect() call, along with the values of  $\delta_p$  from the permutation test. The  $\delta_p$  values can be used to visualize the permutation distribution of your test using the [gDist](#) function. For more information on gPCA, please see [reese](#).

**Value**

delta	test statistic $\delta$ from gPCA.
p.val	$p$ -value associated with $\delta$ resulting from gPCA.
delta.p	nperm length vector of delta values resulting from the permutation test.

<code>batch</code>	returns your length $n$ batch vector.
<code>filt</code>	returns the number of features the variance filter retained.
<code>n</code>	the number of observations
<code>p</code>	the number of features
<code>b</code>	the number of batches
<code>PCu</code>	principal component matrix from unguided PCA.
<code>PCg</code>	principal component matrix from gPCA.
<code>varPCu1</code>	the proportion out of the total variance associated with the first principal component of unguided PCA.
<code>varPCg1</code>	the proportion out of the total variance associated with the first principal component of gPCA.
<code>cumulative.var.u</code>	length $n$ vector of the cumulative variance of the $i = 1, \dots, n$ principal components from unguided PCA.
<code>cumulative.var.g</code>	length $b$ vector of the cumulative variance of the $k = 1, \dots, b$ principal components from gPCA.

**Author(s)**

Sarah Reese <reesese@vcu.edu>

**References**

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

**See Also**

[gDist](#), [PCplot](#), [CumulativeVarPlot](#),

**Examples**

```
data(caseDat)
batch<-caseDat$batch
data<-caseDat$data
out<-gPCA.batchdetect(x=data,batch=batch,center=FALSE,nperm=250)
out$delta ; out$p.val

## Plots:
gDist(out)
CumulativeVarPlot(out,ug="unguided",col="blue")
PCplot(out,ug="unguided",type="1v2")
PCplot(out,ug="unguided",type="comp",npcs=4)
```

**Description**

Produces principal component plots from either unguided or guided PCA.

**Usage**

```
PCplot(out, ug = "unguided", type = "1v2", npcs, ...)
```

**Arguments**

out	object resulting from <code>gPCA.batchdetect()</code> call.
ug	"guided" or "unguided". Do you want the cumulative variance from guided or unguided PCA plotted.
type	type of plot. Either "1v2" to plot the first two principal components, or "comp" to compare all principal component up to the level of npcs.
npcs	Number of principal components to plot when "comp" type is chosen.
...	any other plot calls.

**Details**

This function plots either the first principal component versus the second principal component (`type="1v2"`) from guided or unguided PCA, or compares (`type="comp"`) all combinations of the principal components up to the value of npcs.

**Author(s)**

Sarah Reese <reesese@vcu.edu>

**References**

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J. A., and Eckel-Passow, J. E. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, (in review).

**See Also**

[gPCA.batchdetect](#), [gDist](#), [CumulativeVarPlot](#)

**Examples**

```
# PCplot(out,ug="unguided",type="1v2")
# PCplot(out,ug="unguided",type="comp",npcs=4)
```

# Index

\*Topic **\textasciitildekwd1**

CumulativeVarPlot, 3  
gDist, 4  
gPCA.batchdetect, 5  
PCplot, 7

\*Topic **\textasciitildekwd2**

CumulativeVarPlot, 3  
gDist, 4  
gPCA.batchdetect, 5  
PCplot, 7

\*Topic **datasets**

caseDat, 2

\*Topic **package**

gPCA-package, 2

caseDat, 2

CumulativeVarPlot, 3, 4, 6, 7

gDist, 4, 4, 5–7

gPCA (gPCA-package), 2

gPCA-package, 2

gPCA.batchdetect, 2, 4, 5, 7

PCplot, 4, 6, 7