BatchI Tutorial

Anna Papież

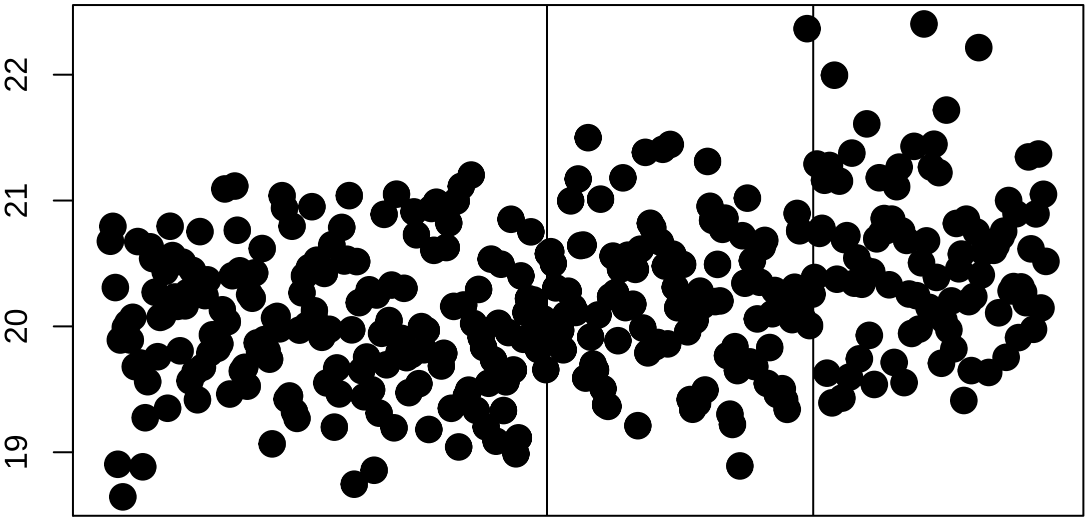
2017-09-30

In this tutorial examples are shown for basic and more customized use of the BatchI package.

Batch identification

The basic functionality in this package relies on use of the batchI function which finds the optimal partition of samples based on a quality index summarizing each sample with the use of a dynamic programing algoritm. When such an index is calculated (average intesity for microarrays, median counts for RNAseq data, total ion current for mass spectrometry samples) it is the simplest case and can be illustrated by an MS example:

```
library(batchI)
data(protTIC)
batches <- batchI(protTIC,K_gr = 3, min_seg = 3,"AD")
```



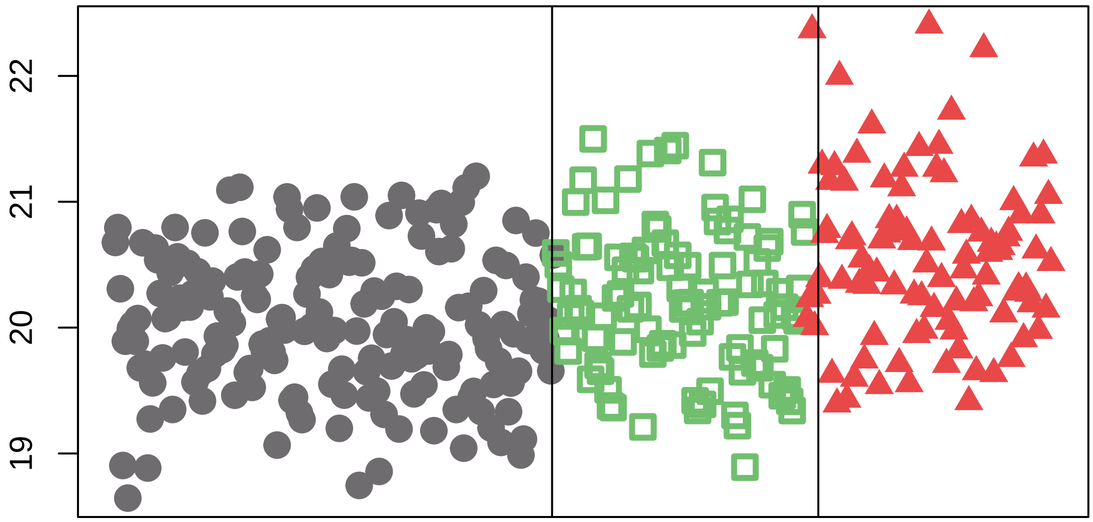
```
batches
```

```
##   opt_part        Q
## 1      177 176.4438
## 2      284 174.4017
```

The protTIC variable is a vector of total ion current calculated for each sample, the rest of the arguments are: the predefined number of batches (3), the minimum number of samples that should form a single batch (3) and the optimization statistic (Absolute Deviation). The function will return the index of the first element in batches 2:K__gr and their respective optimization score. It will also plot the data with the optimal division marked in vertical lines.

For the sake of comparison, if the original batch division is known a priori, the plotBatch function allows for coloring the samples according to their labeling:

```
plotBatch(protTIC,batches$opt_part,batchTIC)
```



Establishing the number of batches

If the number of batches is not known or suspected, the package allows for the discovery of the optimal number of batches. This may be accomplished by determining the significance of the gPCA delta statistic for data given a particular set of batch labels. This may be performed using kernel density estimators in the gPCApval function:

```
data(rnaseq)
gPCApval(rnaseqDat,rnaseqBatch)
```

```
## $p_val
## [1] 0.358122
##
## $delta
##           PC1
## PC1 0.9125176
##
## $varexp
## [1] 20.56316
```

The function returns the delta statistic along with its p-value and the percent of total variation explained by batch effects. It the example above it was used on an RNAseq count dataset with the original batch labeling. In order to choose an optimal number of batches, this function would have to be used on a set of labels obtained using the batchI function for different numbers of batches and the division corresponding to the lowest p-value would be considered as the optimal. The batchNum function in this package allows for an automated run of this iterative procedure:

```
batchNum(rnaseqDat,"median",4)
```

The function requires the dataset and optionally a summary quality index for each sample to serve as a basis for the dynamic programming algorithm. If such a vector is not available, the statistic of choice is calculated as the mean, median or sum. The last argument is the maximum number of batches to be tested in the iterative procedure. The function returns - the optimal batch partitionning parameters: indices of the first element in every batch except the first and values of the quality measure for every batch, - the vector of batch labels for each sample in the optimal batch division, - the gPCA statistic set for the optimal partitioning: the gPCA p-value, delta statistic and percent of variation explained by batch.