Clusterlab is a CRAN package for the routine testing of clustering algorithms. It can simulate positive (data-sets with >1 clusters) and negative controls (data-sets with 1 cluster). Why test clustering algorithms? Because they often fail in identifying the true K in practice, published algorithms are not always well tested, and we need to know about ones that have strange behaviour. I’ve found in many own experiments on clustering algorithms that algorithms many people are using are not necessary ones that provide the most sensible results. I can give a good example below.

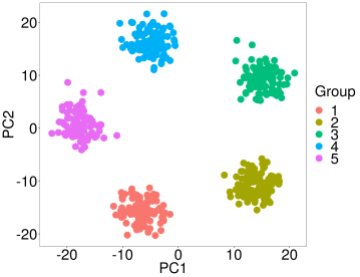
I was interested to see clusterExperiment, a relatively new clustering package on the Bioconductor, cannot detect the ground truth K in my testing so far. Instead yielding solutions with many more clusters than there are in reality. On the other hand, the package I developed with David Watson here at QMUL, does work rather well. It is a derivative of the Monti et al. (2003) consensus clustering algorithm, fitted with a Monte Carlo reference procedure to eliminate overfitting. We called this algorithm M3C.

library(clusterExperiment)

library(clusterlab)

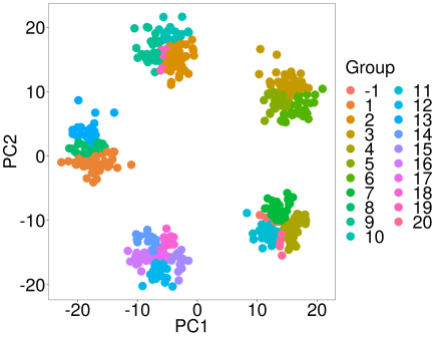
library(M3C)

**Experiment 1: Simulate a positive control dataset (K=5)**



k5 <- clusterlab(centers=5)

**Experiment 2: Test RSEC**

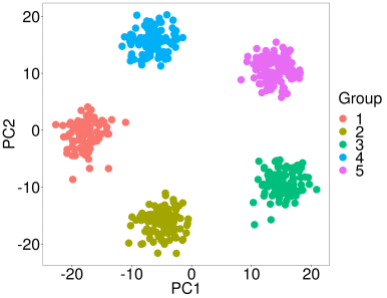


rsec\_test <- RSEC(as.matrix(k5$synthetic\_data),ncores=10)

assignments <- primaryCluster(rsec\_test)

pca(k5$synthetic\_data,labels=assignments)

**Experiment 3: Test M3C**



M3C\_test <- M3C(k5$synthetic\_data,iters=10,cores=10)

optk <- which.max(M3C\_test$scores$RCSI)+1

pca(M3C\_test$realdataresults[[optk]]$ordered\_data,

labels=M3C\_test$realdataresults[[optk]]$ordered\_annotation$consensuscluster,printres=TRUE)

Interesting isn't it R readers.

Well all I can say is I recommend comparing different machine learning methods and if in doubt, run your own control data-sets (positive and negative controls) to test various methods.

**If you are clustering omic’ data from a single platform (RNAseq, protein, methylation, etc) I have tested and recommend the following packages:**

And to be honest, that is about it. I have also tested PINSplus, GAP-statistic, and SNF, but they did not provide satisfactory results in my experiments on single platform clustering (currently unpublished data). Multi-omic and single cell RNA-seq analysis is another story, there will be more on that to come in the future R readers.

Remember there is a darkness out there R readers, not just in Washington, but in the world of statistics. It is there because of the positive results bias in science, because of people not checking methods and comparing them with one another, and because of several other reasons I can’t even be bothered to mention.