**1) Theory:** Explain dimensionality reduction in the context of PCA (possibly illustrate your explanation with a drawing, 10-20 lines)

**2) True or false**. Read the questions carefully. Motivate your answer (try to be concise and complete). The statement you have to motivate is given in italics.

a) *Hierarchical clustering with a correlation-derived distance metric applied on non-rescaled data x (data not mean centered nor variance rescaled) gives exactly the same results as hierarchical clustering with a Euclidian distance applied on the same non-rescaled data.*

or

d.correlation <- as.dist(1 - cor(x)))

d.euclidian <- dist(x, method="euclidian")

hclust(d.correlation, method = ‘complete’)

hclust(d.euclidian, method = ‘complete’)

b) Given a cancer dataset, containing the expression of 3051 genes in 38 patients. The dataset was rescaled and the Euclidean distance was calculated as follows

golub\_m = golub – rowMeans(golub)

SD = apply(golub, 1, sd, na.rm=TRUE)

golub\_r = golub\_m/SD

d.euclidian <- dist(golub\_r, method="euclidian")

m <- as.matrix(dist(golub))

dim(m)

[1] 3051 3051

kmeangolub10= kmeans(d.euclidian,10)

*Kmeans clustering is executed on the matrix golub to group together patients based on the similarity of their (rescaled) expression profiles . 10 patient groups are generated in total*

1. PCA was performed on a dataset that contains the measurements of 154 metabolites profiled at several time points after treatment (more specifically at time 0 h, 4, 12, 24, 48 and 96 h after treatment). For each time point several replicates were measured) : The total datamatrix mDC (dim 154 X 52) contains 52 columns (corresponding to the timepoints and their replicates) and 154 metabolites (rows).

See an example below: X12h.1 refers to e.g. timepoint 12 first replicate

mDC[1,]

X0h X0h.1 X0h.2 X0h.3 X0h.4 X0h.5 X0h.6 X1h X1h.1 X1h.2 X1h.3 X1h.4 X1h.5 X1h.6 X1h.7 X4h X4h.1 X4h.2 X4h.3 X4h.4

0.0015 0.0252 0.0159 0.0116 -0.0431 -0.0053 -0.0058 0.0127 0.0309 0.0092 0.0430 -0.0457 -0.0108 -0.0126 -0.0831 0.0192 0.0156 -0.0015 -0.0500 -0.0481

X4h.5 X12h X12h.1 X12h.2 X12h.3 X12h.4 X12h.5 X12h.6 X12h.7 X24h X24h.1 X24h.2 X24h.3 X24h.4 X24h.5 X24h.6 X48h X48h.1 X48h.2 X48h.3

0.0127 0.1563 0.1174 0.1112 0.1411 0.0287 0.0597 0.0641 0.0839 0.1388 0.1160 0.1008 0.0276 0.0917 0.0629 0.0574 0.2348 0.2672 0.2599 0.2314

X48h.4 X48h.5 X48h.6 X48h.7 X96h X96h.1 X96h.2 X96h.3 X96h.4 X96h.5 X96h.6 X96h.7

0.1901 0.2015 0.2209 0.1881 0.3410 0.3206 0.2923 0.3276 0.2742 0.2560 0.2424 0.2655

>

PCA was applied tot his dataset as follows

PCAres<-prcomp(t(mDC), scale = TRUE, center=TRUE)

And the results were plotted as follows

plot(predict(PCAres)[,1],predict(PCAres)[,2])

abline(v=0, col="gray")

abline(h=0, col="gray")

text(predict(PCAres)[,1],predict(PCAres)[,2], labels=sub("X(.+h)(\\..)?","\\1",rownames(t(mDC))),cex=1, adj = c(0,0))



Explain what is plotted in the figure above, give a complete answer by first explaining what PCA does and then illustrating your explanation with the example shown in the figure, your explanation should define what the variables are and the observations. Is it logical what you observe in the figure? ).

2) What would this code be doing? Golub is a gene-patient dataset. The dataset contains x patients and 3051 genes. Explain the code line per line. Explain why you need to perform the different steps. Explain what you expect in the output.

golub\_m = golub – rowMeans(golub) [code on the whole dataset]

SD = apply(golub, 1, sd, na.rm=TRUE)

golub\_r = golub\_m/SD

d.euclidian <- dist(golub\_r, method="euclidian")

m <- as.matrix(dist(golub))

dim(m)

#(3051,3051)

kmeangolub100= kmeans(d.euclidian\_r,100)

b) what will happen if you change the following parameter setting

kmeangolub10= kmeans(d.euclidian\_r,10)