Homework due May 13, 2021 23:01 +03

We will continue to use this dataset:

```
library(Biobase)
library(GSE5859Subset)
data(GSE5859Subset)
```

and define

```
y = geneExpression - rowMeans(geneExpression)
```

Compute and plot an image of the correlation for each sample. Make two image plots of these correlations. In the first one, plot the correlation as image. In the second, order the samples by date and then plot the an image of the correlation. The only difference in these plots is the order in which the samples are plotted.

Factor Analysis Exercises #1

1/1 point (graded)

Based on these plots, which of the following you would say is true:

- ○ The samples appear to be completely independent of each other.
- ○ Sex seems to be creating structures as evidenced by the two cluster of highly correlated samples.
- ◉ The fact that in the plot ordered by month we see two groups mainly driven by month and within these, we see subgroups driven by date seems to suggest date more than month per se are the hidden factors.
- ○ There appear to be only two factors completely driven by month.

✔

Explanation

```
##simple version
library(rafalib)
sex = sampleInfo$group
mypar(1,2)
cors = cor(y)
image(cors)
o = order(sampleInfo$date)
image(cors[o,o])

##advanced version
library(rafalib)
sex = sampleInfo$group
mypar(1,2)
cols=colorRampPalette(rev(brewer.pal(11,"RdBu")))(100)
cors = cor(y)
image(1:ncol(y),1:ncol(y),cors,col=cols,zlim=c(-1,1),
       xaxt="n",xlab="",yaxt="n",ylab="")
axis(2,1:ncol(y),sex,las=2)
axis(1,1:ncol(y),sex,las=2)
o = order(sampleInfo$date)
image(1:ncol(y),1:ncol(y),cors[o,o],col=cols,zlim=c(-1,1),
      xaxt="n",xlab="",yaxt="n",ylab="")
label = gsub("2005-","",sampleInfo$date[o])
axis(2,1:ncol(y),label,las=2)
axis(1,1:ncol(y),label,las=2)
```

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Factor Analysis Exercises #2

1/1 point (graded)

Based on the correlation plots above, we could argue that there are at least two hidden factors. Using PCA estimate these two factors. Specifically, apply the `svd()` to `y` and use the first two PCs as estimates.

Which command gives us these estimates?

- `pcs = svd(y)$v[1:2,]`
- **(selected)** `pcs = svd(y)$v[,1:2]`
- `pcs = svd(y)$u[,1:2]`
- `pcs = svd(y)$d[1:2]`

✔

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Factor Analysis Exercises #3

1/1 point (graded)

Plot each of the estimated factor ordered by date. Use color to denote month. The first factor is clearly related to date.

Which of the following appear to be most different according to this factor?

- **(selected)** June 23 and June 27
- Oct 07 and Oct 28
- June 10 and June 23
- June 15 and June 24

✔

Explanation

Based on the plot below, we see that the first factor changes:

```
pcs = svd(y)$v[,1:2]
o = order(sampleInfo$date)
cols = as.numeric(month)[o]
mypar(2,1)
for(i in 1:2){
  plot(pcs[o,i],col=cols,xaxt="n",xlab="")
  label = gsub("2005-","",sampleInfo$date[o])
  axis(1,1:ncol(y),label,las=2)
}
```

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Factor Analysis Exercises #4

1/1 point (graded)

Use the `svd()` function to obtain the principal components (PCs) for our detrended gene expression data `y`.

How many principal components (PCs) explain more than 10% each of the variability?

2

✔ **Answer:** 2

2

Explanation

```
s = svd(y)
varexplained = s$d^2/ sum(s$d^2)
plot(varexplained)
sum(varexplained>0.10)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Factor Analysis Exercises #5

2/2 points (graded)

Which PC most correlates (negative or positive correlation) with month?

1 ✔ **Answer:** 1

1

Explanation

```
month = factor( format(sampleInfo$date,"%m"))
cors = cor( as.numeric(month),s$v)
plot(t(cors))
which.max(abs(cors))
```

What is this correlation (in absolute value)?

0.8297915 ✔ **Answer:** 0.8297915

0.8297915

Explanation

```
max(abs(cors))
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Factor Analysis Exercises #6

2/2 points (graded)

Which PC most correlates (negative or positive correlation) with sex?

1 ✔ **Answer:** 1

1

Explanation

```
sex = sampleInfo$group
cors = cor( sex,s$v)
plot(t(cors))
which.max(abs(cors))
```

What is this correlation (in absolute value)?

0.6236858 ✔ **Answer:** 0.6236858

0.6236858

Explanation

```
max(abs(cors))
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

Factor Analysis Exercises #7

2/2 points (graded)

Now instead of using month, which we have shown does not quite describe the batch, add the two estimated factors in Factor Analysis Exercises #6 to the linear model we used in previous exercises:

```
X <- model.matrix(~sex+s$v[,1:2])
```

Apply this model to each gene, and compute q-values for the sex difference.

How many q-values are <0.1 for the sex comparison?

14 ✔ **Answer:** 14

14

Explanation

```
X= model.matrix(~sex+s$v[,1:2])
pvals = sapply(1:nrow(geneExpression),function(i){
  y = geneExpression[i,]
  fit = lm(y~X-1)
  summary(fit)$coef[2,4]
})
qvals = qvalue(pvals)$qvalue
sum(qvals<0.1)
```

What proportion of the genes are on chrX and chrY?

1.0 ✔ **Answer:** 1.0

1.0

Explanation

```
index = geneAnnotation$CHR[qvals<0.1]%in%c("chrX","chrY")
mean(index)
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem