ST340 Lab 2: SVD & PCA

2019–20

1: A simple singular value decomposition

(a) Generate a realization of a 4×5 Gaussian random matrix G.
(b) Look at `?svd`.
(c) Set U, d, and V by using `svd`.
(d) Check that `G` is equal to `U%*%Sigma%*%t(V)` (to machine precision).
(e) Plot the singular values.
(f) Compute G_2, the 2-rank approximation of G, and also compute $||G - G_2||_F$.
(g) Does the value agree with the theory?

2: Image compression via the singular value decomposition

```
load("pictures.rdata")
source("svd.image.compression.R")
```

Take a look at `svd.image.compression.R` and understand what the code is doing. Then run `image.compression()` here to see how well we can compress our images.

3: PCA: Crabs

(a) Load the MASS library to access the crabs data.

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.5.2
```

(b) Read `?crabs`.

(c) Read in the FL, RW, CL, CW, and BD measurements.

```
Crabs <- crabs[,4:8]
Crabs.class <- factor(paste(crabs[,1],crabs[,2],sep=""))
plot(Crabs,col=Crabs.class,pch=20)
```

(d) Read `?prcomp` and use it to obtain the principal components of a centred and scaled version of `Crabs`. Call the output of `prcomp` '`Crabs.pca`'.

(e) If you `plot(Crabs.pca)` it visualizes the variances associated with the components.

```
plot(Crabs.pca)
```

(f) Plot PC2 against PC1.

(g) Read `?pairs` and use it to find a pair of components with good separation of the classes.

(h) Read `?scale`. Check that you can obtain the principal components by using the singular value decomposition on a centred and scaled version of `Crabs`.

4: PCA: Viruses

This is a dataset on 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber and others) described by Fauquet *et al.* (1988) and analysed by Eslava-Gómez (1989). There are 18 measurements on each virus, the number of amino acid residues per molecule of coat protein.

```
load("viruses.rdata")
```

(a) Obtain the principal components of a centred and scaled version of allviruses.

```
groups <- rep(0,61)
groups[1:3] <- 1
groups[4:9] <- 2
groups[10:48] <- 3
groups[49:61] <- 4
group.names <- c("Hordeviruses","Tobraviruses","Tobamoviruses","furoviruses")
```

If you colour by groups (i.e. `col=groups` in plot) then black is horde, red is tobra, green is tobamo, blue is furo.

(b) Do the principal components show some separation between the viruses?

(c) The largest group of viruses is the tobamoviruses. Does a principal component analysis suggest there might be subgroups within this group of viruses?