DA410_Project6_MattGraham

Principal Component Analysis (PCA)

```
library(nnspat)  # used for dist2full()
library("dplyr")  # used to select numeric datatypes
library("ggplot2")
library(reshape)  # used for melting matricies
library(klaR)
library(ggvis)
library(class)
library(gmodels)
library(MASS)
library(readxl)
library(psych)
```

Get data

```
beetles <- read.table("C:/mattgraham93.github.io/school/22_3_DA410/data/T5_5_FBEETLES.DAT", h
eader=FALSE)
colnames(beetles) <- c("Experiment", 'Species', 'post_dist', 'l_elytra','len_sec_an', 'len_th
i_an')

oleracea <- beetles[beetles$Species==1, -c(1:2)]  # we assume that species 1 = oleracea
carduorum <- beetles[beetles$Species==2, -c(1:2)] # we assume that species 2 = carduorum

beetles
```

Experiment <int>	Species <int>	post_dist <int>	l_elytra <int>	len_sec_an <int>	len_thi_an <int>
1	1	189	245	137	163
2	1	192	260	132	217
3	1	217	276	141	192
4	1	221	299	142	213
5	1	171	239	128	158
6	1	192	262	147	173
7	1	213	278	136	201
8	1	192	255	128	185
9	1	170	244	128	192
10	1	201	276	146	186

1-10 of 39 rows Previous **1** 2 3 4 Next

Oleracea PCA

a. Get S and R

```
S1 = cov(oleracea)
R1 = cor(oleracea)
```

S1

```
as.data.frame(S1)
```

	post_dist <dbl>	l_elytra <dbl>	len_sec_an <dbl>	len_thi_an <dbl>
post_dist	187.59649	176.86257	48.37135	113.58187
l_elytra	176.86257	345.38596	75.97953	118.78070
len_sec_an	48.37135	75.97953	66.35673	16.24269
len_thi_an	113.58187	118.78070	16.24269	239.94152

4 rows

R1

```
as.data.frame(R1)
```

	post_dist <dbl>	l_elytra <dbl>	len_sec_an <dbl>	len_thi_an <dbl>
post_dist	1.0000000	0.6948183	0.4335442	0.5353576
l_elytra	0.6948183	1.0000000	0.5018822	0.4126110
len_sec_an	0.4335442	0.5018822	1.0000000	0.1287250
len_thi_an	0.5353576	0.4126110	0.1287250	1.0000000

4 rows

b. Get eigenvalues and eigenvectors of S and R

```
eig.S1 <- eigen(S1)
eig.R1 <- eigen(R1)
```

Eigen S1

```
eig.S1
```

```
## eigen() decomposition
## $values
## [1] 561.30574 168.98584  65.27709  43.71203
##
## $vectors
##            [,1]         [,2]       [,3]       [,4]
## [1,] -0.4997445  0.009204574  0.8230272  0.2698089
## [2,] -0.7187015 -0.484408702 -0.4778690  0.1430301
## [3,] -0.1739702 -0.220296505  0.2042647 -0.9378058
## [4,] -0.4510631  0.846600812 -0.2292234 -0.1651236
```

Eigen R1

```
eig.R1
```

```
## eigen() decomposition
## $values
## [1] 2.3952353 0.8869271 0.4318977 0.2859399
##
## $vectors
##           [,1]       [,2]       [,3]        [,4]
## [1,] 0.5729764  0.1117770 -0.3104311  0.75022433
## [2,] 0.5617458 -0.1172002 -0.5248012 -0.62872048
## [3,] 0.4192879 -0.7000880  0.5775318  0.02305369
## [4,] 0.4246614  0.6954472  0.5428382 -0.20332864
```

c. Percent variance explained and plot S and R

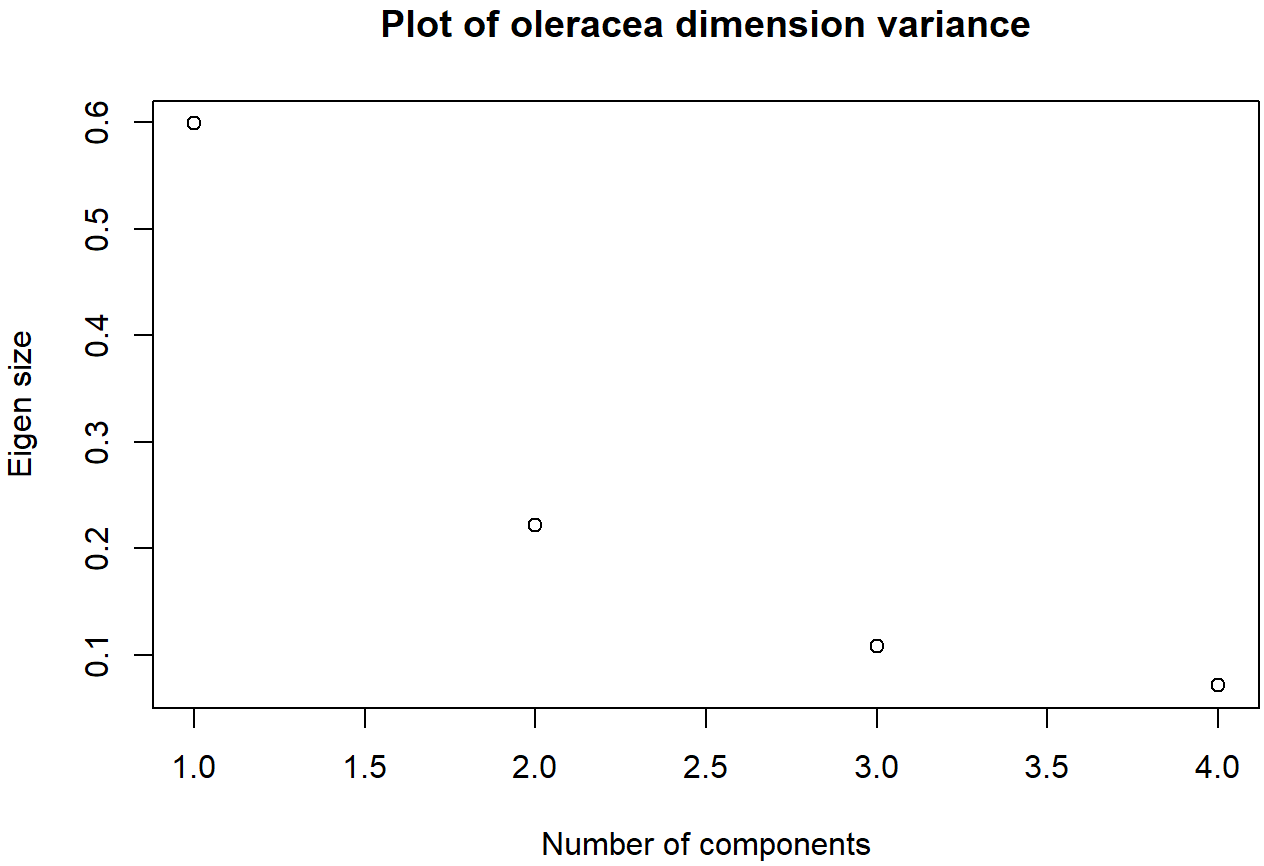
```
for (r in eig.R1$values) {
  print(r/sum(eig.R1$values))
}
```

```
## [1] 0.5988088
## [1] 0.2217318
## [1] 0.1079744
## [1] 0.07148499
```

We can see that of our eigen values, ~60% of our variance is explained with just one dimension, while ~32% is explained with 2 and 3 dimensions. This can be seen below.

Plot

```
plot(eig.R1$values/sum(eig.R1$values), xlab = 'Number of components', ylab='Eigen size', main
='Plot of oleracea dimension variance')
```



d. Decide compontent retention and show reasoning

Based on the percentage of total variance, it makes sense to keep the first 3 components as component 4 makes up less over 10% of all explained variance.

e. Interpretation

Over the duration of subjects' time, as we model our data, we can conclude that most of our variation happens within the post_dist, l_elytra, and len_sec_an.

Carduorum PCA

a. Get S and R

```
S2 = cov(carduorum)
R2 = cor(carduorum)
```

S2

```
as.data.frame(S2)
```

	post_dist <dbl>	l_elytra <dbl>	len_sec_an <dbl>	len_thi_an <dbl>
post_dist	101.83947	128.06316	36.98947	32.59211
l_elytra	128.06316	389.01053	165.35789	94.36842
len_sec_an	36.98947	165.35789	167.53684	66.52632
len_thi_an	32.59211	94.36842	66.52632	177.88158

4 rows

R2

```
as.data.frame(R2)
```

	post_dist <dbl>	l_elytra <dbl>	len_sec_an <dbl>	len_thi_an <dbl>
post_dist	1.0000000	0.6434065	0.2831815	0.2421524
l_elytra	0.6434065	1.0000000	0.6477227	0.3587406
len_sec_an	0.2831815	0.6477227	1.0000000	0.3853655
len_thi_an	0.2421524	0.3587406	0.3853655	1.0000000

4 rows

b. Get eigenvalues and eigenvectors of S and R

```
eig.S2 <- eigen(S2)
eig.R2 <- eigen(R2)
```

Eigen S2

```
eig.S2
```

```
## eigen() decomposition
## $values
## [1] 555.69314 145.44632  93.46372  41.66524
##
## $vectors
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.2836552 -0.2007357  0.5315166 -0.77248627
## [2,] -0.8068689 -0.3389760  0.1218433  0.46820095
## [3,] -0.4222422  0.1359900 -0.7897513 -0.42368751
## [4,] -0.3003563  0.9090144  0.2809577  0.06739234
```

Eigen R2

```
eig.R2
```

```
## eigen() decomposition
## $values
## [1] 2.3143372 0.8280076 0.6404353 0.2172199
##
## $vectors
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.4746835  0.6147015  0.4332980 -0.45724233
## [2,] -0.5938453  0.2236769 -0.1730128  0.75324823
## [3,] -0.5121195 -0.2810781 -0.6603022 -0.47194250
## [4,] -0.3996960 -0.7022150  0.5884899  0.02858017
```

c. Percent variance explained and plot S and R

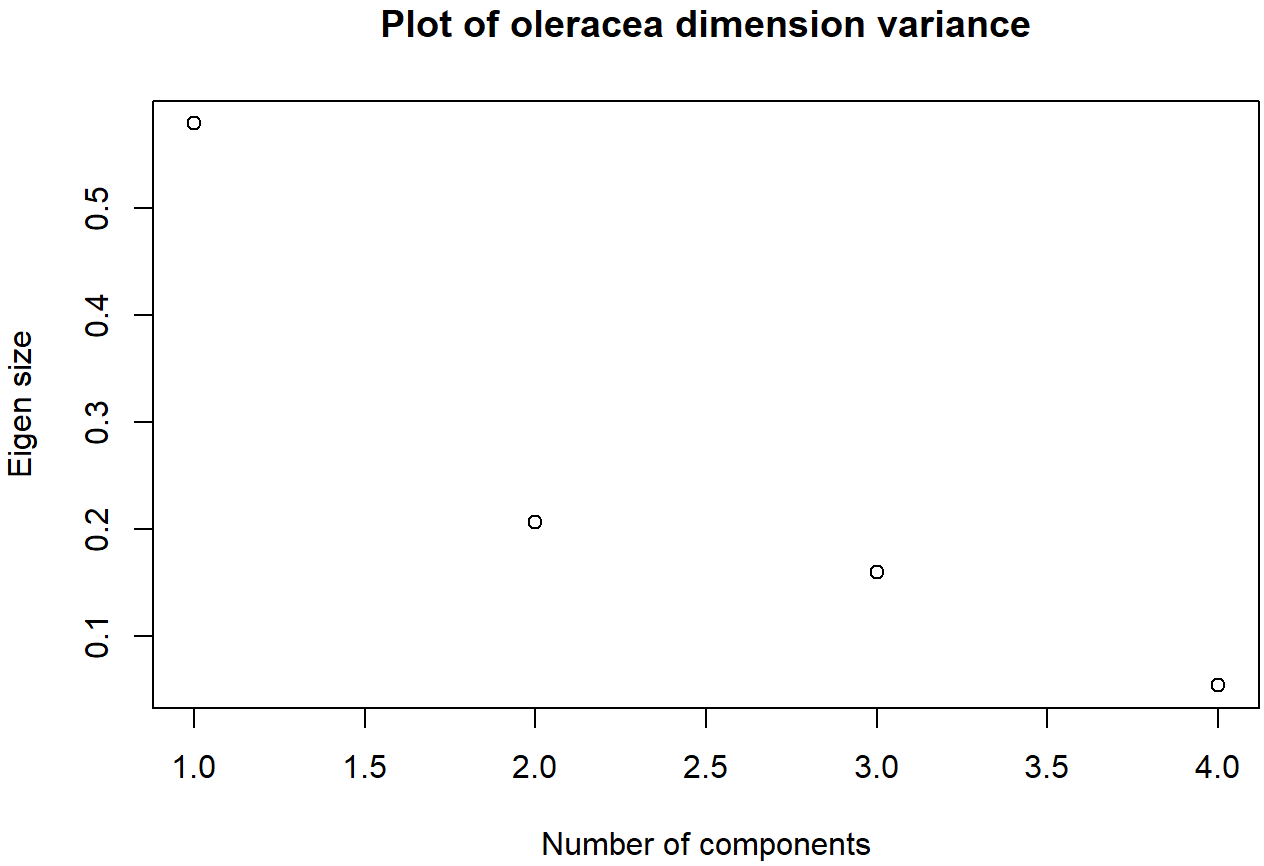
```
for (r in eig.R2$values) {
  print(r/sum(eig.R2$values))
}
```

```
## [1] 0.5785843
## [1] 0.2070019
## [1] 0.1601088
## [1] 0.05430499
```

We can see that of our eigen values, ~57.9% of our variance is explained with just one dimension, while ~36% is explained with 2 and 3 dimensions. This can be seen below.

Plot

```
plot(eig.R2$values/sum(eig.R2$values), xlab = 'Number of components', ylab='Eigen size', main
='Plot of oleracea dimension variance')
```



d. Decide compontent retention and show reasoning

Based on the percentage of total variance, it makes sense to keep the first 3 components as component 4 makes up less over 10% of all explained variance.

e. Interpretation

Over the duration of subjects' time, as we model our data, we can conclude that most of our variation happens within the post_dist, l_elytra, and len_sec_an. THough it is interesting to see we have less variation in post_dist.