MAD Assignment 3

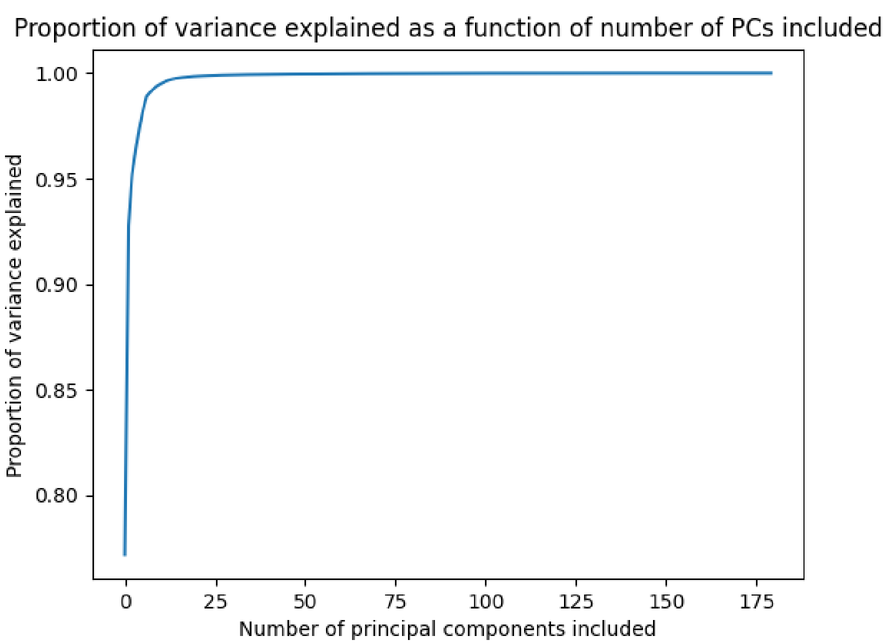
Xingrong Zong

December 7, 2021

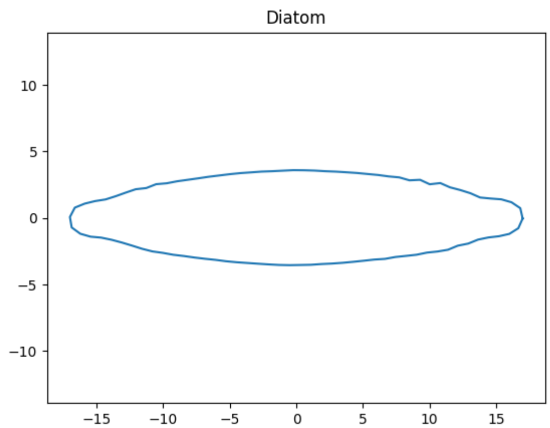
Exercise 1 (Implement PCA)

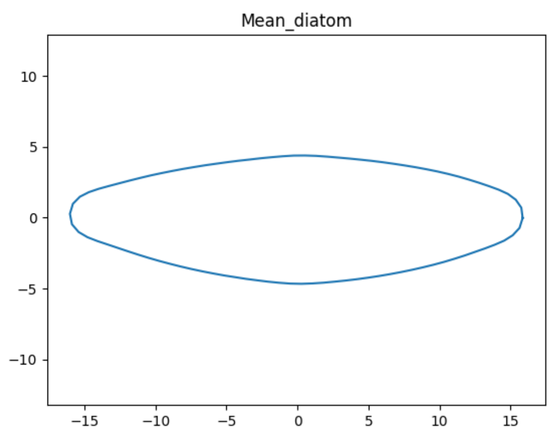
a)

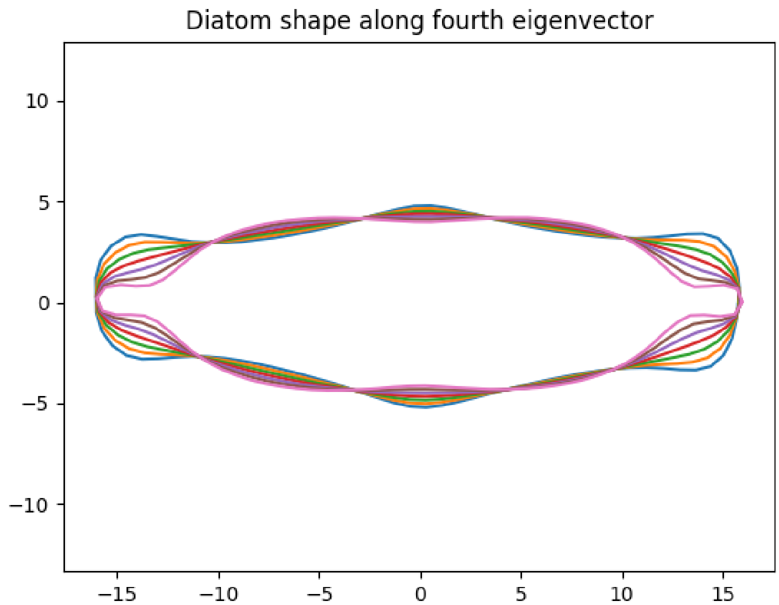
```
Proportion of variance explained by the first 1 principal components: 0.7718721493017529
Proportion of variance explained by the first 2 principal components: 0.9276996293043025
Proportion of variance explained by the first 3 principal components: 0.9521198453942007
Proportion of variance explained by the first 4 principal components: 0.9637878603999529
Proportion of variance explained by the first 5 principal components: 0.9739084497954094
Proportion of variance explained by the first 6 principal components: 0.98236065164916
Proportion of variance explained by the first 7 principal components: 0.9889975933245944
Proportion of variance explained by the first 8 principal components: 0.9910287023941854
Proportion of variance explained by the first 9 principal components: 0.9926692113360289
Proportion of variance explained by the first 10 principal components: 0.9939926229665051
```



b)


Diatom
10
5
0
−5
−10
−15
−10
−5
0
5
10
15


Mean_diatom
10
5
0
−5
−10
−15
−10
−5
0
5
10
15



Exercise 2 (Inequalities)

Let $X - \mu = Y$

then, $E[(X-\mu)^4] = E[Y^4]$

$\sqrt{E[Y^4]}^2 \ \geq \ \sqrt{E[Y^4] - Var(Y^2)}^2 \ \geq \ (\sqrt{E[Y^4] - Var(Y^2)} - (E[Y])^2)^2,$ when

$Var(Y^2) \geq 0, (E[Y])^2 \geq 0$

Since $Var(Y^2) = E[Y^4] - (E[Y^2])^2$,

so, $E[Y^2] = \sqrt{E[Y^4] - Var(Y^2)}$,

replace to the above step, $(\sqrt{E[Y^4] - Var(Y^2)} - (E[Y])^2)^2 = (E[Y^2] - (E[Y])^2)^2 =$

$(Var(Y))^2 = \sigma^4$

therefore, $E[(X-\mu)^4] \geq \sigma^4$

Exercise 3 (Confidence Intervals)

$\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma} \sim \mathrm{N}(0,1)$

$\sigma = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \hat{\mu})}^2$

```
99.0%-confidence interval:
b) Not matching in 98 (out of 10000) experiments, 0.98%
c) Not matching in 98 (out of 10000) experiments, 0.98%
```

Exercise 4 (Hypothesis Testing)

a)

Null hypothesis $H_0 := 0$, as the single gene has no influence on the flowering time of a plant, so the differences will be 0

Alternative hypothesis $H_A :\neq 0$, as the scientist claims the single gene has an influence on the flowering time of a plant, so the differences will not be 0

b)

$T = \frac{\bar{X}-H_0}{\sigma/\sqrt{n}}$

$\bar{X} = \frac{(4.1-3.1)+(4.8-4.3)+(4-4.5)+(4.5-3)+(4-3.5)}{5} = \frac{1+0.5-0.5+1.5+0.5}{5} = \frac{3}{5} = 0.6$

$\sigma = \sqrt{\frac{(1-0.6)^2+(0.5-0.6)^2+(-0.5-0.6)^2+(1.5-0.6)^2+(0.5-0.6)^2}{5-1}} = \sqrt{\frac{0.16+0.01+1.21+0.81+0.01}{4}} =$
$\sqrt{\frac{2.2}{4}} = \sqrt{0.55} \approx 0.74$

$n = 5$

$t = \frac{0.6-0}{0.74/\sqrt{5}} \approx 1.81$

Since doing a two-sided t-test at significance level 0.05, $c_1 = -2.37, c_2 = 2.37$

$c_1 < t < c_2$

c)

$T = \frac{\bar{X}-H_0}{\sigma/\sqrt{n}}$

$\bar{X} = \frac{k((4.1-3.1)+(4.8-4.3)+(4-4.5)+(4.5-3)+(4-3.5))}{5k} = \frac{1+0.5-0.5+1.5+0.5}{5} = \frac{3}{5} = 0.6$

$\sigma = \sqrt{\frac{k((1-0.6)^2+(0.5-0.6)^2+(-0.5-0.6)^2+(1.5-0.6)^2+(0.5-0.6)^2)}{5k-1}} = \sqrt{\frac{k(0.16+0.01+1.21+0.81+0.01)}{5k-1}} =$
$\sqrt{\frac{2.2k}{5k-1}}$

$n = 5k$

$t = \frac{0.6}{\sqrt{\frac{2.2k}{5k-1}}/\sqrt{5k}} = \frac{0.6\sqrt{5k}}{\sqrt{\frac{2.2k}{5k-1}}} = \frac{3\sqrt{\frac{11k^2}{5k-1}}(5k-1)}{11k} = \frac{3\sqrt{11(5k-1)}}{11}$

As $k \to \infty$, therefore, $t \to \infty$, eventually it will be $\geq c_2$, and reject the hypothesis. So the scientist cannot change the test result by copying the data set k times.