# Assignment 3

Marion Rosec

Decembre 2023

## Exercise 1

### a)

Here are the variances explained by the first 1 to 10 principal components (cumulative) :

```python
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
```

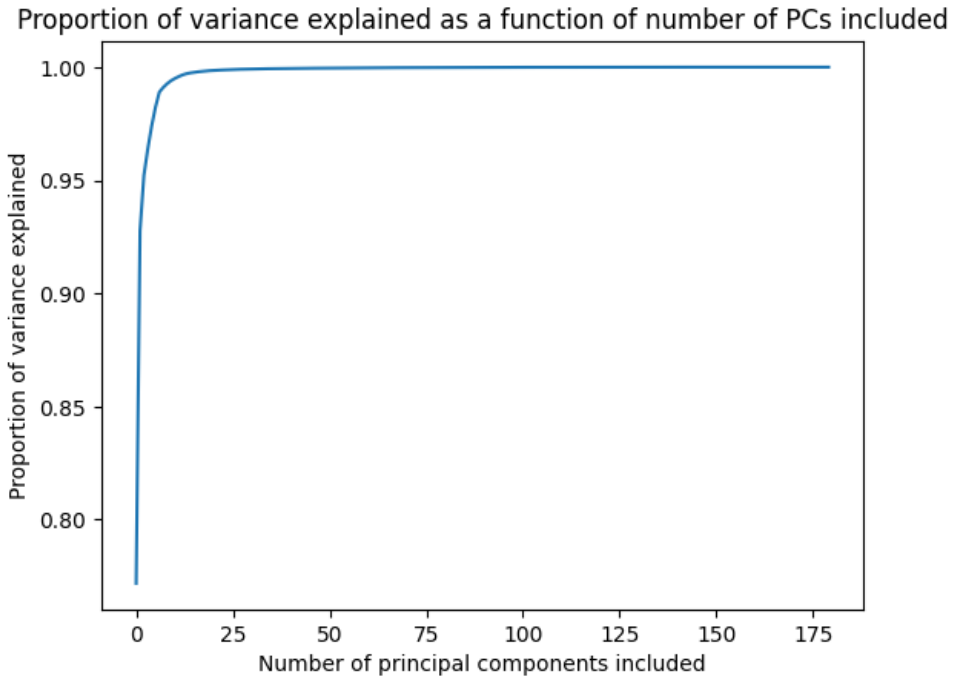


Figure 1 – Cumulative variance explained by number of principal components

**b)**

In the following figure (figure 2), we can see the effect of the fourth eigenvector on the shape of the diatom. We can see that the central part of the diatom's shape seems to remain more or less the same, while the tip of the diatom varies much more than the rest. We can therefore consider that the fourth component affects/ is partly responsible for this specific part of the diatom shape.

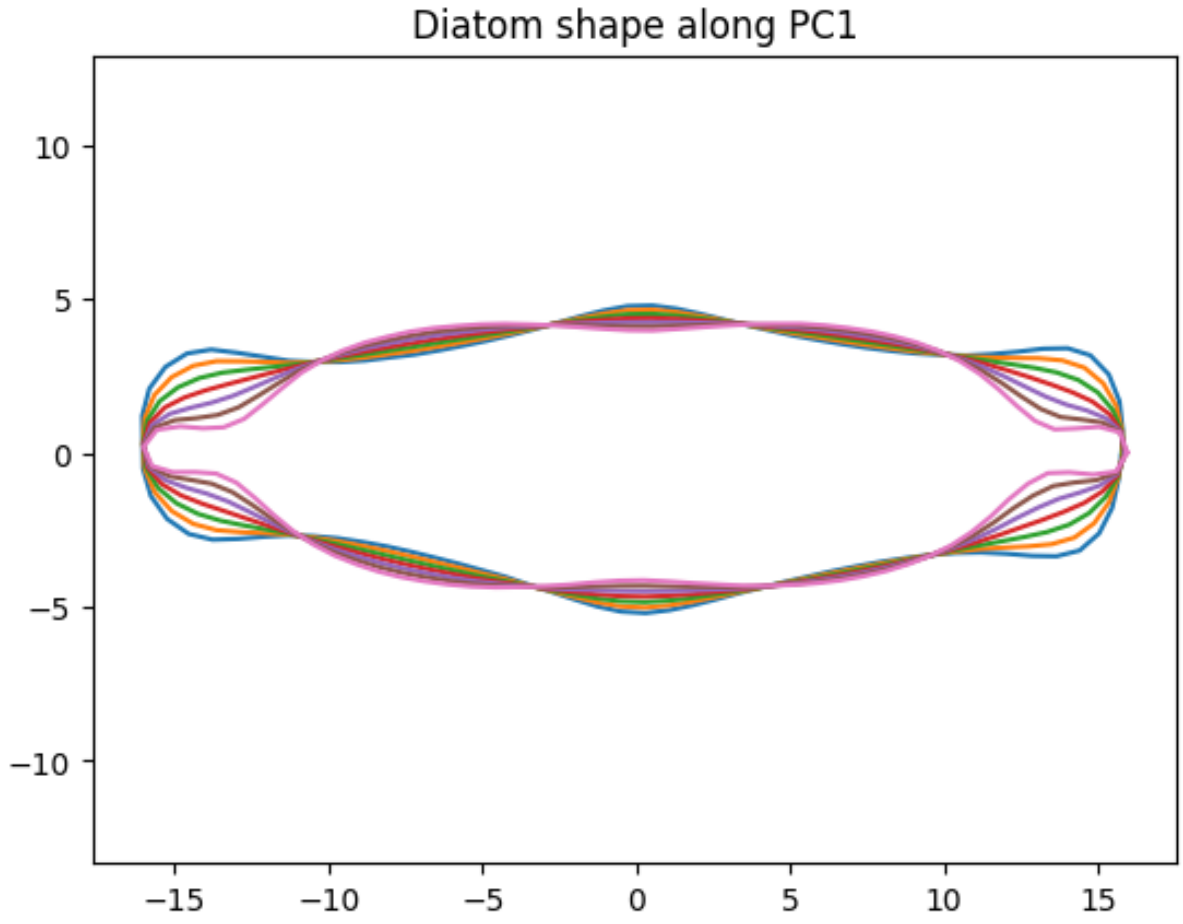


Figure 2 – Effect of the fourth eigenvector on variation in the shape of diatoms

## Exercise 2

Let : $g(X) = X^2$ and $Y = (X - \mu)^2$

We know that the square function is a convex function. Therefore, we can write the following expression using Jensen's formula :

$$E[(Y)^2] \geq [E(Y)]^2$$

We can now replace with the original variable :

$$E[((X - \mu)^2)^2] \geq [E((X - \mu)^2)]^2$$

We know that $E((X - \mu)^2 = \sigma^2$ so we can rewrite the inequality as :

$$E[((X - \mu)^2)^2] \geq (\sigma^2)^2 \Leftrightarrow E[(X - \mu)^4] \geq \sigma^4$$

## Exercise 3

### a)

Since we consider that this formula holds for the estimator, we can use it as it is to create our interval :

$$P(-c \leq \sqrt{n}\frac{\hat{\mu} - \mu}{\hat{\sigma}} \leq c) = \gamma$$

We can now transform the formula to isolate mu in the interval :

$$P(-c\frac{\hat{\sigma}}{\sqrt{n}} \leq \hat{\mu} - \mu \leq c\frac{\hat{\sigma}}{\sqrt{n}}) = \gamma$$
$$P(-\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}} \leq -\mu \leq -\hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n}}) = \gamma$$
$$P(\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}} \geq \mu \geq \hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n}}) = \gamma$$

### b)

Python

```
b) Not matching in 360 (out of 10000) experiments, 3.6%
```

### c)

Python

```
c) Not matching in 97 (out of 10000) experiments, 0.97%
```

Since (1) doesn't hold for $\hat{\sigma}$ here, we can no longer use the critical value $c$ of the Normal distribution, which must therefore be replaced by the critical value $t$ of the Student distribution. We then have :

$$P(\hat{\mu} - t\frac{\hat{\sigma}}{\sqrt{n}} \geq \mu \geq \hat{\mu} - t\frac{\hat{\sigma}}{\sqrt{n}}) = \gamma$$

When computing the actual values for this interval, we will need to pay attention to the degree of freedom needed to compute the critical value of the Student t-distribution.

Python

```
# [ ... ]
# Number of experiments to carry out
nexp = 10000


counter = 0
counter_c = 0
for i in range(nexp):
    x = np.random.normal(
        mu, sigma, n
    )  # simulates n realizations from a Gaussian with mean mu and var sigma^2
    sig = np.sqrt(np.var(x, ddof=1))
    fac1 = scipy.stats.norm.ppf(
```

```
        (1 - gamma) / 2, 0, 1
    )  # computes the 0.5% quantile of a Gaussian, roughly -2.576
    fac2 = scipy.stats.norm.ppf(
        (1 - gamma) / 2 + gamma, 0, 1
    )  # computes the 99.5% quantile of a Gaussian, roughly 2.576
    fac3 = scipy.stats.t.ppf(
        (1 - gamma) / 2, n - 1, 0, 1
    )  # 0.5% quantile for c) with student
    fac4 = scipy.stats.t.ppf(
        (1 - gamma) / 2 + gamma, n - 1, 0, 1
    )  # 99.5% quantile for c) with student
    xmean = np.mean(x)  # Sample mean
    a = xmean - fac2 * sig / np.sqrt(n)
    b = xmean - fac1 * sig / np.sqrt(n)
    ac = xmean - fac4 * sig / np.sqrt(n)  # for c)
    bc = xmean - fac3 * sig / np.sqrt(n)  # for c)
# [ ... ]
```

## Exercise 4

Let : $D = X - Y$ We then know that : $D \sim \mathcal{N}(\mu, \sigma)$
We can then compute the values for $D$ :

| Plant | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Replicate 1 without knockout ($X$) | 4.1 | 4.8 | 4.0 | 4.5 | 4.0 |
| Replicate 2 with knockout ($Y$) | 3.1 | 4.3 | 4.5 | 3.0 | 3.5 |
| Difference between the 2 replicates ($D$) | 1.0 | 0.5 | $-0.5$ | 1.5 | 0.5 |

In this situation, we want to know if the gene has an impact on the flowering time. If there is no effect of the gene knockout, we expect $\mu_X$ (the mean flowering time without knockout in population) to be equal to $\mu_Y$ (the mean flowering time with knockout in population) and therefore, the null hypothesis can be written as :

$$H0 : \mu_D = 0$$

In opposition to the null hypothesis and by watching the value for D, we can suppose that $\mu_X$ is bigger than $\mu_Y$ and therefore write the second hypothesis as :

$$H1 : \mu_D \geq 0$$

which means we will be using a right-tailed t-test. We can now do a t-test using the formula for paired samples :

$$T = \frac{\bar{D}}{S/\sqrt{n}} \rightarrow \text{ Student with } n - 1 \text{ dF}$$

To compute it, we first need to calculate $S^2$ and $\bar{D}$ :

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$
$$\bar{D} = \frac{1 + 0.5 - 0.5 + 1.5 + 0.5}{5} = 0.6$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$
$$S^2 = 0.55$$

We can now compute $T$ :

$$T = \frac{0.6}{\sqrt{0.55/5}} = 1.809$$

To know if the null hypothesis is rejected or not, we need to find the critical value for this test, knowing that this is a right-tailed t-test with a degree of freedom $dF = n - 1 = 4$. We also choose the significant level of $\alpha = 0.05$. we then need to find the following critical value : $t_{1-0.05;4}$

Using SciPy we can compute this value :

Python

```
>>> print(scipy.stats.t.ppf(1-0.05, 4))
 2.13184678133629
```

So we have $t_{1-0.05;4} = 2.13$ and can see that $T < t_{1-0.05;4}$ under $H0$.
So $H0$ is not rejected, and we can say that, in population, the mean flowering time for plant without gene knockout is not significantly higher than the mean flowering time for plant with gene knockout. Therefore, there is no significant effect of the said gene on the plant flowering time.

To see if the scientist can change the test result by copying the data set $k$ times, we have 2 options, computing the limit of the $T$ value and the limit of the critical value and see if, for a certain value of $n$, $T$ is greater than the critical value. We could see that, as $n$ approaches $+\infty$, $T$ approaches $+\infty$ too, and the critical value approaches $0$.

We can also, compute the results we could obtain by gradually increasing k. As expected, $T \to_{n\to+\infty}$ $+\infty$ and the critical value approaches $0$ when $n \to +\infty$. With this method (figure 3), we can also see that the test is rejected, i.e. when $T$ is greater than the critical value and this append for all $k$ greater or equal to $2$, $k \in \mathbb{N}^{*+}$.

Python

```
def compute_T(k):
    D = np.array([1.0, 0.5, -0.5, 1.5, 0.5])
    D = np.concatenate([D for i in range(k)], axis=0)
    return np.mean(D)/np.sqrt(np.var(D,ddof=1)/D.size)

T = [compute_T(k) for k in range(1,10)]
t = [scipy.stats.t.ppf(1-0.05, 5*k-1) for k in range(1,10)]
k = range (1,10)


plt.plot(k,T,'r')
plt.plot(k,t,'b')
plt.xlabel('number of duplicates of D')
plt.ylabel('T (red) and critical value (blue)')
plt.title(
    'computed T and critical value as a function of the number of duplication of D'
```

```
)
```

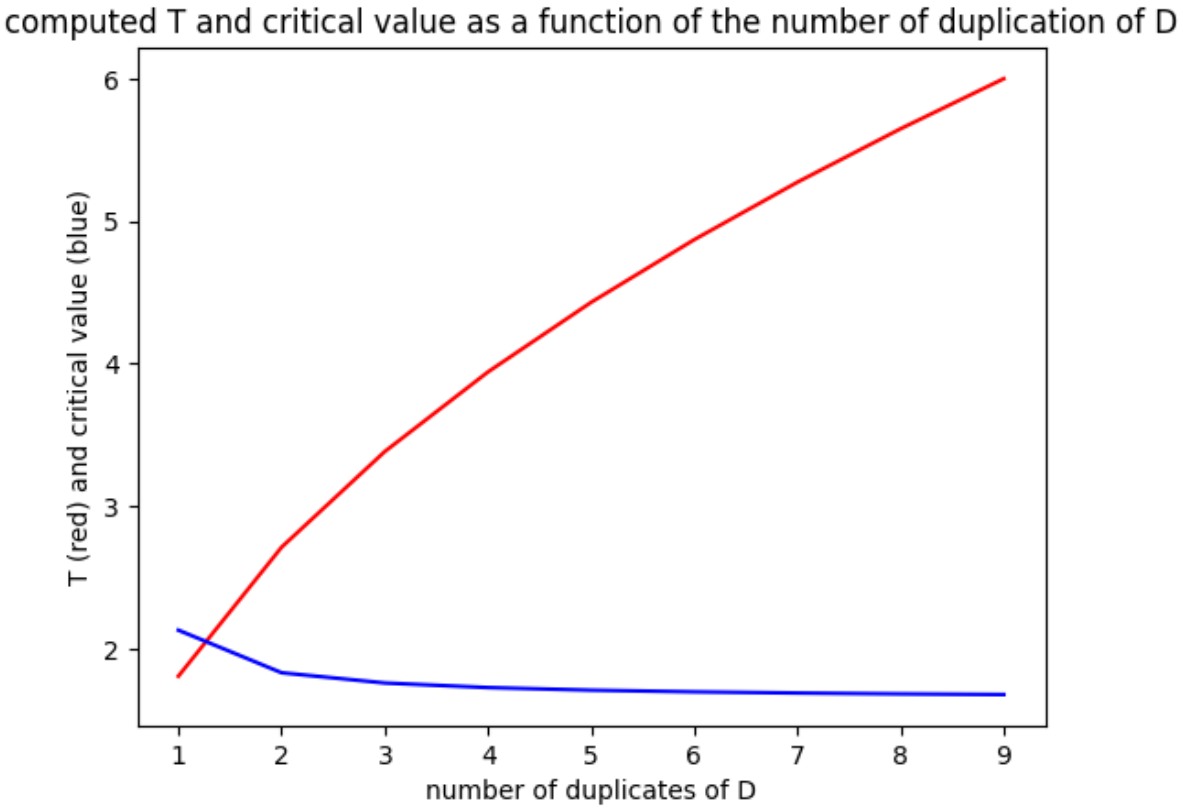


Figure 3 – T values (red) and critical values (blue) as a function of D duplicates