Estimate $P(X_j \mid Y)$ and $P(Y)$ directly from the training data by counting!

Sky	Temp	Humid	Wind	Water	Forecast	Play?
sunny	warm	normal	strong	warm	same	*yes*
sunny	warm	high	strong	warm	same	*yes*
rainy	cold	high	strong	warm	change	*no*
sunny	warm	high	strong	cool	change	*yes*

a)

P(water=warm|Play=Yes)=2/3

P(water=warm|play=no)=1

b) P(play=yes|water=warm)=2/3

p(play=no|water=warm)=1/3

c) p(play=yes|forecast=same)=1

p(play=yes|forecast-change=1/2

d) p(water=warm|play=yes)=2+1/3+2=3/5

p(water=warm|play=no)=1+1/2+1=2/3

2) Kernels

a) k(x, z) = a1k1(x, z) − a2k2(x, z), where a1, a2 > 0 are real numbers

This kernel is not semidefinite since if a1 and a2 are real numbers, there can be real numbers where a2k2(x,z) > a1k1(x,z) meaning the resulting kernel will no longer be positive definite. According to the Mercer's theorem, a kernel is only valid when it is positive definite and symmetric which in this case k is not positive definite for all terms, which makes it a invalid kernel function.

b) k(x, z) = f1(x)f1(z) + f2(x)f2(z), where f1, f2 : R n → R are a real-valued functions

we can look at the kernel matrix K_ij= k(xi,xj)=f1(xi)f1(xj)+f2(xi)f2(xj), we can show K will be positive semi -definite for any vectors. Let's use a=[a1,a2,…,an] to represent a set of input and apply to the kernel which we will get a^T K a >=0 (positive semi-definite)

If we evaluate the above function, we can get:

$a^T K a = \sum_1^i \sum_1^j a_i a_j k(x_i x_j)$

$=\sum_1^i \sum_1^j a_i a_j (f1(x_i)f1(x_j) + f2(x_i)f2(x_j))$

In this case, x_i and x_j can be further be simplify since for every pair of $a_i a_j$, there is a corresponding pair of $a_j a_i$ for the same $x_i x_j$. Thus j can be eliminated and the terms can simply be squared. For the same reason this can proof the kernel's symmetry.

$=\sum_1^i (a_i(f1(x_i))^2 + \sum_1^i (a_i(f2(x_i))^2$

Since all of the terms are getting squared , it would mean that the values for the kernel has to be positive meaning that it will be positive semi-definite. According to Mercer's theorem, a kernel will be valid if it is positive semi-definite and symmetric which this kernel is .

(c) [10 points] If $k(x, z) = e^{\frac{x^\top z}{\sigma^2}}$ is a valid kernel, prove that the Gaussian kernel $k(x, z) = e^{-\frac{||x-z||_2^2}{2\sigma^2}}$ is also a valid kernel

Comparing the two kernel by setting

$K_1(x,z)= e^{\frac{-x^T z}{\sigma 2}}$

$A_1=x^T z$

$K_2(x,z) = e^{\frac{-||x-z||_2^2}{2\sigma 2}}$

$A_2=||x - z||_2^2$

(we can omit the ½ term for now)

when trying to find equivalence between A_1 and A_2, A_2 is in squared Euclidean normal form which would equal to:

$A_2= ||x - z||_2^2=(x_1-z_1)^2+\ldots+(x_n-z_n)^2=(x_1+z_1-2x_1z_1)+\ldots+(x_n^2+z_n^2-2x_nz_n)$

Now to convert a_2 into vector form:

$A_2=Sum(x^2)+Sum(z^2)-2(x^T z)$

Substitute a2 into our original Gaussian kernel will yield:

$K_2(x,z)= e^{-\frac{Sum(x^2)+Sum(z^2)-2(x^T z)}{2\sigma 2}}$

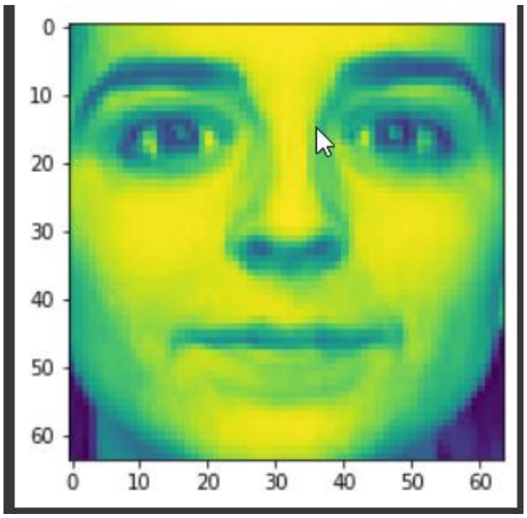
Simplify the equation be isolating the $-2(x^T z)$ will yield:

$K_2(x,z)= e^{-\frac{Sum(x^2)+Sum(z^2)}{2\sigma 2}} e^{\frac{(x^T z)}{\sigma 2}}$

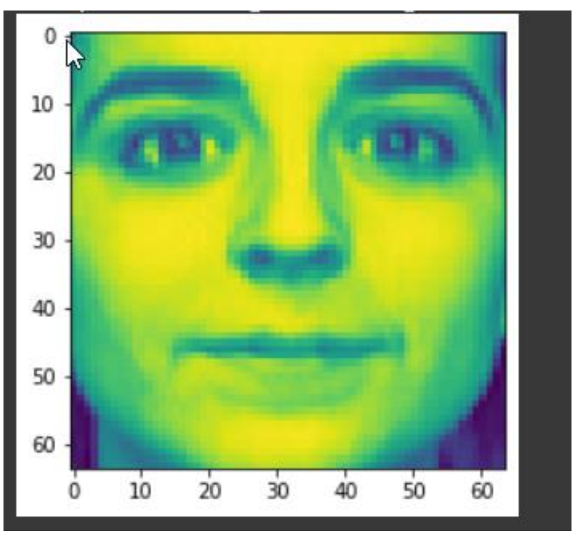
Notice that the second term is almost identical to k1 just missing the negative sign. This can be inferred to be inverse proportionality and since the parameter are evaluated exponentially, this means the kernel matrix will be positive semidefinite.

Symmetry can be found with $e^{\frac{(x^T z)}{\sigma 2}}$ since k1 is also a valid kernel and we can tell that $x_i z_j = x_j z_i$ which means $x_i^2 + z_j^2 = x_j^2 + z_i^2$. K_2 is already positive definite that exhibits symmetry which means the Mercer's theorem is satisfied and k2 is a valid kernel.

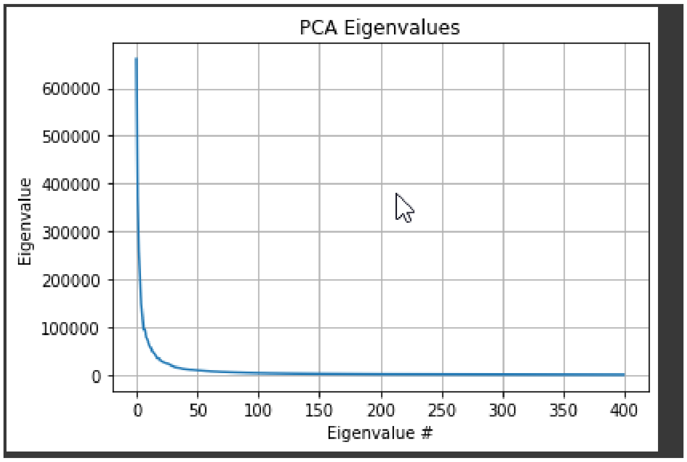
3a) The 100th face



b) With mean removed



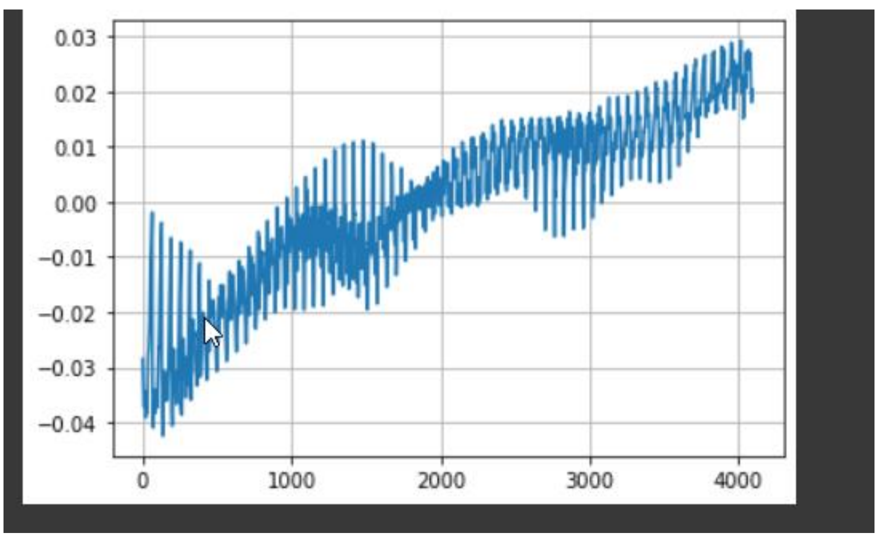
c) PCA on data matrix

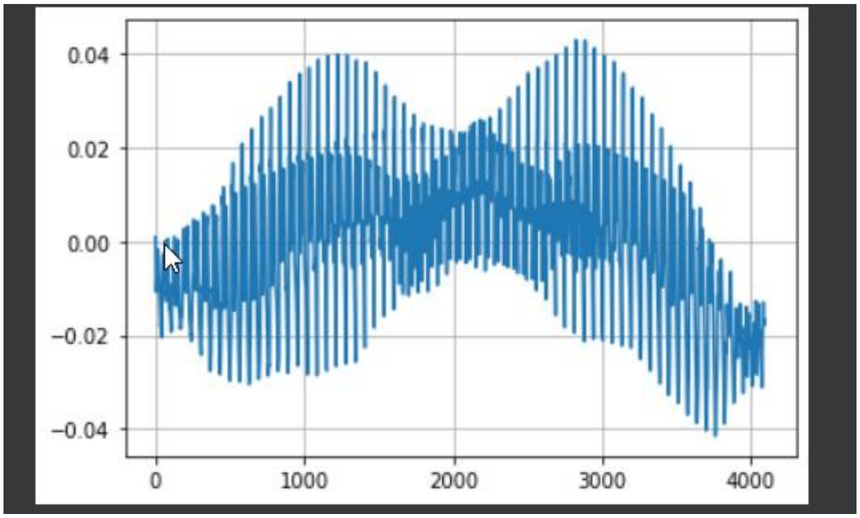


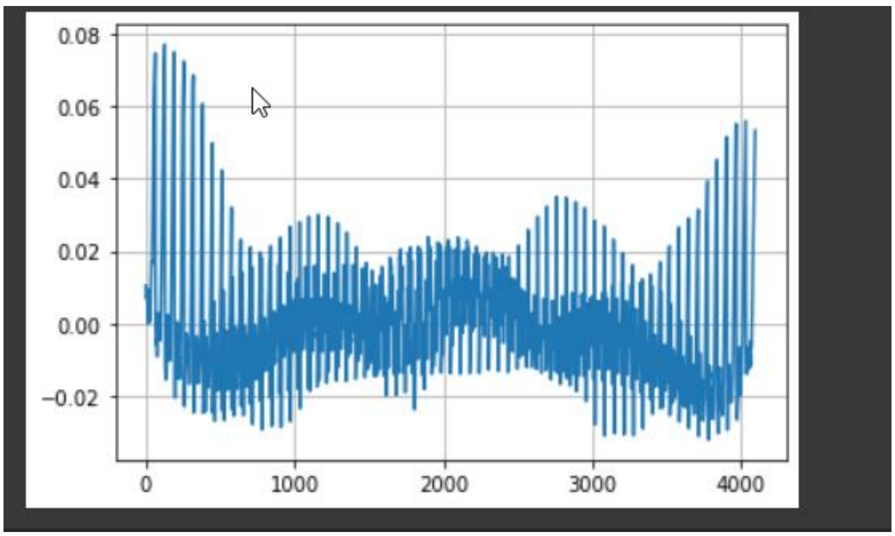
d) The 400th value is 0 because there is no more compression of data provided meaning very little variance. It can also be interpret as the other variance already been explained by the other 399 eigenvalues and the last one can't provide variance between faces

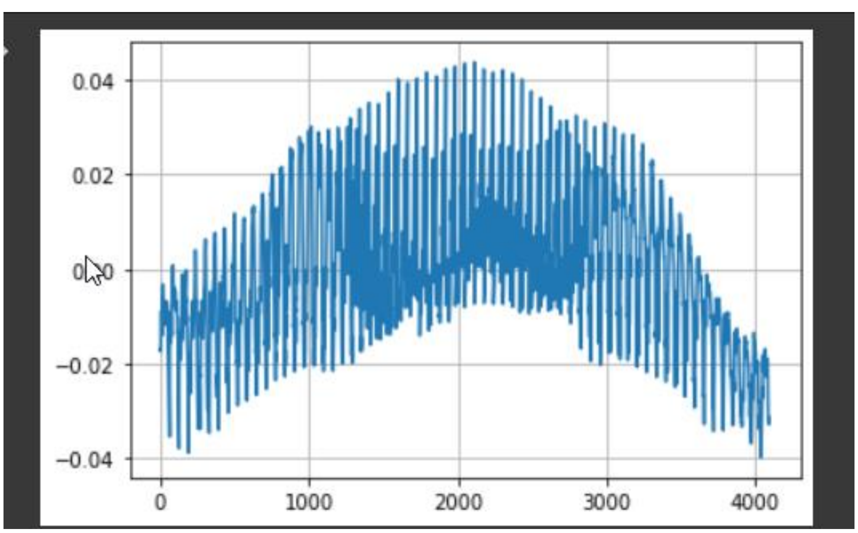
e) By looking at the previous plot, it seems the graph starts to flat out around 50 meaning there is a diminishing return on analyzing more components which costs additional computation power. Ideally it should be the elbow point on the graph but I think 50 can be used just in case.

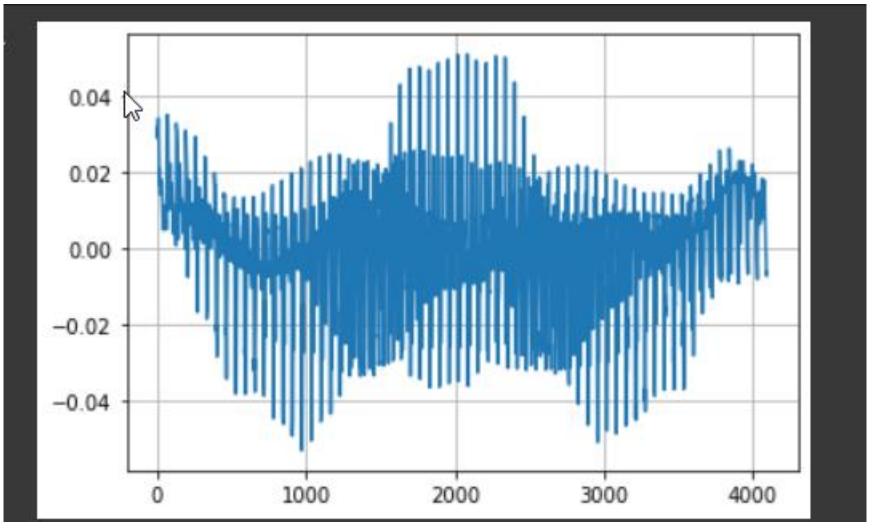
f)




0.04
0.02
0.00
−0.02
−0.04
0
1000
2000
3000
4000

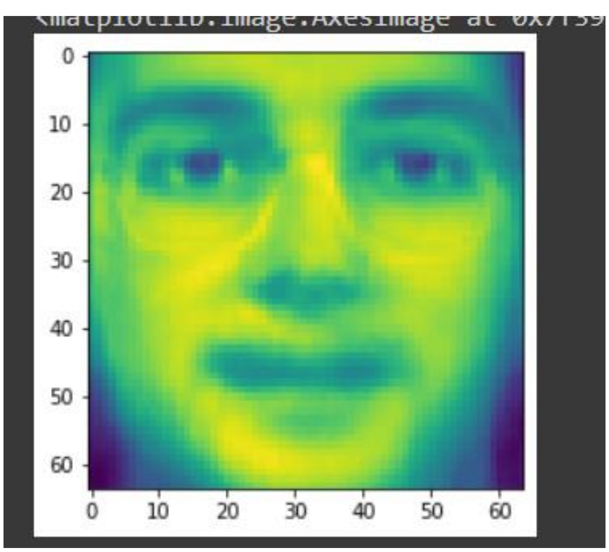

0.08
0.06
0.04
0.02
0.00
−0.02
0
1000
2000
3000
4000


0.04
0.02
−0.02
−0.04
0
1000
2000
3000
4000

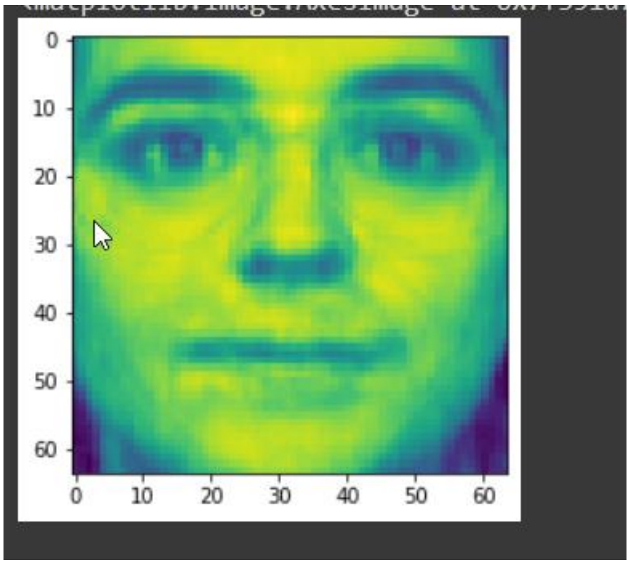


g)

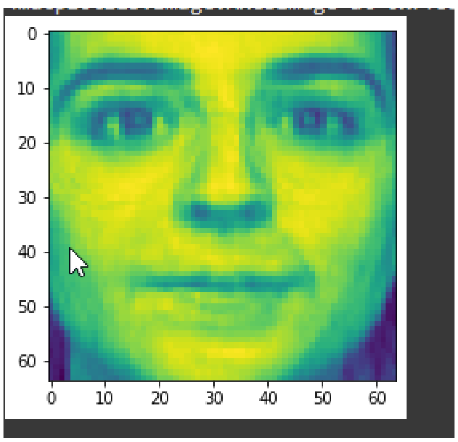
PCA#10



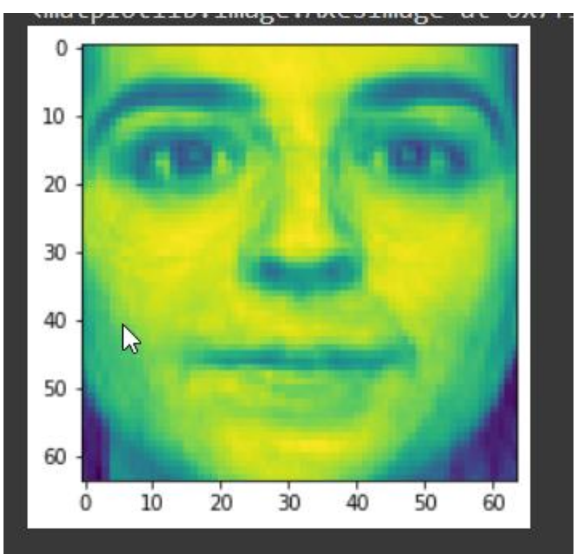
PCA#100

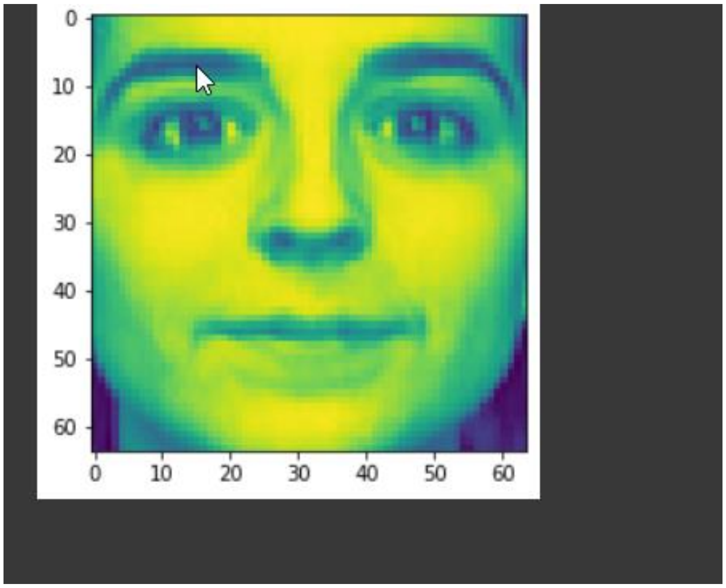


PCA#200



PCA#399




0
10
20
30
40
50
60
0
10
20
30
40
50
60