Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 2

Victor Manuel Gómez Espinosa

4 de marzo de 2020

1. Problema 1

1.1 Escalamiento

El procedimiento para contestar la pregunta sobre cómo cambian los componentes principales al realizar un cambio de escala en una variable consistió en tomar una matriz de varianzas y covarianzas (S), de esta obtener la matriz de correlación (R)(1.1) y a continuación obtenemos los componentes principales para ambas (1.2), es decir aplicando PCA sin estandarización y con estandarización.

$$S = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, R = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{1.1}$$

$$\begin{bmatrix} 0.7 & -0.7 \\ 0.7 & 0.7 \end{bmatrix}, \begin{bmatrix} 0.7 & -0.7 \\ 0.7 & 0.7 \end{bmatrix} \tag{1.2}$$

Posteriormente definimos una segunda matriz de varianzas y covarianzas (S_2) mediante una constante de escalamiento $(c = 10)$, la cual aplicamos sobre la primera variable y a esta matriz (1.3) le aplicamos el mismo procedimiento anteriormente descrito para obtener sus componentes principales para el casi sin estandarización y con estandarización (1.4).

$$S_2 = \begin{bmatrix} c^2 1 & c0.5 \\ c0.5 & 1 \end{bmatrix} = \begin{bmatrix} 100 & 5 \\ 5 & 1 \end{bmatrix}, R_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{1.3}$$

$$\begin{bmatrix} -0.99 & 0.05 \\ -0.05 & -0.99 \end{bmatrix}, \begin{bmatrix} 0.7 & -0.7 \\ 0.7 & 0.7 \end{bmatrix} \tag{1.4}$$

Observe que, aunque las matrices de varianzas y covarianzas (S, S_2) cambian debido al escalamiento, las matrices de correlación permanecen igual (R, R_2), por lo que los componentes de esta ultima no cambian, mientras que para los componentes de ambas sin estandarizar, si obtenemos resultados diferentes.

Por lo tanto, podemos concluir que aplicar PCA posterior a haber realizado un cambio de escala nos dará los mismos resultados sólo después de haber estandarizado.

1.2 PCA con y sin normalización

Para este problema se utilizaron los datos de numero reportado de muertes en los 50 estados de los Estados Unidos durante 1985, clasificado de a cuerdo a 7 categorías, que son, accidente, cardiovascular, cáncer, pulmonar, neumonía, diabetes e hígado.

Primero se realizó una exploración visual a los datos, para buscar si había variaciones considerables entre sus variables y de aquí considerar si es conveniente estandarizar o no. Observe en la Figura 1.1 que los rangos para las variables cardiovascular y cáncer sobresalen de las demás, por lo que en este caso se recomienda estandarizar.

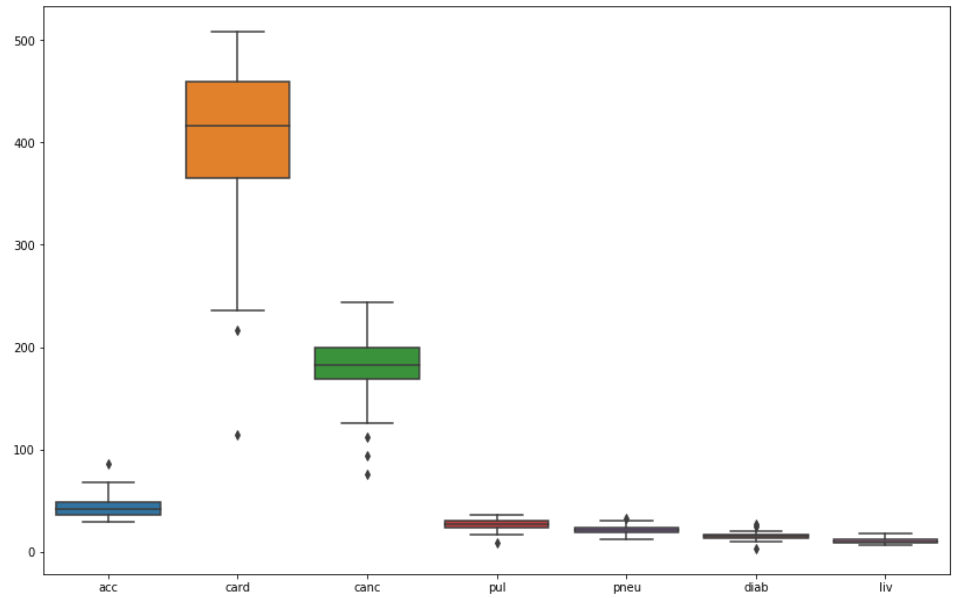

Figura 1.1: Boxplot de las 7 causas de muertes.

Aunque la recomendación es estandarizar, aplicamos PCA con y sin estandarización para explorar las diferencias. Observe en la Figura 1.2 a) que, sin estandarizar, la varianza acumulada para sólo un componente es mayor al 90%, mientras que estandarizando (Figura 1.2 b), para obtener aproximadamente el mismo nivel de varianza acumulada necesitamos más de 4 componentes.

Dicho de otra forma, para el primer caso, un solo componente explica la mayor parte de la información mientras que los otros prácticamente no aportan nada.

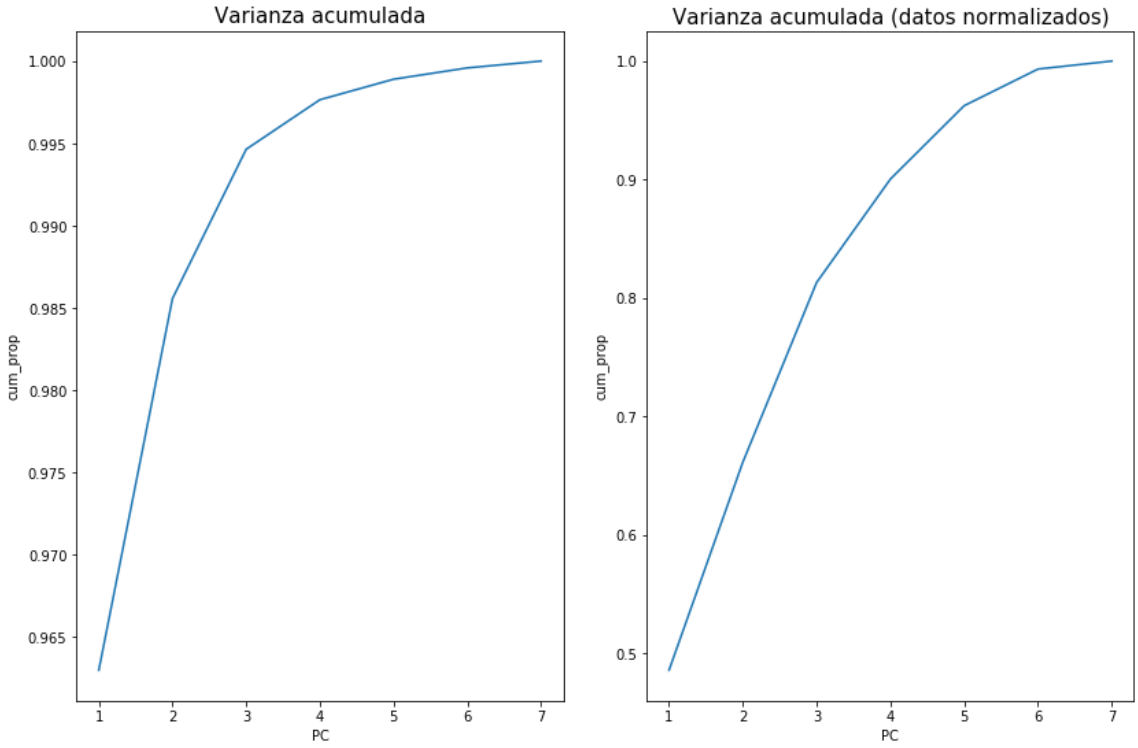

Figura 1.2: a) Varianza acumulada sin estandarizar y b) estandarizando.

Respecto a los componentes, el efecto de no estandarizar o si, se ve reflejado en los pesos de las variables, es decir si no estandarizamos ciertas variables pueden tener más peso que otras. Por ejemplo observe los componentes del ejercicio 1 (1.4) y note como sin estandarizar el peso de la primer variable (a la que se le aplico el escalamiento) es mucho mayor que para la segunda $(-0.99, -0.05)$, mientras que estandarizando es equilibrado $(0.7, 0.7)$. Por lo tanto, a fin de no dar más peso a ciertas variables que a otras, se recomienda estandarizar.

Posteriormente graficamos un Biplot (Figura 1.3) sobre los dos primeros componentes principales y lo primero que saltó a la vista es un posible dato atípico para AK, por lo que fue removido (Figura 1.4) para ver el efecto sobre los componentes (para esto se debe consultar al experto en el tema de estudio para saber si quitar o no los datos atípicos).

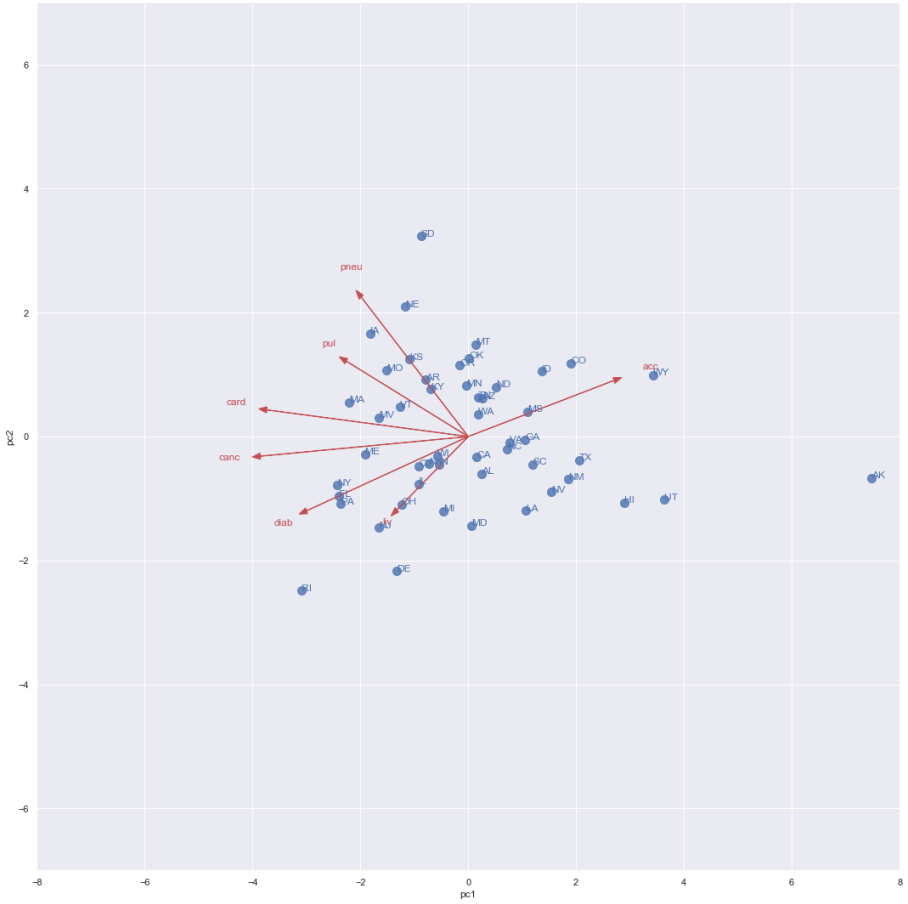

Figura 1.3: Biplot para los primeros dos componentes.

Observe que después de remover el dato atípico (Figura 1.4), el Biplot cambia, note la dirección de las flechas rojas que son las direcciones de variación de las variables, en este caso, las causas de muerte.

Note que las 7 causas de muerte se pueden catalogar en accidentes o por enfermedad y dentro de esta última podemos catalogar por vías respiratorias (pulmonar y neumonía), las cuales parecen prácticamente una misma flecha, lo cual nos muestra su relación, de igual forma, observe que cáncer e hígado están muy relacionadas, y de igual manera, aunque en menor grado, cáncer con causas cardiovasculares y diabetes con el hígado, lo cual tiene sentido.

Por otro lado, se puede observar también que las causas por accidente van en una dirección un poco diferente a la de las enfermedades (de las cuales cáncer cardio y diabetes son las que mayor peso tienen), sólo un poco cerca respecto a las causas por vías respiratorias (¿habrá casos que estén relacionados?).

Estas relaciones entre las causas de muerte parecen interesantes, pero nuevamente, hay que consultar al experto en el tema si en verdad estas relaciones tienen sentido y si ayudan a explicar el fenómeno.

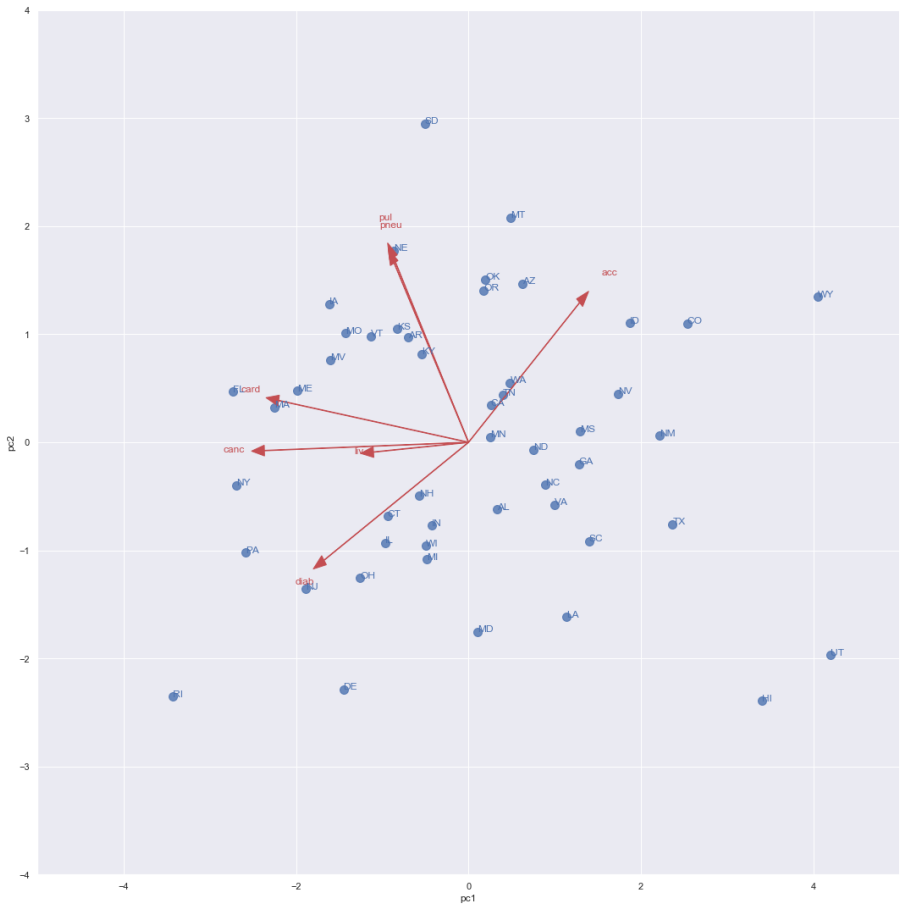

Figura 1.4: Biplot para los primeros dos componentes, después de remover datos atípicos.

2. Problema 2

En este problema se utilizaron las imágenes de Olivetti_faces del sklearn que contiene 10 fotos de 40 sujetos, el cual se separó en dos conjuntos, uno de entrenamiento (primeras 9 fotos de cada sujeto) y otro de prueba (ultima foto de cada sujeto).

Primero se estandarizó el conjunto de prueba, al cual se le aplico PCA y se determinó utilizar 30 componentes ya que con estos tenemos aproximadamente el 80% de la varianza acumulada (Figura 2.1).

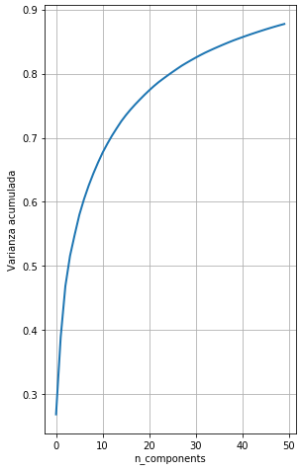

Figura 2.1: Varianza acumulada contra componentes.

Observe en la Figura 2.2, que para el primer componente lo que más se resalta (lo obscuro), es la forma general del rostro, sin mucho detalle en ojos boca y nariz, mientras que para el segundo componente parece resaltar más el lado izquierdo y menos el derecho, en otras palabras, esto me sugiere que lo que esta tomando en cuenta es la rotación del rostro en el eje vertical.

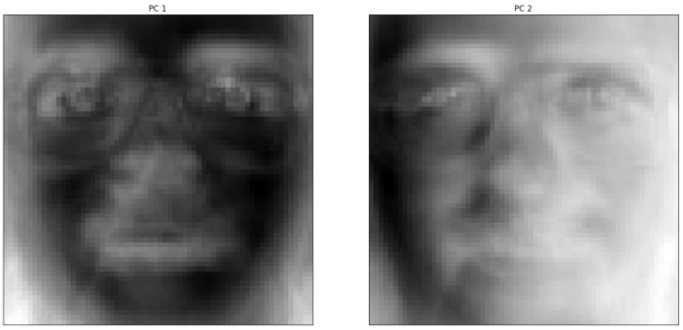

Figura 2.2: Primeros dos componentes principales.

Posteriormente se graficaron las proyecciones en los primero dos componentes (Figura 2.3), donde se pueden observar algunos patrones de agrupamiento por color, que nos indican al mismo sujeto.

También se pudo observar otros patrones interesantes, por ejemplo, lo que parecerían datos atípicos en los extremos, pero en realidad lo que vemos es la rotación de los rostros sobre el eje vertical, de igual forma para los puntos en los extremos del eje vertical, muestran un grado de inclinación, por ejemplo, los rostros que se encuentran en el extremo superior parecen tener una ligera inclinación hacia abajo y los de abajo hacia arriba.

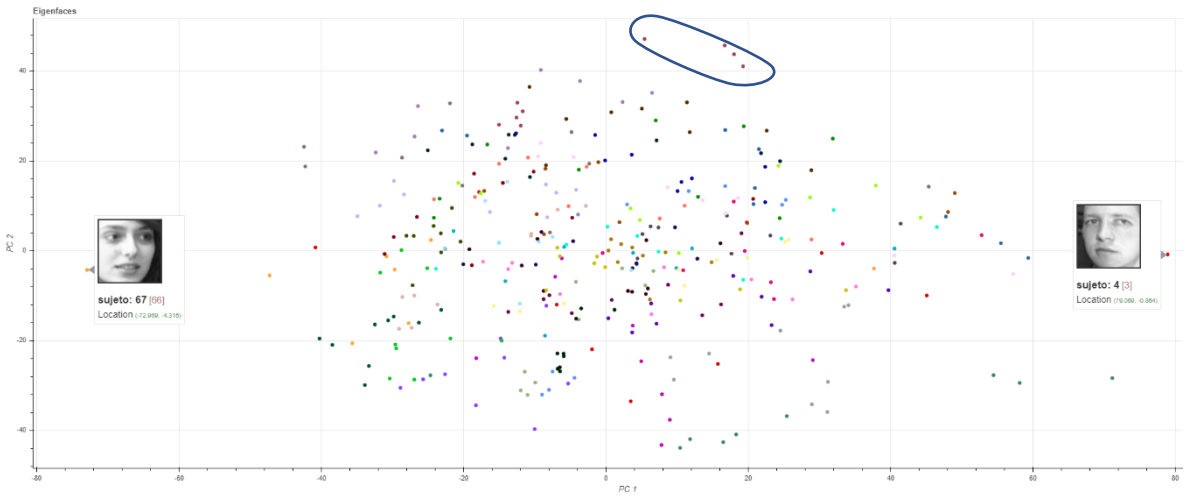

Figura 2.3: Proyección de los primeros dos componentes.

Posteriormente utilizamos el conjunto de prueba, al cual primero se estandarizó y después proyectó sobre los componentes principales y mediante el método del vecino más cercano, se buscó el rostro más parecido, es decir partiendo del hecho de que sabemos que el sujeto se encuentra en el conjunto de entrenamiento, probamos si se podía encontrar (Figura 2.4).

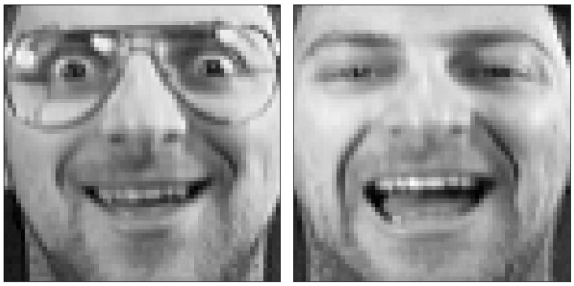

Figura 2.4: a) Foto de prueba, b) foto más cercana en el conjunto de entrenamiento, con datos estandarizados y utilizando 30 componentes principales.

Se encontró que para la mayoría de los sujetos funcionó, y solo hubo pequeños errores para el primer vecino cercano, pero para el segundo vecino se acertó (Figura 2.5).

Figura 2.5: Con datos estandarizados y 30 componentes, parte superior error para el primer vecino mas cercano, parte inferior, acierto para el segundo más cercano.

Posteriormente surgió la duda de con cuantos componentes seguiría siendo buena la predicción para el sujeto de la Figura 2.4, y se encontró que a partir de 7 componentes con los datos estandarizados (aproximadamente el 60% de la varianza acumulada, Figura 2.1) funciona correctamente (Figura 2.6).

Figura 2.6: a) Foto de prueba, b) foto más cercana en el conjunto de entrenamiento, con datos estandarizados y utilizando 7 componentes principales.

Ahora la duda fue, si continuaría acertando para 7 componentes, pero sin estandarizar los datos, y se encontró que no fue el caso (Figura 2.7) para más de un sujeto.

Figura 2.7: a) Foto de prueba, b) foto más cercana en el conjunto de entrenamiento, con datos no estandarizados y utilizando 7 componentes principales.

Ahora la pregunta es, ¿que pasa si ahora comparamos un sujeto que no está dentro del conjunto de prueba?, para esto se siguió el mismo procedimiento, se estandarizó y se proyectó utilizando 30 componentes del conjunto de entrenamiento y posteriormente se buscó el vecino más cercano.

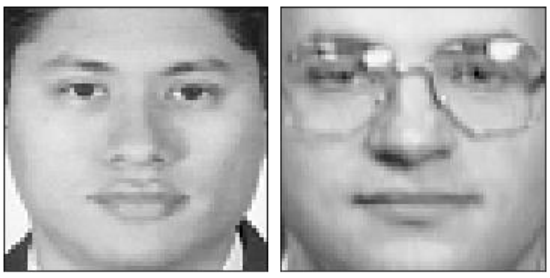

Figura 2.8: a) Foto fuera del conjunto de prueba, b) foto más cercana en el conjunto de entrenamiento, con datos estandarizados y utilizando 30 componentes principales.

Observe la Figura 2.8 y note que dio como resultado al más parecido dentro del conjunto de prueba, obviamente ya se esperaba este resultado, puesto que ya se sabía que el sujeto no pertenece al conjunto de entrenamiento, pero ¿de que forma podríamos detectar automáticamente estos casos?

Una solución podría consistir en utilizar la distancia máxima que hay entre el conjunto de prueba y el de entrenamiento (en este ejemplo 39.5), y si la distancia es considerablemente mayor a esta, podríamos considerar que no se encuentra en los datos, por ejemplo la distancia del sujeto fuera del conjunto de entrenamiento fue de 56, es decir 56>39.5, y por lo tanto no se encuentra en los datos.