Estadística Multivariada
Cuestionario PCA

Valeria Zamora Israel Garcia Mateo Valencia Marian Medina André Chavéz

15 de Abril de 2024

Ejercicio 1.

Diga si los siguientes enunciados son verdaderos o falsos. Argumente su respuesta.

1. La *i-ésima* componente principal se toma como la dirección que es ortogonal al (i-1)-*ésimo* componetnte principal y maximiza la variabilidad restante.
2. Distimtos componentes principales están linealmente no correlacionadas.
3. La dimensión de los datos originales es siempre mayor que la dimensión de los datos transformados por un PCA.

Respuestas

1. **Verdadero.** Se escoge la i-ésima componente principal a manera de que sea ortogonal a las componentes principales anteriores y que maximize la variabilidad restante.
   Esto asegura que cada componente capture la mayor cantidad de varianza posible que no ha sido explicada en los componentes anteriores.
2. **Verdadero.** Si los componentes principales son ortogonales, eso quiere decir que son linealmente no correlacionados.
   Cada componente captura una parte única de la varianza sin influenciarse por los otros componentes.
3. **Falso.** La dimensión puede ser igual o menor, dependiendo de cuanta varianza se tenga en los componentes principales.
   Normalmente se conservan solo los componentes más importantes.

Ejercicio 2.

Suponga que se tiene la siguiente matriz de covarianza

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{1}$$

Calcula la primer componente principal **Respuesta**
Método:

1. Conseguir datos
2. Quitarle la media a cada dato por fila y columna para obtener un dataset con media cero.
3. Calcular la matriz de covarianza.
4. Calcular los eigenvalores y los eigenvectores de la matriz de covarianza.

5. Escoger los componentes y formar un vector característico.
    - Aquí entra la compresión de datos.
    - Es importante ordenar los eigenvalores de mayor a menor. Esto indica el orden de significancia.
    - El eigenvector con el eigenvalue más alto es el componente principal de ese dataset.
    - Hacer el vector característico, que es crear una matriz con los eigenvectores.

Como en este caso ya contamos con la matriz de covarianza, solo necesitamos calcular los eigenvalores y eigenvectores. El Primer componente será el eigenvector del eigenvalor más grande.

$$det(P - \lambda I) \tag{2}$$

$$\begin{pmatrix} 1-\lambda & 0 & 0 \\ 0 & 2-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{pmatrix} = (-1)^2(1-\lambda)\begin{pmatrix} 2-\lambda & 0 \\ 0 & 2-\lambda \end{pmatrix} \tag{3}$$

$$= (1-\lambda)[(2-\lambda)(2-\lambda)] \tag{4}$$

$$= (1-\lambda)(4-4\lambda+\lambda^2) \tag{5}$$

$$= (1-\lambda)(\lambda-2)(\lambda-2) = 0 \tag{6}$$

$$\lambda_1 = 1 \tag{7}$$

$$\lambda_2 = 2 \tag{8}$$

$$\lambda_3 = 2 \tag{9}$$

Código para PCA

```
import numpy as np

X = np.random.randint(10, 50, 100)
np.mean(X)
Y = X.reshape(20,5)
np.mean(Y, axis = 0) # Calcula la media por columnas y las muestra en un arregl
np.mean(Y, axis = 1) # Calcula la media por renglones y las muestra en un arreg

# Centramos a los datos
Y_media = Y - np.mean(Y, axis = 0)

# Matriz de covarianza
S = np.cov(Y_media, rowvar = False)
print(S)

# Vamos a calcular eigenvalores y eigenvectores
eigen_val, eigen_vec = np.linalg.eigh(S)

# Accesos
eigen_val[::-1]

# Reescribimos a los eigenvalores de manera decreciente
eigen_val = eigen_val[np.argsort(eigen_val)[::-1]]
eigen_vec = eigen_vec[:,np.argsort(eigen_val)[::-1]]
n_components = 2
eigenvector_2 = eigen_vec[:,0:n_components]
```

```
# Descomposicion espectral
Y_red = np.dot(eigenvector_2.transpose(),Y_media.transpose()).transpose()

# Informacion total
eigen_val_total = sum(eigen_val)

varianza_explicada = [(i/ eigen_val_total )*100 for i in eigen_val ]

varianza_explicada = np.round(varianza_explicada, 2)

varianza_explicada_acumulada = np.cumsum(varianza_explicada)

print("Varianza explicada: {}".format(varianza_explicada))

print("Varianza explicada acumulada: {}".format(varianza_explicada_acumulada))

# Escribimos lo anterior como funcion
def PCA(X , num_componentes):

    X_media = X - np.mean(X , axis = 0)

    cov_mat = np.cov(X_media , rowvar = False)

    eigen_val , eigen_vec = np.linalg.eigh(cov_mat)

    sorted_index = np.argsort(eigen_val)[::-1]
    sorted_eigenval = eigen_val[sorted_index]
    sorted_eigenvec = eigen_vec[:,sorted_index]

    eigenvector_ = sorted_eigenvec[:,0:num_componentes]

    X_red = np.dot(eigenvector_.transpose() , X_media.transpose() ).transpose()

    return X_red
```

Ejercicio 3.

Suponga que se tiene la tabla.

Observación	X_1	X_2
1	-2	2
2	2	-2

(10)

$b_{11} \rightarrow 0{,}7071$

$$b_1 = \begin{pmatrix} 0{,}7071 \\ b_{12} \end{pmatrix} \tag{11}$$

$$b_{12} < 0$$

$$b_{11}^2 + b_{12}^2 = 1$$

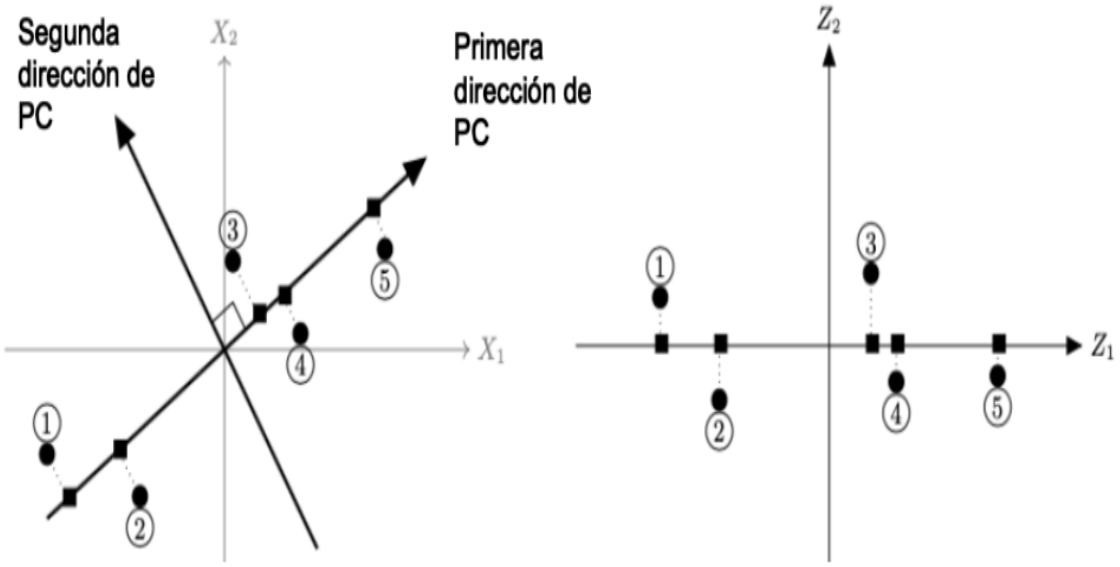
$$b_{12} = -0{,}7071$$

$$Xb_1 = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 0{,}7071 \\ -0{,}7071 \end{pmatrix} \tag{12}$$

$$Xb_1 = \begin{pmatrix} -2{,}8284 \\ 2{,}8284 \end{pmatrix} \tag{13}$$

Ejercicio 4.

Explica, con todo detalle, la siguiente figura:



Si comparamos los ejes de la perimera dirección de PC y la segunda dirección de PC, se forma un ángulo de 90^{o}, lo que quiere decir que los datos están correlacionados. Pero, como estamos realizando un análisis de componentes principales, no queremos que los datos estén correlacionados. Por lo tanto, tenemos que realizar la rotación de los datos para quitar la correlación de los datos y hacer que estos sean ortogonales. Esto último se ve en la gráfica de la derecha.

Por otra parte, hay datos que tienen una parte positiva, cómo el 1 y el 3, pero hay otros que tienen loadings negativos, cómo el 5, quien tiene el más negativo, mientras que 3 es el que tiene el más positivo.

PCA
Su objetivo es reducir la dimensión
No considera información de clases
Usa la máxima varianza total
Transforma las variables en componentes principales
No es tan sensible porque no considera clases
Reduce dimensiones de grandes cantidades de datos
Usa eigenvalores y eigenvectores

(14)

LDA
Su objetivo es separar las clases
Sí utiliza información de clases
Máxima varianza entre clases y minimiza la varianza dentro de las clases
Transforma las variables en una combinación lineal
Sensible al desbalance de las clases
También reduce dimensiones de datos, clasifica y reconoce patrones
Usa eigenvalores y eigenvectores

(15)

Ejercicio 5.

Sea (X_1, X_2, X_3) un vector gaussiano N (0,1) y sean $\beta \in \mathbb{R}, \sigma \geq 0$. Definase,

$$Y_1 = 0{,}5X + 2X_1, Y_2 = -0{,}5X + 2X_2 \tag{16}$$

Sea $Y = (Y_1, Y_2)^T$

(1) Explica, con todo detalle, que Y tiene una distribución gaussiana con parámetros que deberás encontrar. Calcula los eigenvalores de la matroz de covarianza de Y.

(2) Calcula, en función de Y_1 y Y_2 y luego en función de las componentes de X, las componentes principales ξ_1 y ξ_2 asociadas a Y. Muestra además que $Var(\xi_i) = \lambda_i$ y $Cov(\xi_1, \xi_2) = 0$.

(3) Calcula $\rho_{i,j}$, y verifica que

$$\rho_{i,1}^{-2} + \rho_{i,2}^{-2} = 1, i = 1, 2 \tag{17}$$

Respuestas: Dado que X,X_1, X_2 son v.a. gaussianas con media cero y matriz de covarianza identidad W(0,I) y $Y_1 = 0{,}5X + 2X_1$ y $Y_2 = -0{,}5X + 2X_2$, podemos expresar Y como

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 0{,}5 & 2 \\ -0{,}5 & 2 \end{pmatrix} \begin{pmatrix} X \\ X_1 \end{pmatrix} \tag{18}$$

Y como Y es una combinación lineal de variables gaussianas, Y es gaussiano.
Para mi matriz de covarianza:

$$Cov(Y) = \begin{pmatrix} 0{,}5 & 2 \\ -0{,}5 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0{,}5 & -0{,}5 \\ 2 & 2 \end{pmatrix} \tag{19}$$

$$Cov(Y) = \begin{pmatrix} 5{,}25 & 4 \\ 4 & 5{,}25 \end{pmatrix} \tag{20}$$

Los valores propios se encuentran con: $|Cov(Y) - \lambda I = 0|$.
Y da como resultado: $\lambda_1 = 9{,}25 y \lambda_2 = 0{,}75$
Y los eigenvectores que tenemos son:
Para $\lambda_1 = 9{,}25$:

$$v_1 = \begin{pmatrix} 0{,}803 \\ -0{,}596 \end{pmatrix} \tag{21}$$

Para $\lambda_2 = 0{,}75$:

$$v_2 = \begin{pmatrix} 0{,}596 \\ 0{,}803 \end{pmatrix} \tag{22}$$

Y como los eigenvectores son ortogonales,$Cov(\xi_1, \xi_2) = 0$.

Ejercicio 6.

Tienes los siguientes datos estandarizados: Supongamos que el eigenvector de la primera componente principal del conjunto de datos es (0.707,-0.5,-0.5). Calcula la proporción de la varianza explicada por el primer componente.

i	X_1	X_2	X_3
1	-0.577	1	-1
2	-0.577	1	1
3	-0.577	-1	1
4	1.732	-1	-1

(23)

1. Encontrar La varianza total de los datos. Como los datos están estandarizados, la varianza total es la suma de las varianzas de cada variable.

$$Vardecadavariable = 1 Vartotal = 4 \tag{24}$$

2. Calcular la varianza del primer componente principal, que es el cuadrado del componente correspondiente del eigenvector proporcionado.

$$Var(PC_1) = (0{,}707)^2 = 0{,}499849 \tag{25}$$

3. Calcular la proporción de la varianza explicada por el primer componente principal.

$$\frac{Var(PC_1)}{Var_{total}} = \frac{0{,}499849}{4} = 0{,}124962 = 12{,}49\,\% \tag{26}$$

Ejercicio 7.

Diga si las siguientes enunciados son verdaderos o falsos. Argumente su respuesta.

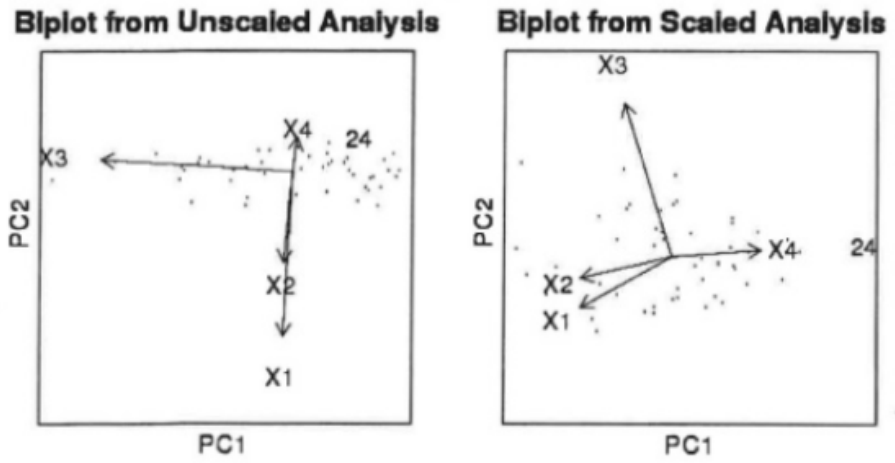
1. La proporción de varianza explicada por un componente principal adicional nunca disminuye a medida que se agregan más componentes principales.
2. La proporción acumulativa de la varianza explicada nunca decrece a medida que se agregan más componentes principales.
3. Al usar todas las posibles componentes principales, nos provee de un mejor entendimiento de los datos.
4. La *gráfica sree* provee un método para determinar el número de componentes principales a usar.

Respuestas:

1. Falso. Sí puede disminuir porque los componentes capturarían menos varianza o puede haber un error en el modelado por los nuevos componentes.
2. Verdadero. La proporción acumulativa aumenta o se mantiene igual ya que cada componente captura una parte de la varianza.
3. Falso. Usar todas puede hacer lo contrario porque algunos componentes pueden acumular muy poca variabilidad. La selección de los mejores componentes se basa en la proporción de varianza que expliquen.
4. Verdadero. Esta gráfica muestra la varianza que explica cada componente. El punto en la gráfica donde cambia la pendiente indica el número que se debe trabajar.

Ejercicio 8.

Imagina que realizas un análisis por componentes principales sobre las mismas cuatro variables en un conjunto de datos particular: X_1, X_2, X_3 y X_4. El primer análisis sólo centra las variables pero no los escala, y el segundo análisis centra y escala las variables. Los *biplots* de las primeras dos componentes principales que se producen en ambos análisis se muestran a continuación. Se etiqueta una locación 24 también. Dadas las siguientes aseveraciones:



I. X_1 está más correlacionado con X_2 que con X_3.

II. X_3 tiene la varianza más alta de las cuatro variables.

III. La observación 24 tiene un valor positivo, relativamente grande, para X_4.

¿Cuáes de las aseveraciones anteriores pueden ser demostradas visualmente en las gráficas anteriores? Justifica tu respuesta.

Observe, de la gráfica escalada, que PC2 es una medida de X_3. Ya que la observación 2 tiene un *score* cercano a cero, su valor X_3 debe estar cercano al valor promedio.

Respuestas: La primera es cierta, porque el ángulo entre X_1 y X_3 es de 90°o incluso más, con lo que indica una relación negativa entre componentes.
La segunda también es cierta debido a que este componente es el más cercano al margen de los biplots, lo que indica una varianza muy alta. // La tercera es falsa, puesto que no se puede interpretar el valor de un número de esa manera.

Ejercicio 9.

Imagina que, como actuario, estás revisando un conjunto de datos con 100 observaciones en cuatro variables: X_1, X_2, X_3 y X_4. Quieres analizar tales datos realizando usando dos componentes principales:

$$\xi_1 = b_{11}X_1 + b_{21}X_2 + b_{31}X_3 + b_{41}X_4$$
$$\xi_2 = b_{12}X_1 + b_{22}X_2 + b_{32}X_3 + b_{42}X_4$$

Tienes las siguientes aseveraciones:

I. $\sum_{i=1}^{100}\left(\sum_{j=1}^{4} b_{j1}x_{ij}\right)^2 = \sum_{i=1}^{100}\left(\sum_{j=1}^{4} b_{j2}x_{ij}\right)$.

II. $\sum_{i=1}^{4} b_{j1}b_{j2} = 0$.

III. $\sum_{i=1}^{4} b_{j1}^2 + \sum_{j=1}^{4} b_{j2}^2 = 1$.

Determina, justificadamente, cuáles aseveraciones son verdaderas. **Respuestas:**

I. Es verdadero si los 2 primeros componentes principales capturan la misma cantidad total de variabilidad en los datos. Dado que ξ_1 y ξ_2 son ortogonales y representan direcciones distintas de variabilidad, la cantidad total de variabilidad capturada por ambos componentes principales es igual a la varianza total de los datos.

II. Esta aseveración se refiere a la covarianza entre los coeficientes de los primeros dos componentes principales. Si los dos componentes son ortogonales, su covarianza es cero. Como sabemos que ξ_1 y ξ_2 sí lo son, es verdadero.

III. Aquí se habla de la suma de los cuadrados de los coeficientes de cada componente principal. Si los componentes están estandarizados, la suma de los cuadrados de los coeficientes de cada componente principal es igual a 1 y eso hace que la varianza de cada componente principal sea 1. Entonces es verdadero.

Ejercicio 10.

Imagina que, como actuario, te dan un conjunto de datos en donde cada observación consiste de la edad, peso, altura e ingresos de un cierto asegurado. Determina cuáles de los siguientes eigen-vectores representa mejor la primer componente principal.

(A) (1,1,1,1).

(B) (0.5,-0.5,0.5,-0.5).

(C) (1,-1,1,-1).

(D) (0.7071,0,-0.7071,0).

(E) (0.5,0.5,0.5,0.5).

Justifica tu respuesta.

Respuesta:

Queremos encontrar la dirección en la cual la variabilidad de las variables es máxima. Viendo las variables podemos ver que la edad, el peso, la altura y los ingresos no están tan correlacionados entre si, así que el vector D parece ser la mejor respuesta, con una relación inversa entre la edad y la altura y una relación de cero entre el peso y los ingresos, que no se relacionan en lo más mínimo.