Cuestionario

Carlitos Vega, Isa Rumi, Abel Cano Days, Lucas Muecas, Nadia Flowers

May 1, 2024

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

   (a) La i-ésima componente principal se toma como la dirección que es ortogonal al $(i-1)$-esimo componente principal y maximiza la variabilidad restante.

   Respuesta: Verdadero, ya que se encuentran en la Cámara de Weyl, además que el $(i-1)$-esimo componente principal es ortogonal a todos los anteriores

   (b) Distintos componentes principales están linealmente no correlacionados

   Respuesta: Verdadero, por el Teorema de Maximización, donde especifica que los componentes están linealmente no correlacionados al ser combinación lineal unos de los otros. Además que los eigenvectores son ortogonales entre sí.

   (c) La dimensión de los datos originales es siempre mayor que la dimensión de los datos transformados

   Respuesta: Verdadero, por el Teorema de Hotteling-Fisher.

2. Suponga que se tiene la siguiente matriz de covarianza XX^T:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

   Calcula la primer componente principal

   Primer paso: Calcular los eigenvalores
   $\det(XX^T - \lambda I) = 0$

$$\det\left(\begin{bmatrix}1 & 0 & 0\\0 & 2 & 0\\0 & 0 & 2\end{bmatrix} - \lambda\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix}1-\lambda & 0 & 0\\0 & 2-\lambda & 0\\0 & 0 & 2-\lambda\end{bmatrix}\right) = 0$$

$$\lambda_1 = 2\,\text{con multiplicidad } 2$$

$$\lambda_2 = 1$$

Segundo paso: Meterlos a la cámara de Weyl

Tercer paso: Carcular eigenvectores
$\det(XX^T - 2I) * v1 = 0$

$$\left(\begin{bmatrix}-1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}\right) = 0$$

Así, el eigenvector puede ser cualquier vector en el espacio generado por el primer eje

i.e. v= (1, 0, 0)

3. Suponga que se tiene la siguiente tabla

Observación	X_1	X_2
1	-2	2
2	2	-2

- La primer componente (loading) del eigenvector que resuelve el problema de optimización del PCA para X_1 es 0.7071
- La primer componente (loading) del eigenvector que resuelve el problema de optimización del PCA para X_2 es negativa

Calcula Xb_1

$$\mathrm{b}_1 = \begin{pmatrix}0.7071\\ b_{1_2}\end{pmatrix}$$

Debemos encontrar la entrada

$$b_{1_2} \leqslant 0$$

Así, el eigenvector puede ser cualquier vector en el espacio generado por el primer eje
También sabemos que

$$b_{1_1}^2 + b_{1_2}^2 = 1$$

$$1 - b_{1_1}^2 = b_{1_2}^2$$

$$b_{1_2}^2 = 1 - 0.7071^2$$

$$b_{1_2} = \sqrt{1 - 0.7071^2}$$

$$X_{b_1} = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 0.7071 \\ -0.7071 \end{pmatrix}$$

$$X_{b_1} = \begin{pmatrix} 1 - 2(0.7071) + 2(-0.7071) \\ 2(0.7071) - 2(-0.7071) \end{pmatrix}$$

$$X_{b_1} = \begin{pmatrix} -1.8284 \\ 2.8284 \end{pmatrix}$$

4. Explica, con todo detalle, la siguiente figura:

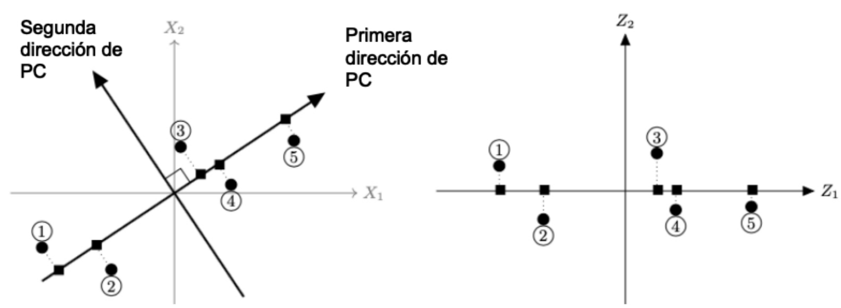

Figure 1: Ejercicio 4

Table 1: Comparación entre PCA y LDA

Característica	**PCA**	**LDA**
Objetivo	Maximizar la varianza de los datos	Maximizar la separación entre clases
Supervisión	No supervisado	Supervisado
Aplicación	Reducción de dimensionalidad	Reducción de dimensionalidad y clasificación
Sensibilidad a datos	Sensible a la escala de los datos	Menos sensible a la escala de los datos
Matriz de Covarianza	Utiliza la matriz de covarianza de los datos	Utiliza la matriz de dispersión intra-clase e inter-clase
Componentes	Vectores propios de la matriz de covarianza	Vectores propios de la matriz de dispersión inversa
Utilidad	Útil para explorar la estructura de los datos	Útil para la clasificación y maximizar la separabilidad de las clases
Robustez	Sensible a outliers	Más robusto ante outliers debido a su naturaleza supervisada
Interpretación	Las componentes no tienen una interpretación directa en términos de clases	Las componentes representan las direcciones que maximizan la separación entre clases

Respuesta: Podemos observar en la figura de la izquierda que son los ejes coordenados de las componentes principales, lo que buscamos al rotar las componentes principales es buscar la ortogonalidad y así que no haya correlación entre las componentes.

5. Sea (X, X_1, X_2) un vector gaussiano $\mathcal{N}(0, I)$ y sean $\beta \in \mathbb{R}, \sigma >= 0$,

   Defínase: $Y_1 = 0.5X + 2X_1 \wedge Y_2 = 0.5X + 2X_{2'}$

$$\text{Sea } Y = (Y_1, Y_2)^T$$

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

   (b) 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 $\mathrm{Var}(\xi_i) = \lambda_i$ y $\mathrm{Cov}(\xi_1, \xi_2) = 0$

   (c) Calcula $\tilde{p}_{i,j}$ y verifica que

$$\tilde{p}_{i,1}^2 = \tilde{p}_{i,2}^2 = 1,\ i = 1, 2$$

Respuesta: Dado que X, X_1, X_2 son v.a gaussianas con media cero y matriz de covarianza identidad W(0,1) y $Y_1 = 0.5X + 2X_1$ y $Y_2 = -0.5X + 2X_2$, podemos poner a Y como:

$$Y = \begin{pmatrix} Y_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} 0.5 & 2 \\ -0.5 & 2 \end{pmatrix}$$

Y comp Y es una combinación lineal de variables gaussianas, Y es gaussiano. Ahora para la matriz de covarianza:

$$Cov(Y) = \begin{pmatrix} 0.5 & 2 \\ -0.5 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ 2 & 2 \end{pmatrix}$$

$$Cov(Y) = \begin{pmatrix} 5.25 & 4 \\ 4 & 5.25 \end{pmatrix}$$

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:

$$\text{Para } \lambda_1 = 9.25 :$$

$$v_1 = \begin{pmatrix} 0.803 \\ -0.596 \end{pmatrix}$$

Para $\lambda_2 = 0.75$:

$$v_2 = \begin{pmatrix} 0.596 \\ 0.803 \end{pmatrix}$$

Y como los eigenvectores son ortogonales, Cov(ξ_1, ξ_2)=0

6. Tienes los siguientes datos estandarizados

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

Supongamos que el eigenvector de la primera componente principal del conjunto de datos es $(0.707, -0.5, -0.5)$. Calcula la proposición de la varianza explicada por el primer componente.

$$\mathrm{v}_1 = (0.707)^2 + (-0.5)^2 + (-0.5)^2$$

$$\mathrm{v}_1 = 0.5 + 0.25 + 0.25$$

$$\lambda_1 = 1$$

La proporción de la varianza es:

$$\tfrac{1}{1} = 1$$

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

   (a) La proposición de varianza explicada por un componente principal adicional nunca disminuye a medida que se agregan más componentes principales.

(b) La proposición acumulativa de la varianza explicada nunca decrece a medida que se agregan más componentes principales.
(c) Al usar todas las posibles componentes principales, nos provee de un mejor entendimiento de los datos.
(d) La gráfica scree provee un método para determinar el número de componentes principales a usar.

Respuestas.

(a) Falso. La proposición de varianza explicada por un componente principal adicional sí puede disminuir a medida que se agregan más componentes principales.
(b) Verdadero. La proposición acumulativa de la varianza explicada nunca decrece a medida que se agregan más componentes principales.
(c) Falso. Usar todas las posibles componentes principales no necesariamente provee de un mejor entendimiento de los datos.
(d) Verdadero. La gráfica scree sí provee un método para determinar el número de componentes principales a usar.

8. Imagina que realizas un análisis por componentes principales sobre las misma cuatro variables en un conjunto de datos particular: X_1, X_2, X_3 y X_4. El primer análisis solo 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 aserveraciones:

   - X_1 está más correlacionado con X_2 que con X_3
   - X_3 tiene la varianza más alta de las cuatro variables.
   - La observación 24 tiene un valor positivo, relativamente grande, para X_4

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

   Nótese que, 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.

   Respuesta: Visualmente, todas las aseveraciones son falsas, ya que en el PCA se hace una transformación ortogonal y en ninguna de las gráficas proporcionadas se puede observar dicha transformación ortogonal.

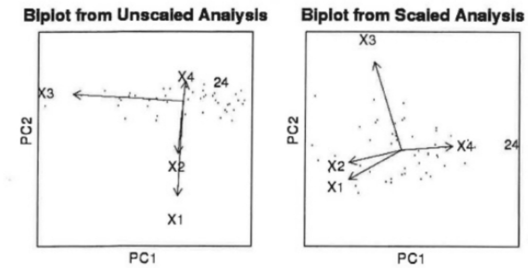

Figure 2: Ejercicio 8

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 usando dos componentes principales:

$$\xi_1 = b_{11}X_{11} + b_{21}X_{21} + b_{31}X_{31} + b_{41}X_{41}$$
$$\xi_2 = b_{12}X_{12} + b_{22}X_{22} + b_{32}X_{32} + b_{42}X_{42}$$

Tienes las siguientes aseveraciones

- $\Sigma_{i=1}^{100}(\Sigma_{j=1}^{4}b_{j1}x_{ij})^2 = \Sigma_{i=1}^{100}(\Sigma_{j=1}^{4}b_{j2}x_{ij})$
- $\Sigma_{i=1}^{4}b_{j1}b_{j2} = 0$
- $\Sigma_{i=1}^{4}b_{j1}^2 + \Sigma_{j=1}^{4}b_{j2}^2 = 1$

Determina, justificadamente, cuales aserveraciones son verdaderas

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,-0.7072,0)
(e) (0.5,0.5,0.5,0.5)

Justifica tu respuesta Para determinar cuál de los siguientes vectores representa mejor la primera componente principal, necesitamos calcular los valores propios y los vectores propios de la matriz de covarianza de los datos.

Pero primero, normalmente normalizamos los datos para que tengan media cero y desviación estándar uno antes de calcular la matriz de covarianza. Supongamos que ya hemos realizado esta normalización.

Dado un conjunto de datos X con n observaciones y p características, la matriz de covarianza S se calcula como:

$$S = \frac{1}{n} X^T X$$

Donde X^T es la matriz transpuesta de X.

Luego, calculamos los valores y vectores propios de S.

Después de calcular los valores propios y los vectores propios, el vector propio asociado con el valor propio más grande es la dirección de la primera componente principal.

Dado que no tenemos la matriz de covarianza ni los datos, podemos examinar las características de los vectores propuestos:

- **A:** (1,1,1,1) - Esta dirección sugiere que todas las características tienen el mismo peso, lo que significa que no está dando más importancia a ninguna característica en particular.
- **B:** (0.5, -0.5,0.5,-0.5) - Este vector parece estar equilibrando positiva y negativamente las características, lo que podría ser útil para identificar relaciones lineales opuestas.
- **C:** (1,-1,1,-1) - Similar al anterior, pero esta vez las características están contrapuestas con un peso absoluto de 1.
- **D:** (0.7071,0,-0.7072,0) - Este vector parece estar ponderando más las características diagonales de la matriz de covarianza.
- **E:** (0.5,0.5,0.5,0.5) - Todas las características tienen el mismo peso, similar a A.

Si la matriz de covarianza refleja una relación lineal más fuerte entre las características, esperaríamos que el vector propio esté más alineado con esas características.

Por lo tanto, el vector propio que representa mejor la primera componente principal probablemente sea **D** o **B**, ya que parecen sugerir una dirección donde las características están relacionadas de manera más significativa. Sin embargo, la elección final dependería de los datos y de cómo están distribuidas las características.