Tarea 01 - Reconocimiento de patrones
Giovanni Gamaliel López Padilla

Problema 01

Para un estudio se mide la temperatura en diferentes partes del cuerpo de una muestra de personas. Un investigador expresa todas las temperaturas en grados Celcius. Otro investigador convierte primero todas estas temperaturas a grados Fahrenheit. ¿Cómo se relacionan sus matrices de covarianza? Si ambos deciden proyectar en la dirección de máxima varianza, ¿obtendrán las mismas direcciones de proyección? Explica tu respuesta de manera formal.

En la ecuación se muesra la transformación de grados Celcius a Fahrenheit.

$$F = \frac{5}{9}C - \frac{160}{9} \tag{1}$$

La cual se puede reducir a una transformación lineal de la forma:

$$Y = aX + b$$

donde

- a es un número real igual a $\frac{5}{9}$.
- b es un número real igual a $\frac{-160}{9}$.
- X es un vector con los grados en Celcius.
- Y es un vector con los grados en Fahrenheit.

Entonces, se tiene que:

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov(aX_i + b, aX_j + b) \\ &= Cov(aX_i, aX_j) \\ &= a^2 Cov(X_i, X_j) \end{aligned}$$

Por lo tanto, la relación entre las matrices de covarianzas es proporcional a $\frac{25}{81}$. Al ser proporcionales las direcciones de su máxima varianza serán la misma. El cambio se encontrará en los valores propios encontrados.

Problema 2

Supongamos que $\mathbf{X}=(\mathbf{X_1}, \mathbf{X_2})$, $\mathbf{Var(X_1)}=\mathbf{Var(X_2)}=\mathbf{1}$.

a) Supongamos que X_1 y X_2 son v.a. independientes con promedio 0. Verifica que cualquier dirección l da máxima varianza en las proyecciones.

Como X_1 y X_2 son v.a independientes entonces, la matriz de covarianza Cov(X) es:

$$Cov(X) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

por lo que se obtiene que $Cov(X) = \mathbb{I}$. Entonces, se obtiene que:

$$\begin{aligned} \max_{||l||} \frac{l^t Cov(X) l}{l^t l} &= \max_{||l||} \frac{l^t \mathbb{I} l}{l^t l} \\ &= \max_{||l||} \frac{l^t l}{l^t l} \\ &= \max_{||l||} 1 \\ &= 1 \end{aligned}$$

por lo tanto, se maximiza la varianza para cualquier dirección de l en las proyecciones.

b) Supongamos que X_1 y X_2 son v.a. dependientes. Calcula la primer componente principal a mano. ¿Qué particularidad tiene?

Suponiendo de la covarianza de X_1 Y X_2 es a, entonces, la matriz de covarianza es:

$$Cov(X) = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$$

Calculando la primer componente l, se obtiene que los valores propios de $Cov(X)$ es:

$$\begin{aligned} |Cov(X) - \lambda \mathbb{I}| &= 0 \\ \begin{vmatrix} 1-\lambda & a \\ a & 1-\lambda \end{vmatrix} &= 0 \\ (1-\lambda)^2 - a^2 &= 0 \\ (1-\lambda-a)(1-\lambda+a) &= 0 \\ \lambda_1 &= 1-a \\ \lambda_2 &= 1+a \end{aligned}$$

Suponiendo que $a > 0$, entonces λ_2 es el eigenvalor mayor. Calculando los vectores propios relacionados a λ_2, se obtiene que:

$$\begin{pmatrix} -a & a \\ a & -a \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} -1 & 1 \\ a & -a \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$-x_1 + x_2 = 0$$
$$x_1 = x_2$$

por lo tanto, el vector propio asociado a λ_2 es $v_2 = [x_1, x_1]^T$. La particularidad que tiene es que las componetes no tienen un valor determinado por lo que es necesario elegir el parámetro x_1 y en seguida normalizar el vector.

Problema 3

Haz unos pequeños cambios necesarios para demostrar que el segundo vector propio de Cov(X) es la solución del problema de maximizar el cociente bajo la restricción adicional de ser ortogonal primer vector propio.

Al final del video se obtiene que una solución del problema descrito en la ecuación 2 es el primer vector propio de Cov(X).

$$\max_{||l||} \frac{l^t Cov(X) l}{l^t l} \tag{2}$$

Realizando un cambio de base a la ecuación 2, se obtiene la ecuación 3.

$$\max_{||y||} \frac{y^t \Lambda y}{y^t y} \rightarrow \max_{||y||} \frac{\sum_i \mu_i y_i^2}{\sum_i y_i^2} \tag{3}$$

Tomando en cuenta que $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_i$, donde $y_1 = (1, 0, \ldots, 0)$. Entonces se propone que $y_2 = (0, 1, 0, \ldots, 0)$, esto con el proposito que y_1 y y_2 sean ortogonales. Usando y_2 en la ecuación 3, se obtiene que una solución es μ_2, el cual es el segundo eigenvalor de la matriz Cov(X). Devolviendo a la base original a y_2 se obtiene que:

$$y_2 = U^t l_2$$
$$U y_2 = U U^t l_2$$
$$U y_2 = l_2$$
$$u_2 = l_2$$

donde u_2 es el segundo eigenvector de la matriz Cov(X) el cual es ortogonal a u_1.

Problema 4

Considera los datos oef2.data. Se trata de los promedios mensuales de la temperatura (en Celsius) en 35 estaciones canadienses de monitoreo. El interés es comparar las estaciones entre sí en base de sus curvas de temperatura. Considerando las 12 mediciones por estación como un vector X, aplica un análisis de componentes principales. Como X representa (un muestreo de) una curva, este tipo de datos se llama datos funcionales.Interpreta y dibuja (como curva) los primeros dos componentes, p_1, p_2, es decir grafica $\{(i, p_{1i})\}$ y $\{(i, p_{2i})\}$. Agrupa e interpreta las estacionesen el biplot (ten en mente un mapa de Canada).

En la figura 1 se muestra las gráficas lineales de los valores de cada componente. Se observa que el comportamiento es semejante en las dos componentes. La primer componente conserva una menor variación en sus valores a comparación de la segunda componente.

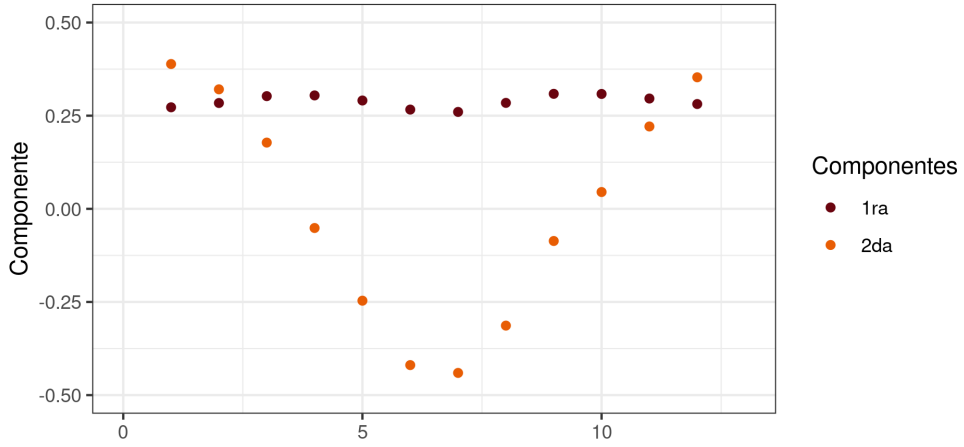
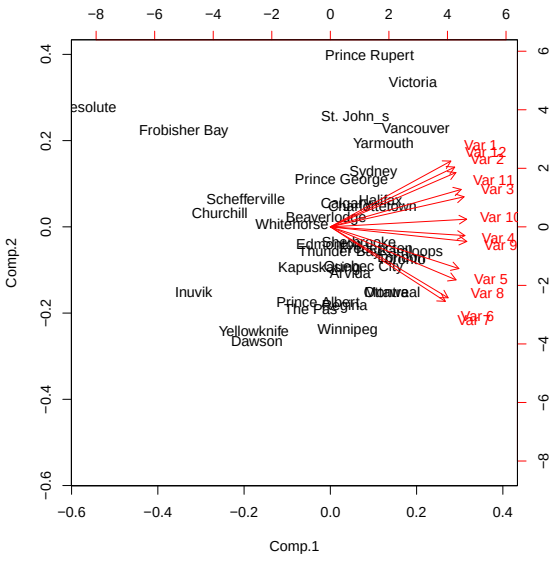


Figura 1: Valores de la primer y segunda componente obtenida con PCA.



(a) Biplot de la primer y segunda componente obtenido de los datos del archivo oef2.data.

(b) Mapa de Canada. Obtenido de Geology.com

Figura 2

En la figura 2a se muestra la gráfica de biplots obtenida al aplicar PCA a los datos del archivo oef2.data. Se observa como la distribución de valores se concentra para las estaciones que tienen una latitud semejante (figura 2b). En cambio para la estación ubicada en Resolute se encuentra lejana de las demás. Esto puede dar la interpretación que mientras más se encuentren en la izquierda de la gráfica las estaciones reportaran una temperatura menor. En caso contrario, estas reportaran una temperatura mayor.

Problema 5

En este ejercicio usamos resultados del heptatlón feminina de los pasados juegos olímpicos de Tokyo (2021). En el archivo heptatlonTokyose pueden consultar los tiempos/distancias y el puntaje final (score) de 20 atletas.

a) Describe de manera general los datos sin considerar la columna con los puntajes finales, usando visualizaciones ilustrativas. Toma en cuenta que son pocas observaciones. Asi, no será posible llegar a conclusiones fuertes.

Los datos obtenidos de los resultados del heptatlón femenino de los juegos olímpicos del 2021 contienen la siguiente información:

1. **Salto de altura (high jump)**
2. **Lanzamiento de peso (shot put)**
3. **Salto de longitud (long jump)**
4. **Lanzamiento de javalina (javelin)**
5. **100 metros planos (100m)**
6. **200 metros planos (200m)**
7. **800 metros planos (800m)**

Las competencias de los puntos 1 al 4 el objetivo es obtener la mayor distancia posible. En cambio las competencias de los puntos 5 al 8 el objetivo es obtener el menor tiempo posible.

Las distancias y tiempos obtenidos por los atletas fueron normalizadas con la ecuación 4.

$$m_n = \frac{m - \mu}{\sigma} \tag{4}$$

donde

- m es la distancia o tiempo obtenido en la prueba del atleta.
- μ es la media de los atletas en cierta prueba.
- σ es la desviación estandar de los atletas en cierta prueba.
- m_n es la medida de la distancia o tiempo del participante estandarizado.

En la figura 3 se visualizan las medidas m_n de cada prueba. En la parte superior se encuentran las pruebas donde es mejor obtener un puntaje alto y en la parte inferior las pruebas donde es mejor obtener un puntaje bajo.

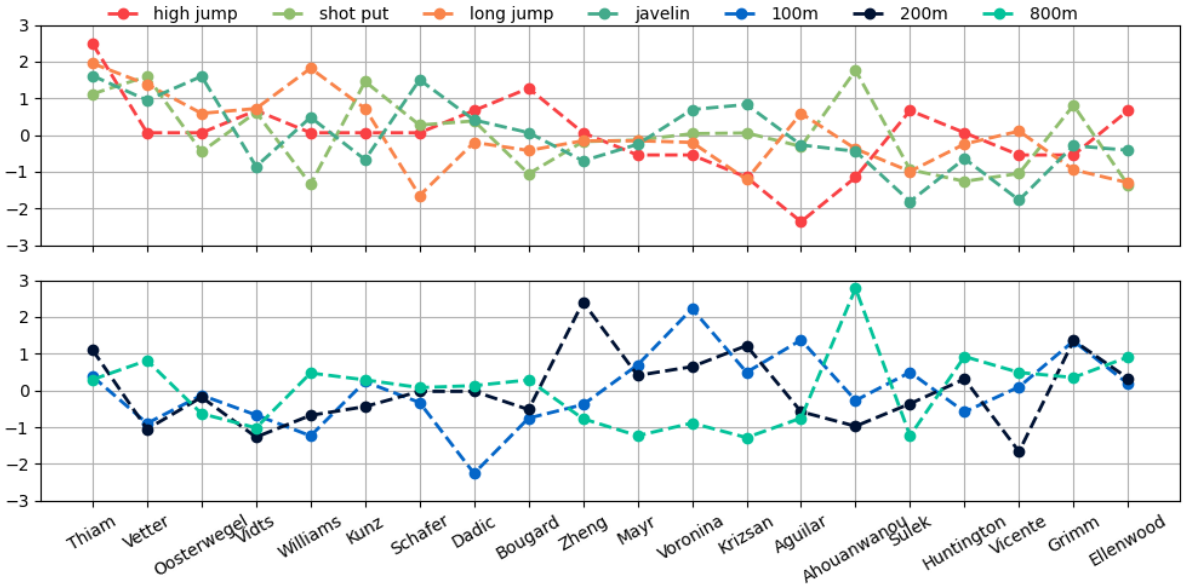


Figura 3: Puntajes estandarizados de los atletas obtenidos en los olímpicos 2021.

Con la figura 3 se puede visualizar que atleta es mejor en que prueba y el rendimiento general de cada uno.

b) Un problema en un heptatlón es cómo convertir los resultados obtenidos en las diferentes pruebas en un puntaje final. Explora la utilidad de PCA usando la proyección de los resultados de las pruebas sobre el primer componente como una alternativa al puntaje final. ¿Cómo se relaciona con el puntaje final oficial? Información sobre cómo se calcula actualmente el puntaje: Puntaje heptatlon.

La relación entre los puntajes oficiales y los resultados de la primer componente de PCA resulta ser semejante. Algunos puestos logran conservar su posición del tablero. En cambio existen atletas que reducen o aumentan su posición con respecto a la posición en el puntaje oficial. Esto es debido los problemas presentados en el link. Algunas pruebas se ven como si tuvieran una mayor importancia y la diferencia entre el primer puesto y ultimo no siempre es la misma. En la figura 4 se ve representado los resultados de aplicar PCA a las mediciones de cada atleta en las pruebas tomadas.

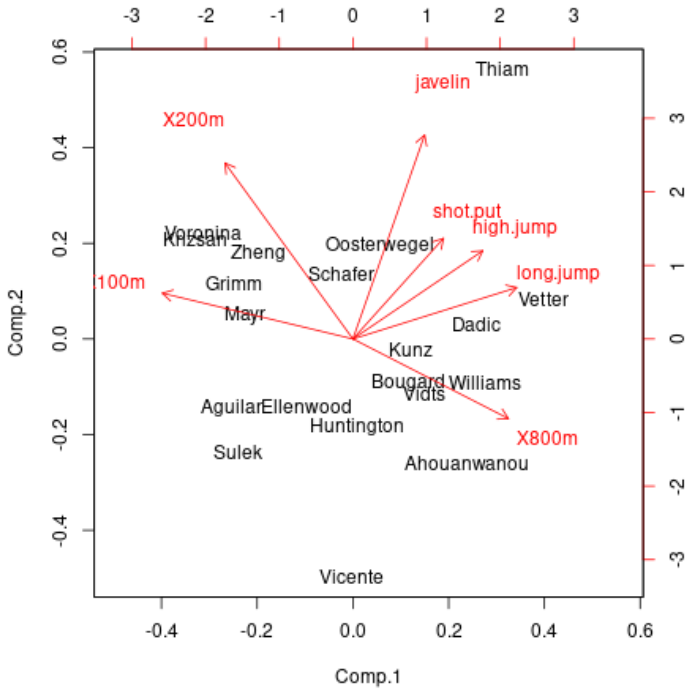


Figura 4: Resultados de PCA a las pruebas competidas en los olimpicos 2021.