TAREA 2

María Pallares Diez

2025-01-08

Introducción

En esta tarea, realizaremos un Análisis de Componentes Principales (ACP) sobre dos bancos de datos: `datos` y `datos_b`. Los datos están contenidos en los archivos `datosp1.RData` y `datosp1_b.RData` respectivamente. Se calcularán las desviaciones típicas de las variables cuantitativas, se determinará el método adecuado para el ACP y se analizarán los resultados.

Carga de datos

```
# Cargar los datos
load("datosp1.RData")
load("datosp1_b.RData")
```

a. Desviaciones típicas y elección del método

Desviaciones típicas

```
# Filtrar solo columnas numéricas para evitar errores con factores
datos_numericos <- datos[sapply(datos, is.numeric)]
datos_b_numericos <- datos_b[sapply(datos_b, is.numeric)]

# Calcular desviaciones típicas para las variables cuantitativas de ambos bancos
sapply(datos_numericos, sd, na.rm = TRUE)
```

```
##                   ID                 Edad        Horas_estudio
##           29.0114920            0.8382979            4.7687462
## Promedio_matematicas    Promedio_ciencias     Promedio_lectura
##           10.2286389            8.8425306           12.2432730
##           Asistencia          Horas_sueño         Nivel_estres
##           11.2869253            1.0321552            1.9232758
##     Uso_dispositivos     Condicion_fisica               Centro
##            1.8609792           47.7802480            2.4452799
```

```
sapply(datos_b_numericos, sd, na.rm = TRUE)
```

```
##                   ID                 Edad        Horas_estudio
##           29.0114920            0.8382979            5.9134713
## Promedio_matematicas    Promedio_ciencias     Promedio_lectura
##           10.2286389           11.2538682           14.0563689
##           Asistencia          Horas_sueño         Nivel_estres
##           14.6701543            4.4103372            1.9232758
##     Uso_dispositivos     Condicion_fisica
##            1.8846981           86.9240511
```

Justificación del método

A partir de las desviaciones típicas, si las variables tienen escalas muy diferentes, se usará la matriz de correlaciones para estandarizar las variables antes del análisis. De lo contrario, se usará la matriz de varianzas-covarianzas.

b. Análisis de Componentes Principales

ACP para `datos`

```r
# Realizar ACP
acp_datos <- prcomp(datos, scale. = TRUE)
summary(acp_datos)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.4525 1.3021 1.1834 1.0979 1.05215 0.98765 0.94652
## Proportion of Variance 0.1758 0.1413 0.1167 0.1004 0.09225 0.08129 0.07466
## Cumulative Proportion  0.1758 0.3171 0.4338 0.5342 0.62650 0.70778 0.78244
##                           PC8     PC9    PC10   PC11    PC12
## Standard deviation     0.85329 0.81256 0.77478 0.7275 0.30468
## Proportion of Variance 0.06068 0.05502 0.05002 0.0441 0.00774
## Cumulative Proportion  0.84312 0.89814 0.94816 0.9923 1.00000
```

Pregunta 1: Porcentaje de varianza explicado La varianza explicada por la primera componente es:

```r
sum(acp_datos$sdev^2 / sum(acp_datos$sdev^2))
```

```
## [1] 1
```

ACP para `datos_b`

```r
# Realizar ACP
acp_datos_b <- prcomp(datos_b_numericos, scale. = TRUE)
summary(acp_datos_b)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
```

```
## Standard deviation     2.4324 1.3345 1.1036 0.95987 0.66354 0.51577 0.4350
## Proportion of Variance 0.5379 0.1619 0.1107 0.08376 0.04003 0.02418 0.0172
## Cumulative Proportion  0.5379 0.6998 0.8105 0.89428 0.93430 0.95849 0.9757
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.3181 0.29445 0.24833 0.13375
## Proportion of Variance 0.0092 0.00788 0.00561 0.00163
## Cumulative Proportion  0.9849 0.99277 0.99837 1.00000
```

```
cumsum(acp_datos_b$sdev^2) / sum(acp_datos_b$sdev^2)
```

Pregunta 2: Número de componentes para el 90% de varianza

```
##  [1] 0.5378821 0.6997907 0.8105167 0.8942766 0.9343028 0.9584867 0.9756877
##  [8] 0.9848856 0.9927677 0.9983738 1.0000000
```

Pregunta 3: Diferencias entre resultados Comparemos los porcentajes de varianza explicados por las primeras componentes entre ambos conjuntos de datos.

Interpretación de las componentes principales para `datos_b`

```
# Cargar coeficientes de las primeras componentes
acp_datos_b$rotation[, 1:2]
```

```
##                              PC1         PC2
## ID                    0.395364841 -0.05210119
## Edad                  0.005112111  0.20708039
## Horas_estudio         0.320546792  0.01956427
## Promedio_matematicas  0.393758942 -0.06443541
## Promedio_ciencias     0.386279064 -0.06794983
## Promedio_lectura      0.321949472 -0.04687938
## Asistencia            0.306390998 -0.04265555
## Horas_sueño           0.290299107  0.11345487
## Nivel_estres          0.057731577  0.68330863
## Uso_dispositivos     -0.379280034  0.10059616
## Condicion_fisica      0.087555041  0.67181237
```

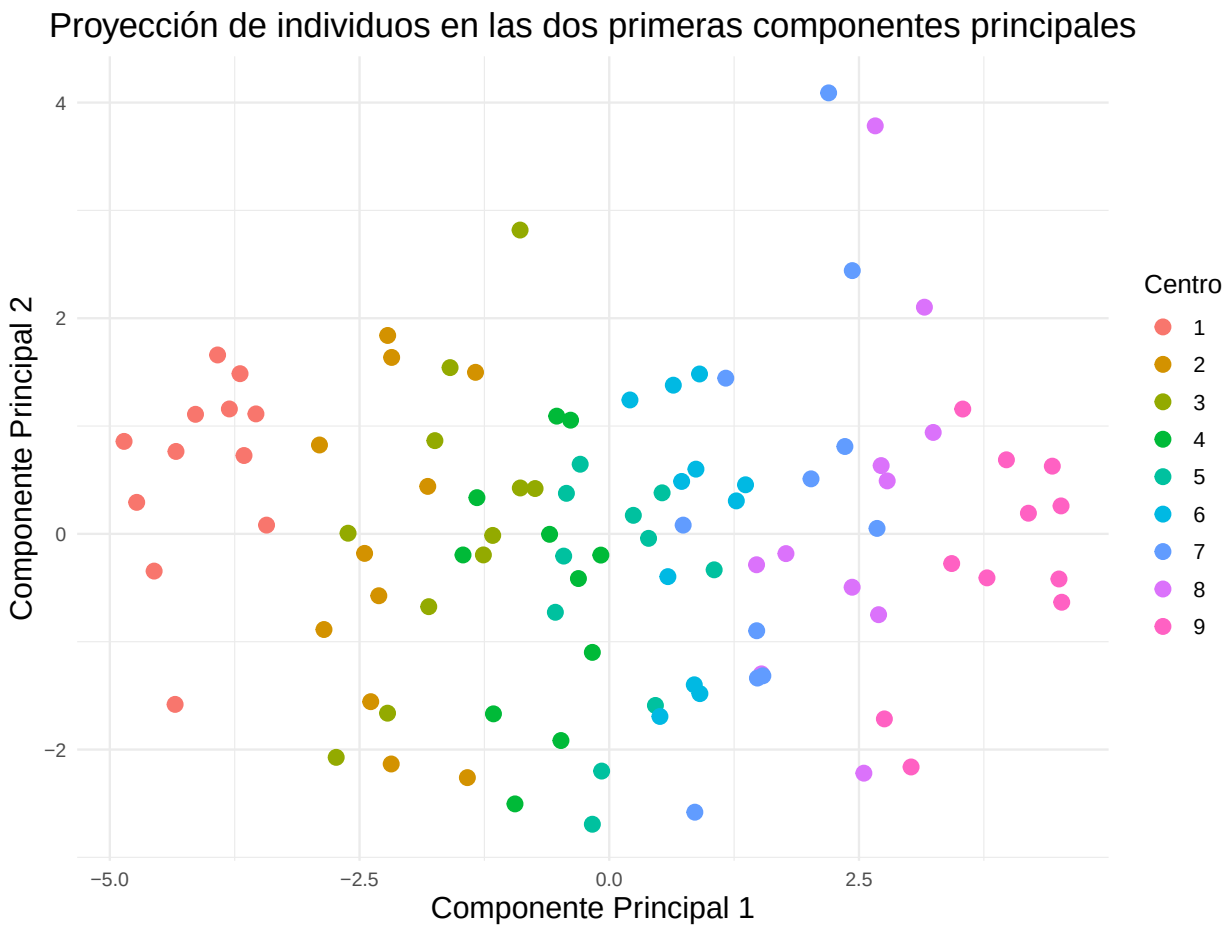
```
library(ggplot2)

# Crear un data frame con las primeras dos componentes
individuos <- data.frame(acp_datos_b$x[, 1:2])
colnames(individuos) <- c("Componente1", "Componente2")
individuos$Centro <- as.factor(datos_b$Centro)  # Suponiendo que existe esta columna

# Gráfico
ggplot(individuos, aes(x = Componente1, y = Componente2, color = Centro)) +
```

```r
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Proyección de individuos en las dos primeras componentes principales",
       x = "Componente Principal 1",
       y = "Componente Principal 2") +
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10))
```

Pregunta 5: Gráfico de individuos



Conclusiones

En esta sección se discutirán las diferencias observadas entre los conjuntos de datos, la importancia de las componentes principales y la interpretación de los resultados obtenidos.