Tarea 1 (B) Reconocimiento de Patrones

1. (no hay que entregar nada) A aquellos que se sienten aun no muy familiarizados con análisis de datos, recomiendo leer la parte del libro `Applied Multivariate Analysis` sobre un análisis de PCA de los datos de (otro) heptatlon a partir de pag. 78 (pag. 92 en el pdf).
   Nota: en el libro se usa `prcomp` y no `princomp`. Ambos calculan PCA; la diferencia es más bien en el método númerico subyacente que se usa: `prcomp()` usa SVD y `princomp()` usa la matriz de covarianza. En general se considera que desde punto numérico `prcomp()` es mejor pero es más dificil sacar las proyecciones y scores. Si `objeto<- prcomp()`, entonces `objeto$rotation[,1]` es el equivalente a lo que da loadings[,1] con `princomp()` (en Python por default se va por SVD).

2. (usaremos este resultado en la siguiente clase)
   Sea $\{x_i\}$ un conjunto de n vectores d dimensional. Definimos las matrices $n \times n$ $[\mathbb{K}_{i,j}]$ con $\mathbb{K}_{i,j} = \langle x_i, x_j \rangle$ y $\mathbb{D}^2$ la matriz de distancias al cuadrada correspondiente: $\mathbb{D}^2_{i,j} = ||x_i - x_j||^2$

   Verifica la identidad:

   $$\mathbb{D}^2 = c1^t + 1c^t - 2\mathbb{X}\mathbb{X}^t,$$

   con 1 un vector de unos de longitud n y c el vector de longitud n con elementos $(\mathbb{K}_{i,i})_{i=1}^n$

   Hint: escribe primero $||x_i - x_j||^2$ en términos de productos puntos.

3. Acerca de la demostración de la maximización del cociente de Rayleigh ($\text{máx}_l \frac{l^t Cov(X) l}{l^t l}$): 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 al primer vector propio. (por si sirve: aquí una grabación vieja de la demostración en tiempos de la pandemia `https://www.youtube.com/watch?v=8TBpSUXcDww`)

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. Consideramos las 12 mediciones por estación como las entradas de un vector (v.a.) X

a) Busca algunas visualizaciones informativos de los datos.

b) Calcula los componentes principales.

c) Aprovechando que las columnas hacen referencia a meses consecutivos, tiene sentido dibujar cada componente principal como una gráfica. Por ejemplo graficar $\{(i, l_{1i})\}$ y $\{(i, l_{2i})\}$ con i de 1 a 12 y l_1 y l_2 el primer y segundo componente principal.

¿ Qué interpretación das al primer y segundo componente?

Para leer los datos en R

```
temp <- matrix(scan("oef2.data"), 35, 12, byrow=T)

nombresestaciones <-   c("St. John_s",    "Charlottetown", "Halifax" ,
                  "Sydney",        "Yarmouth",      "Fredericton",
                  "Arvida",        "Montreal",      "Quebec City",
                  "Schefferville", "Sherbrooke",    "Kapuskasing",
                  "London",        "Ottawa",        "Thunder Bay",
                  "Toronto",       "Churchill",     "The Pas",
                  "Winnipeg",      "Prince Albert", "Regina",
                  "Beaverlodge",   "Calgary",       "Edmonton",
                  "Kamloops",      "Prince George", "Prince Rupert",
                  "Vancouver",     "Victoria",      "Dawson",
                  "Whitehorse",    "Frobisher Bay", "Inuvik",
                  "Resolute",      "Yellowknife")

rownames(temp)<-nombresestaciones
```