Tarea 1 Reconocimiento de Patrones.

Fecha entrega: lunes 31 de enero, 10PM

A. Cultura general (no entregar nada)

1. En caso de no conocerlo, recomiendo visitar el sitio de Gapminder como un ejemplo de cómo visualizar datos de alta dimensión: `https://www.gapminder.org/` y ver los videos y las herramientas en:
   `https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1`

2. Echa un ojo al articulo donde K. Pearson introdujo por primera vez PCA en 1901:
   `http://pca.narod.ru/pearson1901.pdf`

3. Como preparación a la parte de visualización de datos, echa un ojo a la conferencia de Andy Kirk de la semana pasada:
   `https://www.youtube.com/watch?v=QuWM2Chr_DA`

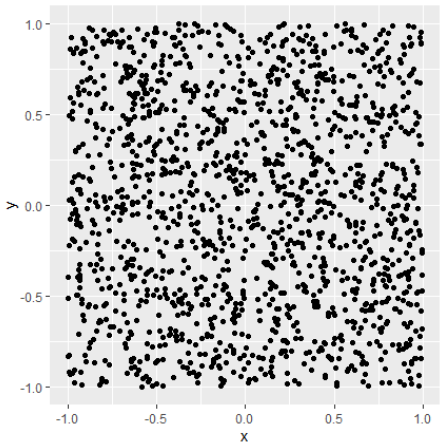
B. Preguntas cortas

1. 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.

2. Supongamos que $X = (X_1, X_2), Var(X_1) = Var(X_2) = 1$.

   a) Supongamos que X_1 y X_2 son v.a. independientes con promedio 0. Por ejemplo, una muestra podria ser:



Verifica que cualquier dirección l da máxima varianza en las proyecciones.

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

3. Revisa el video sobre la maximización del cociente de Rayleigh:
`https://youtu.be/8TBpSUXcDww`
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 ortogonalal primer vector propio.

4. (no entregar) Sea X una v.a. multidimensional con matriz de covarianza $Cov(X)$. Si l_i es el i-esimo vector propio de $Cov(X)$ y $Y_i = \langle l_i, X \rangle$ muestra que:

$$Cov(Y_i, Y_j) = 0, i \neq j$$

C. Análisis de datos (en equipos de dos)

1. 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, $p1, p2$ es decir grafica $\{(i, p1_i)\}$ y $\{(i, p2_i)\}$. Agrupa e interpreta las estaciones en el biplot (ten en mente un mapa de Canada).

Para leer los datos:

```
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
```

2. En este ejercicio usamos resultados del heptatlón feminina de los pasados juegos olímpicos de Tokyo (2021). En el archivo *heptatlonTokyo* se 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.

   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:
http://theaftermatter.blogspot.mx/2012/06/maths-of-heptathlon-why-scoring-system.html