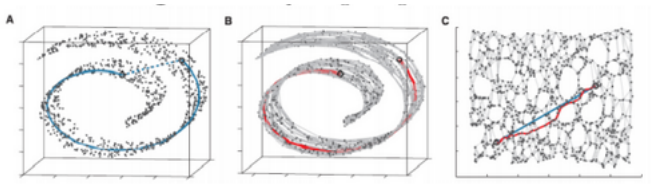
Reconocimiento Estadístico de Patrones **Nota: 10/10**
Tarea 3

Randy Osbaldo Ibarra Cayo

B. Preguntas Cortas

Pregunta 1 **Nota: 1/1**

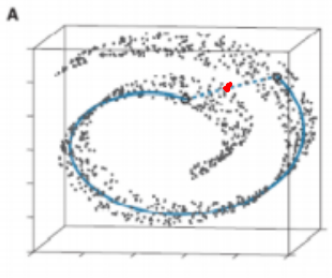
Vimos el siguiente ejemplo para ilustrar ISOMAP



Decidimos que dos observaciones x_i, x_j son conectados por una arista en el grafo correspondiente si y sólo si x_i está entre los k-vecinos más cercanos de x_j, o bien x_j está entre los k-vecinos más cercanos de x_i. Muestra que la adición de una sola observación en este ejemplo puede destruir por completo el desenrollamiento. Márcala en el dibujo y explícalo.

Solución

Notemos que al agregar una observación en el esacio que hay entre dos partes de "la sábana enrollada", esta nueva observación tendrá cerca observaciones que pertenecen a partes diferentes de esta "sábana" por lo que al aplicar ISOMAP el algoritmo intentará desenrollar la misma sábana, pero ahora con dos "pedazos pegados".



Ok

Figure 1: Posible observación que destruiría desenrrollamiento

Por ejemplo, en la figura 1 podemos agregar una observación a la mitad del la recta punteada azul que une a las dos observaciones marcadas. Suponiendo que dos de los vecinos más cercanos a esta nueva observación son estas dos observaciones marcadas, entonces el desenrrollamiento se destruye, pues ahora están conectadas dos regiones que imposibilitan el desenrrollamiento, es decir, con esta nueva observación el grafo que se construye parece no ser "aplanable".

Pregunta 2 **Nota: 1/1**

Vimos el Teorema de Rao en clase: Si $\mathbb{F}$ es una matriz simétrica de rango d y con SVD:

$$\mathbb{F} = \sum_{i}^{d} \lambda_i v_i v_i^T$$

La matriz simétrica de rango $p < d$ que minimiza $||\mathbb{F} - \mathbb{G}||_F$ es:

$$\mathbb{G} = \sum_{i}^{p} \lambda_i v_i v_i^T$$

Muestra que para esta elección, el error $||\mathbb{F} - \mathbb{G}||_F^2$ es igual a $\sum_{i=p+1}^{d} \lambda_i^2$. Hint: usa las propiedades de v_i y recuerda que $||A||_F^2 = \text{Traza}(A^T A)$.

Solución

Si $\mathbb{F} = \sum_i^d \lambda_i v_i v_i^T$ es la descomposición SVD de $\mathbb{F}$ y $\mathbb{G} = \sum_i^p \lambda_i v_i v_i^T$ con $p < d$, entonces

$$\begin{aligned} \mathbb{F} - \mathbb{G} &= \sum_{i=1}^{d} \lambda_i v_i v_i^T - \sum_{i=1}^{p} \lambda_i v_i v_i^T \\ &= \sum_{i=p+1}^{d} \lambda_i v_i v_i^T \qquad \text{Ok} \end{aligned}$$

Luego, dado que $||A||_F^2 = \text{Traza}(A^T A)$ se tiene que

$$\| \sum_{i=p+1}^{d} \lambda_i v_i v_i^T \|_F^2 = \text{Traza} \left[(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)^T \right] \qquad \text{Ok}$$

y notemos que para cada $i = 1, \ldots, d$, la matriz $\lambda_i v_i v_i^T$ es simétrica, entonces $\sum_{i=p+1}^{d} \lambda_i v_i v_i^T$ es simétrica, es decir

$$(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)^T = \sum_{i=p+1}^{d} \lambda_i v_i v_i^T$$

entonces

$$\begin{aligned} (\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)^T &= (\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T) \\ &= (\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)^2 \qquad \text{Ok} \end{aligned}$$

observemos que

$$(\sum_{i=p+1}^{d} \lambda_i v_i v_i^T)^2 = \sum_{p+1 \leq i,j \leq d} (\lambda_i v_i v_i^T)(\lambda_j v_j v_j^T)$$

es decir, la suma del producto de cada pareja de terminos en la suma original. Ahora bien, notemos que v_i y v_j son ortogonales para cada pareja $1 \leq i \neq j \leq d$ (Pues forman parte de la descomposicion SVD de $\mathbb{F}$), entonces $v_i^T v_j = 0$ para $i \neq j$, es decir, si $i \neq j$, entonces

$$(\lambda_i v_i v_i^T)(\lambda_j v_j v_j^T) = \lambda_i \lambda_j v_i (v_i^T v_j) v_j^T = 0 \qquad \text{Ok}$$

por lo anterior

$$\sum_{p+1\leq i,j\leq d}(\lambda_i v_i v_i^T)(\lambda_j v_j v_j^T) = \sum_{p+1\leq i\neq j\leq d}(\lambda_i v_i v_i^T)(\lambda_j v_j v_j^T) + \sum_{i=p+1}^{r}(\lambda_i v_i v_i^T)^2$$
$$= \sum_{i=p+1}^{r}(\lambda_i v_i v_i^T)^2$$
$$= \sum_{i=p+1}^{r}\lambda_i^2(v_i v_i^T)^2 \qquad \text{Ok}$$

Por otro lado notemos que cada v_i es unitario, es decir $v_i^T v_i = \|v_i\| = 1$, entonces

$$(v_i v_i^T)^2 = v_i v_i^T v_i v_i^T = v_i v_i^T$$

por lo que

$$\sum_{i=p+1}^{d}\lambda_i^2(v_i v_i^T)^2 = \sum_{i=p+1}^{d}\lambda_i^2 v_i v_i^T$$

Entonces se tiene la siguiente igualdad

$$(\sum_{i=p+1}^{d}\lambda_i v_i v_i^T)(\sum_{i=p+1}^{d}\lambda_i v_i v_i^T)^T = \sum_{i=p+1}^{d}\lambda_i^2 v_i v_i^T$$

por lo que

$$\text{Traza}\left[(\sum_{i=p+1}^{d}\lambda_i v_i v_i^T)(\sum_{i=p+1}^{d}\lambda_i v_i v_i^T)^T\right] = \text{Traza}\left[\sum_{i=p+1}^{d}\lambda_i^2 v_i v_i^T\right]$$
$$= \sum_{i=p+1}^{d}\text{Traza}\left[\lambda_i^2 v_i v_i^T\right]$$
$$= \sum_{i=p+1}^{d}\lambda_i^2\text{Traza}\left[v_i v_i^T\right] \qquad \text{Ok}$$

Por último, recordemos que $\text{Traza}\left[v_i v_i^T\right] = \|v_i\| = 1$, entonces

$$\sum_{i=p+1}^{d}\lambda_i^2\text{Traza}\left[v_i v_i^T\right] = \sum_{i=p+1}^{d}\lambda_i^2$$

Se concluye que

$$\|\mathbb{F}-\mathbb{G}\| = \sum_{i=p+1}^{d}\lambda_i^2 \qquad \text{Bien!}$$

Pregunta 3 **Nota: 1/1**

Usamos para T-SNE la distancia de Kullback-Leibler. Para distribuciones discretas su definición es

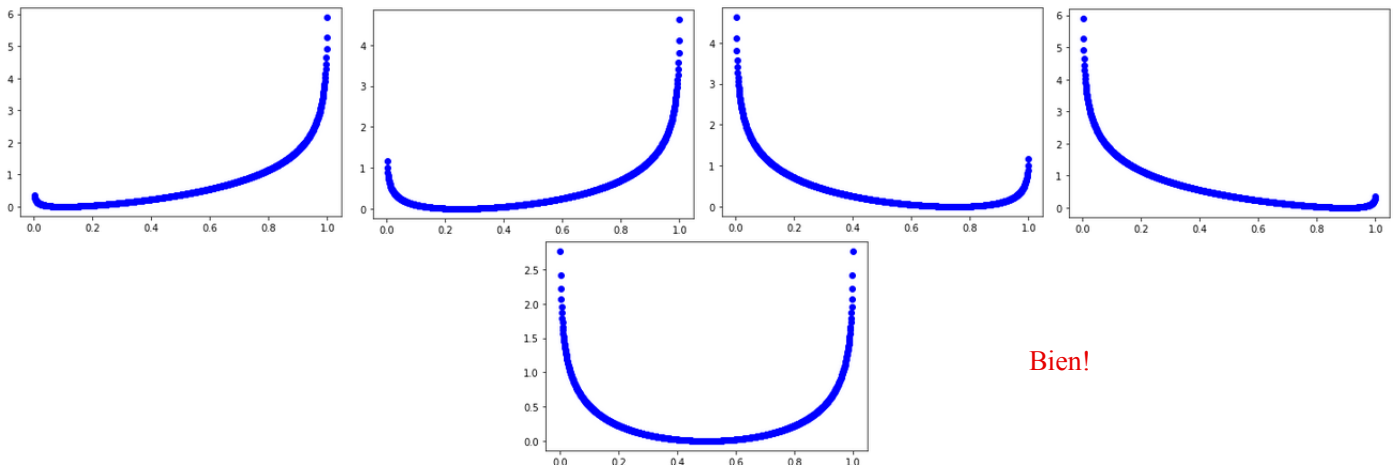
$$d(P^{(1)}, P^{(2)}) = \sum_i P_i^{(1)} \log \frac{P_i^{(1)}}{P_i^{(2)}}$$

- Calcula $d(P^{(1)}, P^{(2)})$ si $P^{(1)} \sim \text{Bern}(\theta_1)$ y $P^{(2)} \sim Bern(\theta_2)$.
  Solución: Si $P^{(1)} \sim \text{Bern}(\theta_1)$ y $P^{(2)} \sim Bern(\theta_2)$, entonces

$$\begin{aligned} d(P^{(1)}, P^{(2)}) &= \sum_{i=0}^{1} P_i^{(1)} \log \frac{P_i^{(1)}}{P_i^{(2)}} \\ &= P_0^{(1)} \log \frac{P_0^{(1)}}{P_0^{(2)}} + P_1^{(1)} \log \frac{P_1^{(1)}}{P_1^{(2)}} \\ &= (1-\theta_1) \log \frac{(1-\theta_1)}{(1-\theta_2)} + \theta_1 \log \frac{\theta_1}{\theta_2} \end{aligned}$$

Ok

- Para θ_1 fija, grafica $d(P^{(1)}, P^{(2)})$ como función de θ_2 y verifica que efectivamente mide de alguna manera la disimilitud entre $P^{(1)}$ y $P^{(2)}$.
  Solución:



Bien!

Figure 2: Graficas para diferentes valores fijos de θ_1 (0.1, 0.25, 0.75, 0.9 y 0.5)

Podemos observar que para valores θ_2 alejados de θ_1 la función se va a infinito, esto lo podemos apreciar en ambos extremos. Por otro lado, para valores θ_2 alejados de θ_1 la función se minimiza, por lo que de alguna manera mide la disimilitud entre $P^{(1)}$ y $P^{(2)}$.

Pregunta 4 **Nota: 1/1**

Sea S un conjunto finito. Definimos como medida de similitud entre dos subconjuntos A y B de S como

$$K(A, B) = |A \cap B|$$

donde $|X|$ es la cardinalidad del conjunto X. Encuentra una función Φ tal que

$$K(A, B) = \langle \Phi(A), \Phi(B) \rangle$$

Solución

Dado que S es finito, podemos suponer $|S| = n$ y entonces podemos escribir a S como

$$S = \{s_i\}_{i=1}^{n}$$

Sea $\mathbb{P}(S)$ el conjunto potencia de S, es decir, la familia de todos los subconjuntos de S y definamos $\Phi : \mathbb{P}(S) \to \mathbb{R}^n$ como

$$\Phi(X) = [\mathbb{I}_X(s_1), \mathbb{I}_X(s_2), \ldots, \mathbb{I}_X(s_n)]$$

donde $\mathbb{I}_X$ es la función indicadora con $\mathbb{I}_X(s) = 1$ si $s \in X$, y $\mathbb{I}_X(s) = 0$ si $s \notin X$. Entonces $\Phi(X)$ es un vector binario, donde la posición i contiene 1 si el elemento s_i está en el subconjunto X de S, y 0 en caso contrario. Observemos que $\Phi(X)$ tiene tantos 1's como la cantidad de elementos en X, es decir

$$\sum_{i=1}^{n} \mathbb{I}_X(s_i) = |X| \qquad \text{Ok}$$

por lo tanto

$$\begin{aligned} \langle \Phi(A), \Phi(B) \rangle &= \sum_{i=1}^{n} \mathbb{I}_A(s_i)\mathbb{I}_B(s_i) \\ &= \sum_{i=1}^{n} \mathbb{I}_{A \cap B}(s_i) \\ &= |A \cap B| \qquad \text{Ok} \end{aligned}$$

Se concluye entonces que

$$K(A, B) = \langle \Phi(A), \Phi(B) \rangle$$

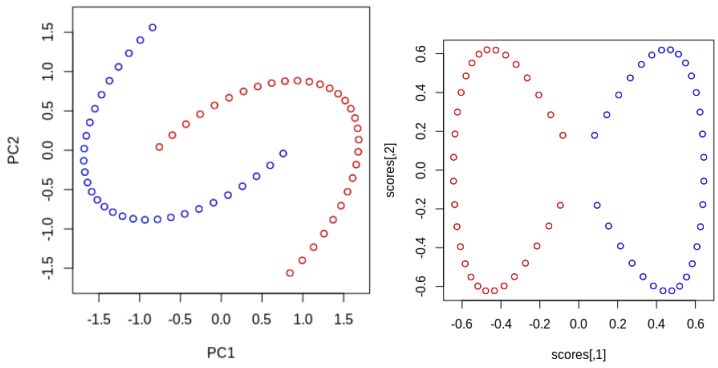
B. Análisis de Datos

Problema 1 **Nota: 3/3**

- Escribe una función en R que implementa KPCA con un kernel centrado de base radial con parámetro σ, $\mathbb{K}_c = \mathbb{C}\mathbb{K}\mathbb{C}$ con $\mathbb{K}_{i,j} = \exp(-\|x_i - x_j\|^2/\sigma)$ y $\mathbb{C}$ la matriz para centrar. Se debe programar todo desde cero apoyándose solamente en una función de R para obtener el SVD. Aplícalo para datos en 2D con la siguiente estructura (toma en cada grupo 30 observaciones). Muestra cómo las proyecciones sobre el primer componente cambian en función de σ y cómo se aproximan a PCA si $\sigma \to \infty$.

  Resultados

  En la figura 3 se muestran las proyecciones obtenidas con PCA y con KPCA (kernel centrado de base radial con $\sigma = 0.05$)



Ok

Figure 3: Resultados de PCA

Las siguientes proyeccciones de KPCA fueron obtenidas usando un kernel centrado de base radial con diferentes valores para el parámetro σ. La figura 4 muestra los resultados obtenidos para diferentes valores (entre 0 y 1) de σ. Podemos notar que ambos subconjuntos se separan para un σ pequeño y conforme σ crece los subconjuntos comienzan a mezclarse.

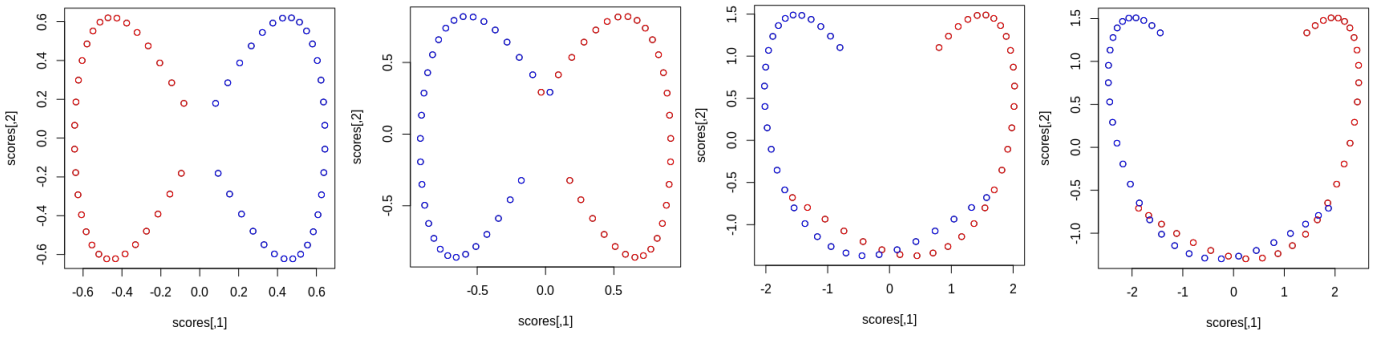


Figure 4: Resultados de KPCA con $\sigma = 0.05, 0.1, 0.5, 1.0$

Bien!

Por otro lado, cuando σ toma valores más grandes, las proyecciones se empiezan a parecer a las proyecciones obtenidas con PCA y también parecen tener la estructura del espacio original. La figura 5 muestra los resultados obtenidos para diferentes valores (grandes) de σ.

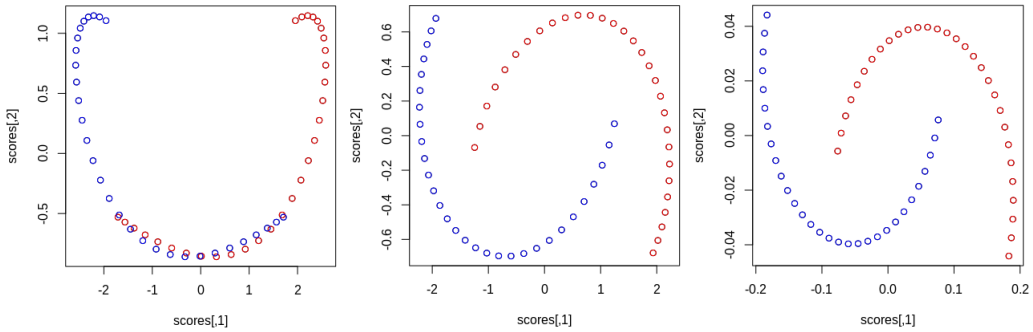


Figure 5: Resultados de KPCA con $\sigma = 2.0, 4.0, 100$

- Compara el resultado también con lo que da T-SNE.

  Resultados En la figura 6 se muestran las proyecciones obtenidas con t-sne usando diferentes valores para el parámetro de perplejidad. Podemos notar que valores de 5 y 16 obtuvieon resultados que logran separar ambos conjuntos de datos.

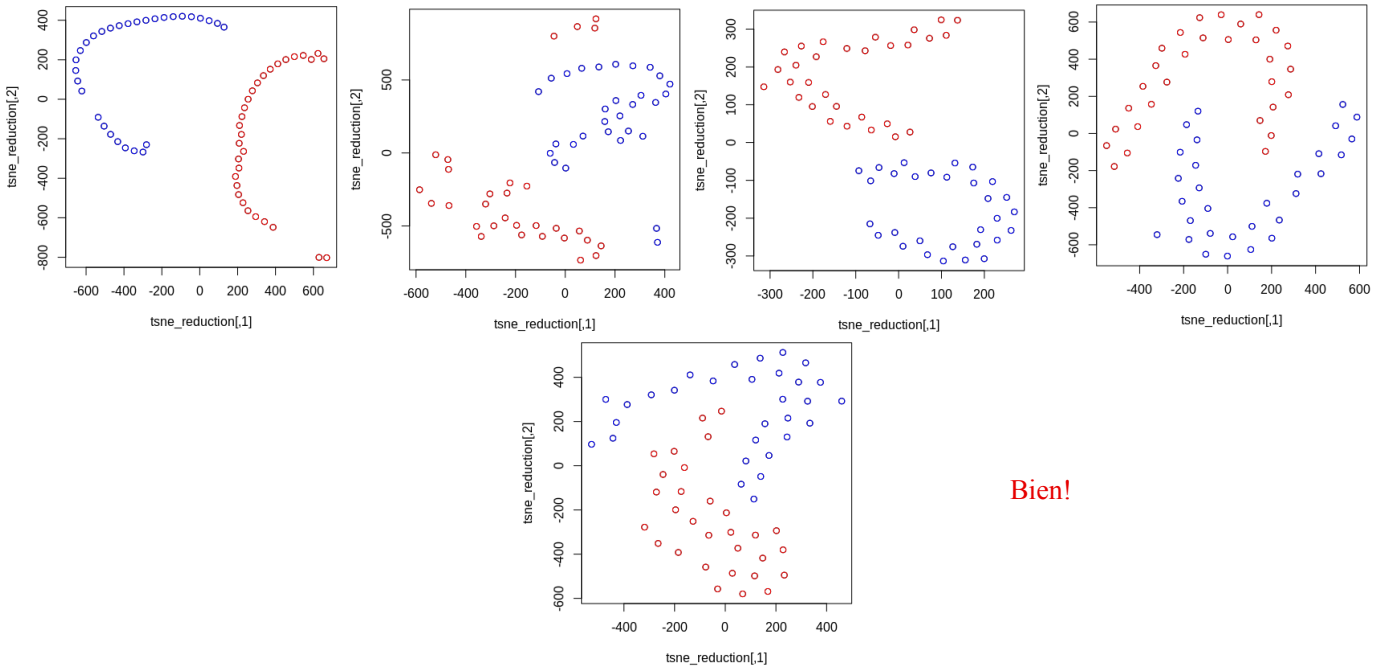


Figure 6: Resultados de T-SNE con perplexity= $5, 10, 16, 20, 25$

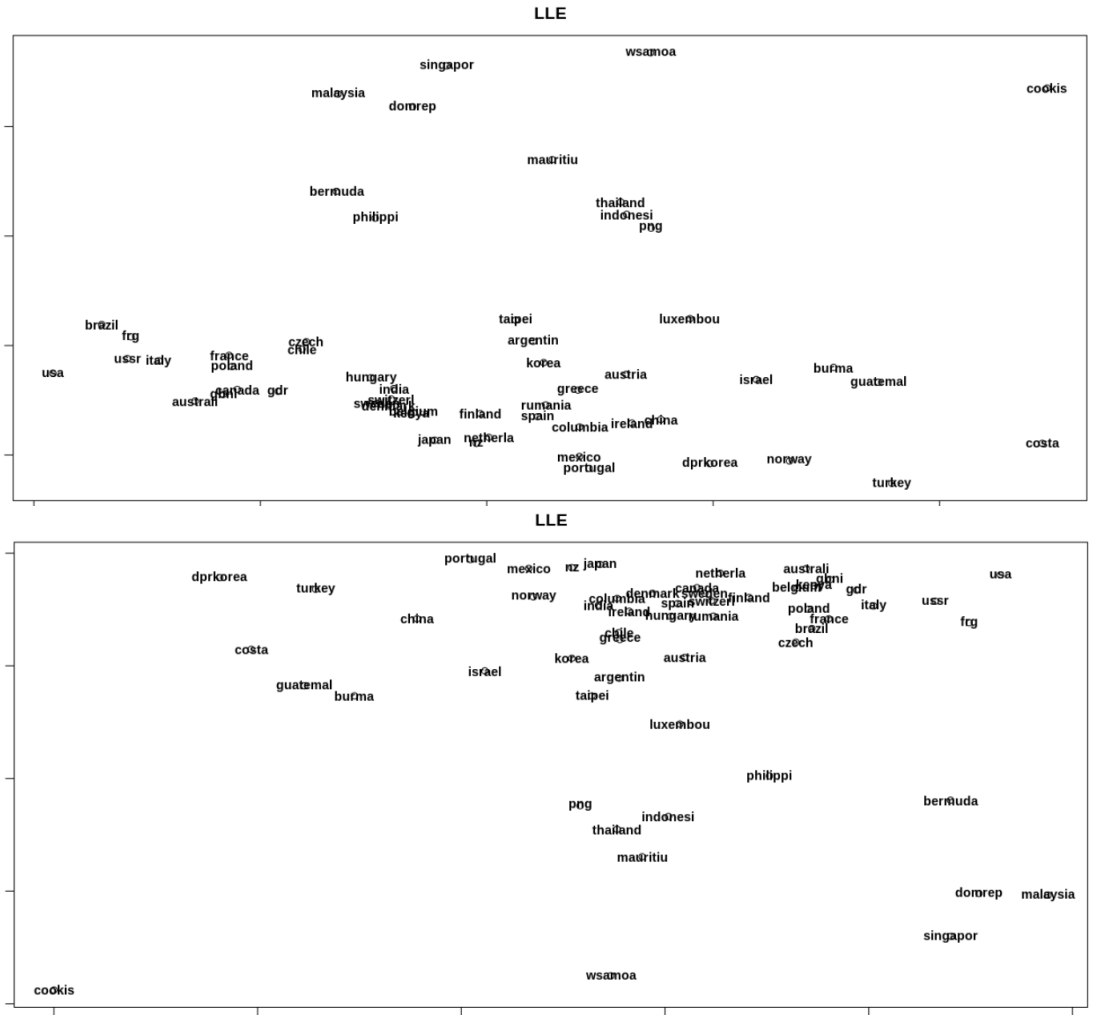
Al comparar resultados de T-SNE con los resultados de PCA y KPCA podemos notar que PCA y KPCA mantienen ciertas estructuras circulares como en los datos originales aún para diferentes valores de σ, mientras que varios de los resultados obtenidos con T-SNE modifican esas estructuras circulares. En conclusión, T-SNE y KPCA logran buenos resultados al transformar los datos para separarlos cuando se tienen los parámetros adecuados.

Ejercicio 2 **Nota: 3/3**

Usa ISOMAP, LLE, T-SNE y SOM's para encontrar visualizaciones informativas de los datos `deport.dat` que usaste en la primera tarea. El resultado debe ser un pequeño reporte con textos e imágenes integrados.

Solución

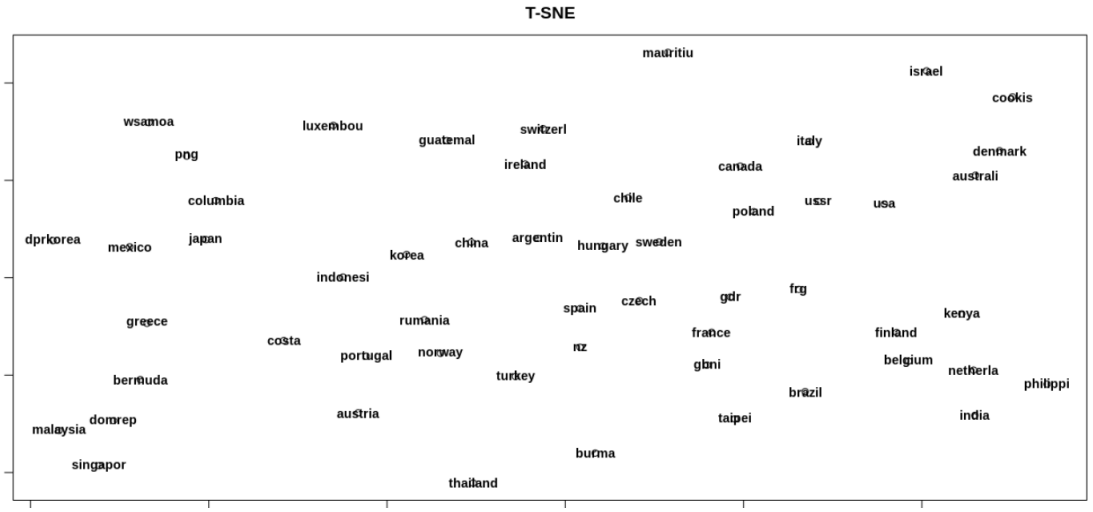
- **LLE:** Se presentan resultados para diferentes valores de k. Se puede apreciar en ambas gráficas que el país de Cookis se encuatra alejado de todos los demás, pues es el país con tiempos más grandes. También podemos notar que al otro extremo de este país se encuentran USA, USSR e Italia, entre otros, que obtuvieron tiempos menores. Otra cosa en común que tienen ambas visualizaciones es que Malasya también se encuentra en una esquina, que se mantiene alejada tanto de USA como de Cookis, mientras que en la esquina restante hay más variabilidad.



Ok

Figure 7: Resultados de LLE con $k = 10, 40$

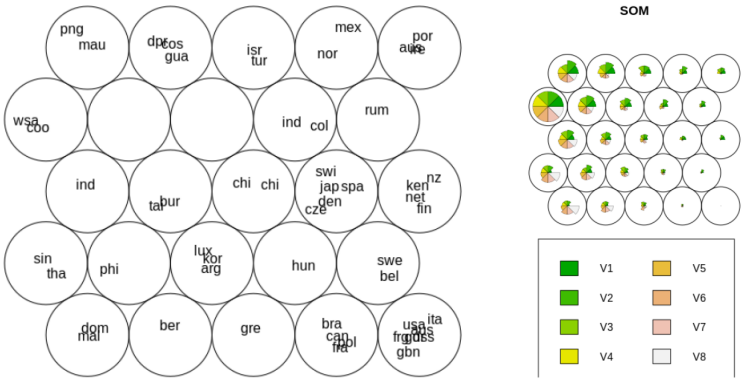
- **T-SNE:** En la figura se presentan resultados para diferentes valores de perplejidad. Podemos apreciar en todas las gráficas que los puntos se distribuyen bien en el plano. Por otra parte podemos notar que para un valor de perplejidad de 10, países con resultados muy distintos como USA y Cookis se encuentran cerca (Esquina superior derecha), mientras que países como Cookis y Wsamboa se encuentran lejos (Los cuales tienen tiempos grandes). Por otro lado, para un valor de perplejidad ??



Ok

Figure 8: Resultados de T-SNE con *perplexity* = 10

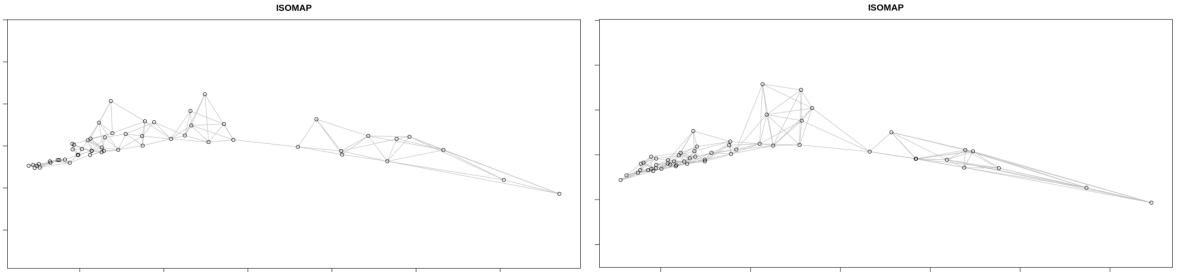
- **SOM:** En la figura se presentan los resultados de SOM. Podemos notar que los países con tiempos menores como USA e Italia pertenecen a la celda inferior derecha, mientras que países con tiempos más grandes, como Cookis o Wsamboa, pertenecen a celdas en de la esquina superior izquierda. Por otro lado, podemos notar que la gráfica derecha nos indica para cada círculo en qué prueba tuviste que tener un tiempo alto para pertenecer a dicho círculo. Por ejemplo podemos notar que Wsamboa y Cookies tivieron tiempos altos en todas las pruebas por igual. Tambien podemos apreciar que en círculos de la esquina inferior izquierda influye mayoritariamente la variable 8 (Gris) mientras que en los círculos de la esquina superior derecha influye más la variable 1, es decir, países con tiempos más altos en la prueba de menor distancia se encuentran cerca de la esquina superior derecha, mientras que países con mayor tiempo en la prueba de mayor distancia se encuentran en la esquina inferior izquierda. Por último, podemos notar que el círculo de la esquina inferior derecha parece estar en blanco, esto se debe a que para estar en este círculo se debe tener poco valor en todas las variables, por lo que en este círculo se encuentran los países con menores tiempos.



Bien!

Figure 9: Resultados de SOM; izquierda los nombres de países en cada rejilla; derecha importancia de cada variable

- **ISOMAP:** En la figura se presentan los resultados de ISOMAP tomando en cuenta diferentes valores para k (Cantidad de vecinos a tomar en cuenta). Observemos que en la parte izquierda se acumula la mayor parte de los países.



Ok

Figure 10: Resultados de ISOMAP con $k = 4, 6$

Comentarios: No logré colocar los nombres de los países a sus correspondientes puntos, pero casi estoy seguro que el punto alejado de todos los demás es el correspondiente a "Cookis", pues en todos los demás métodos destacó de la misma manera. También podemos observar que se acumulan puntos en el otro extremo de este país, los países con tiempos menores en cada prueba.

Conclusiones: Podemos notar que todos los métodos colocan a USA e Cookis en lugares alejados, pues USA es uno de los países con menor tiempo en cada prueba y por el contrario Cookis tiene tiempos mayores en las pruebas. Así mismo podemos observar que países con tiempos pequeños tienden a estar cerca de USA y países con tiempos mayores tienden a estar un poco más cerca de Cookis.