# Práctica 2, Exploración y Análisis de Datos: Reducción Dimensionado.

Universidad del País Vasco Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Julen Rodríguez Meneses, Mikel Salvoch Vilches

Octubre de 2024

## 1 Introducción

En esta memoria se recoge el trabajo práctico realizado sobre la reducción de dimensionado de un conjunto de datos presentado, conociendo únicamente las distancias (disimilitud) entre pares de elementos del conjunto y la clase a la que pertenecen. Adicionalmente, se disponen de nuevos elementos sin clasificar del que se conoce la distancia respecto a los elementos ya clasificados. Los datos provienen de espectroscopias de estiércol sólido heterogéneo de ganado vacuno.

## 2 Reducción mediante coordenadas principales

Para la reducción mediante coordenadas principales se decide tomar un valor de $q = 2$, si representamos los valores obtenidos para las componentes principales (figura 1), vemos que pronto empieza a descender hasta valores cercanos a cero incluso llegando a ser negativo. Para el caso de los valores propios negativos se decide ignorarlos ya que no aparecen en pocas compo-

nentes principales y ya hay muchas de ellas que están por encima de 0 y que consiguen explicar correctamente nuestros datos.

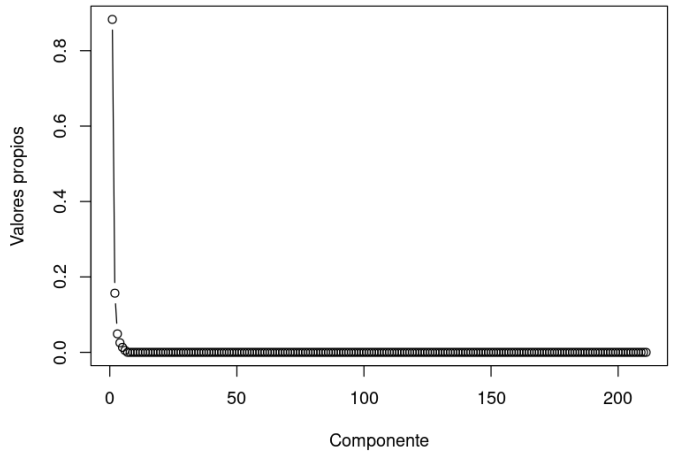


Figure 1: Elección de $q$, vistas todas las componentes principales.

Observando la figura 1 nos cuesta ver algo claro aunque ya podemos intuir que para valores bajos de $q$ podemos explicar la mayor parte de la variabilidad de los datos. Si observamos solamente las 10 primeras componentes principales como en la figura 2 vemos que con las dos primeras componentes es suficiente. Si calculamos el porcentajde variabilidad que explicamos mediante las dos primeras componentes principales nos sale que es del 91.81%.

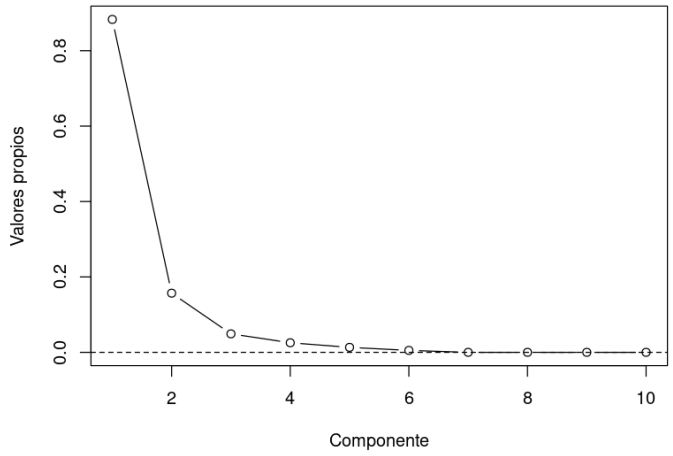


Figure 2: Elección de $q$, para las 10 primeras componentes principales.

A continuación, presentamos los nuevos datos y los datos originales coloreados por clases en el nuevo espacio construido para las dos coordenadas principales (figura 3).

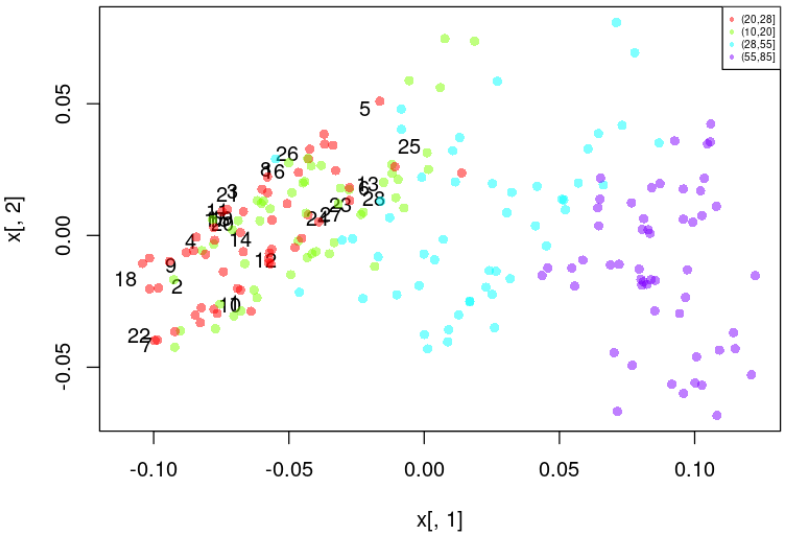


Figure 3: Representación de los datos en el nuevo espacio.

# 3 Reducción mediante kernel PCA

Si ahora hacemos esto mismo pero utilizando un Kernel Gaussiano con $\sigma = 0.2$ vemos en la figura 4 que nuevamente las dos componentes principales son las que más información nos aportan sobre nuestros datos. Concretamente, explicamos una variabiliad del 86,26% con $q = 2$.

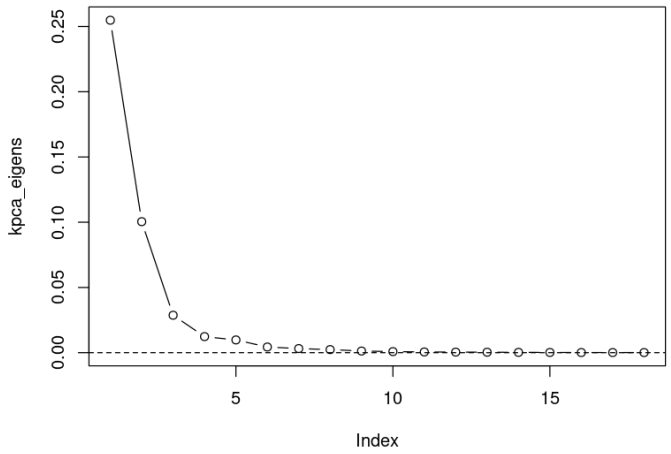


Figure 4: Elección de $q$ viendo los valores propios.

Si ahora representamos nuevamente todos nuestros datos en el nuevo espacio que hemos

construido obtenemos la figura 5. Igualmente que en el caso anterior, colocamos los elementos ya clasificados por colores y los nuevos datos con una etiqueta numérica y sin clasificar.

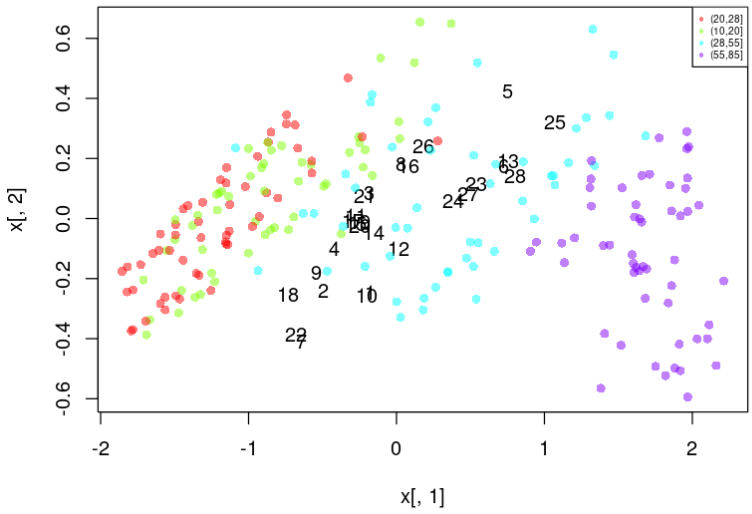


Figure 5: Representación de los datos en el nuevo espacio ($\sigma = 0.2$).

Si aplicamos el mismo procedimiento para un valor más alto de $\sigma = 0.8$ en la figura 6, vemos que el nuevo espacio construido no es muy representativo y no se observa ningún patrón como sí se puede intuir para $\sigma = 0.2$ en la figura 5. Más adelante se entra en mayor detalle sobre este caso particular.

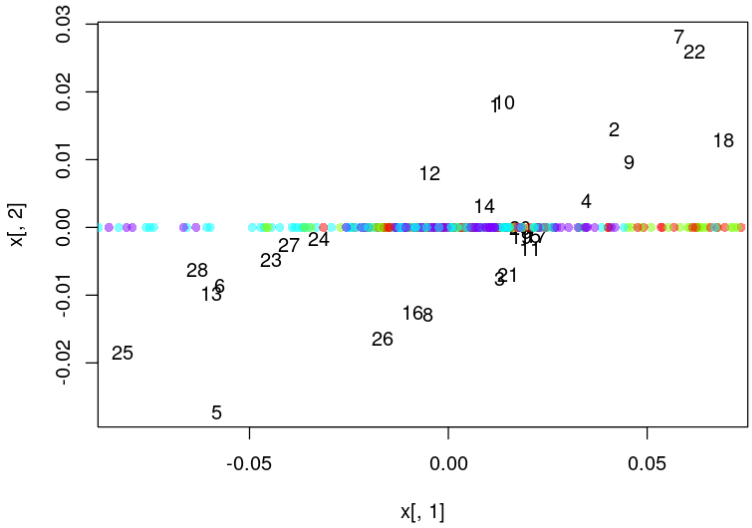


Figure 6: Representación de los datos en el nuevo espacio ($\sigma = 0.8$).

# 4 Medidas de bondad

Para ambas reducciones elegimos $K = 5$ para la medida local. Representamos primero los resultados para la primera representación: la medida global en la figura 7 y la local en la figura 8.

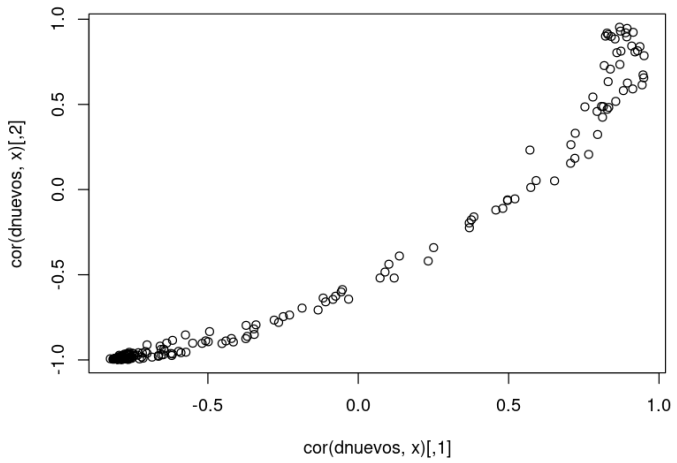


Figure 7: Medida global para la primera representación.

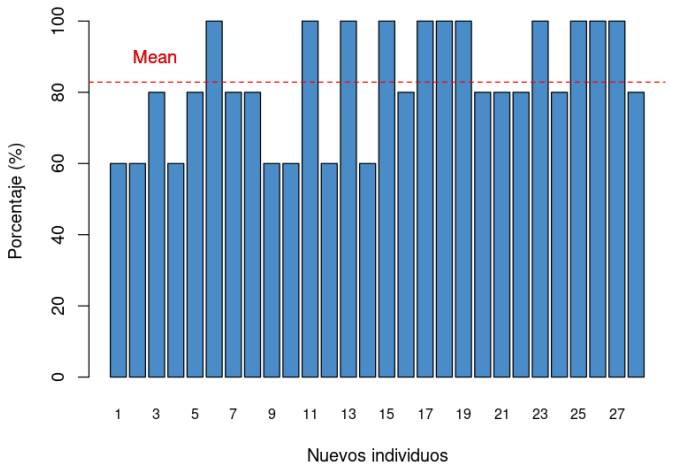


Figure 8: Medida local con $K = 5$ para la primera representación.

Hacemos lo mismo para el caso de la segunda representación. En primer lugar para $\sigma = 0.2$ en las figuras 9 y 10.

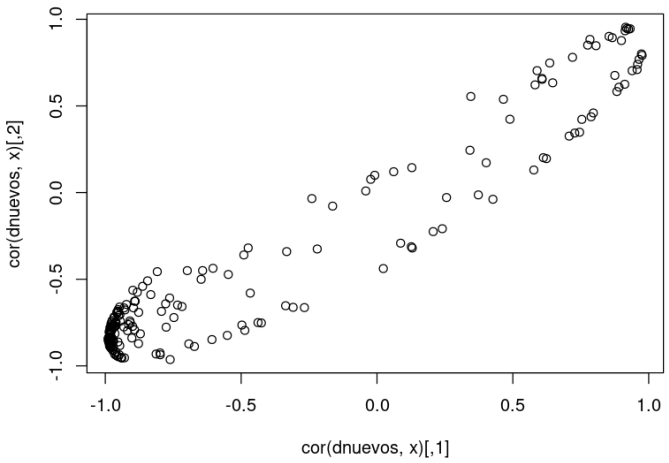


Figure 9: Medida global para la segunda representación con $\sigma = 0.2$.

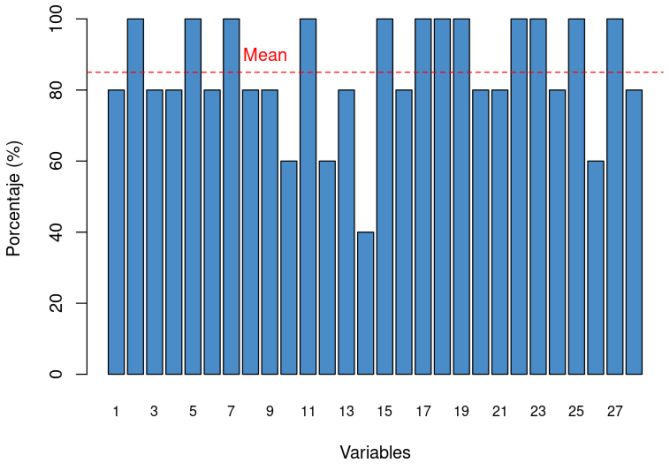


Figure 10: Medida local con $K = 5$ para la segunda representación con $\sigma = 0.2$.

Para $\sigma = 0.8$, todos los valores tienden a cero como se ve en las figura 11, siendo muchos de estos valores también negativos. Por ello, no es posible trabajar adecuadamente en este espacio ni calcular de manera representativa algunas de las métricas como la correlación ya que la proyección de los individuos nuevos resulta en números imaginarios (raíces negativas).

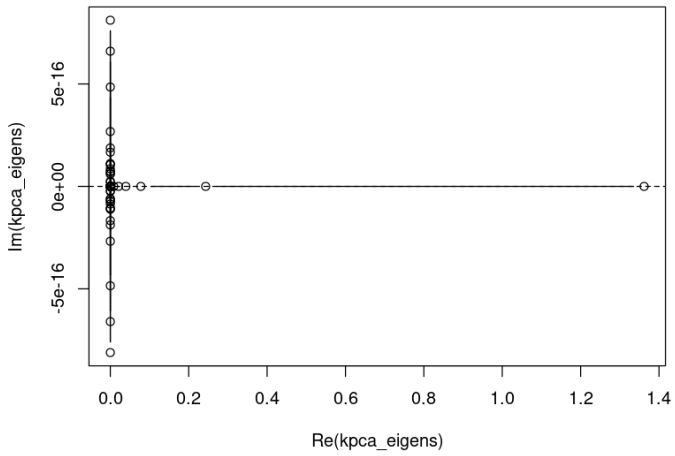


Figure 11: Valores propios para la segunda representación con $\sigma = 0.8$.

# 5 Conclusiones

Analizando tanto los resultados como las representaciones gráficas, vemos que para el primer caso los nuevos individuos en el nuevo espacio parecen tener sentido y estar agrupados al igual que los elementos originales. Si analizamos las medidas vemos que las distancias se conservan bastante bien entre el espacio original y la nueva proyección obteniendo un porcentaje medio de 82.86 para la medida local.

Para el caso de la segunda representación, como ya se ha mencionado, parece que el valor de $\sigma$ influye significativamente en la reducción del dimensionado, en el caso de $\sigma = 0.2$ se obtiene una representación gráfica (figura 5) que parece aportarnos información sobre la distribución y clasificación de los elementos (tanto los nuevos como los originales), sin embargo en la representación para $\sigma = 0.8$ (figura 6) no parece que las distancias entre los nuevos individuos y los originales se mantenga muy fiel. La gráfica de las figuras 7 y 9 muestran la correlación entre las distancias originales y las distancias en el nuevo espacio reducido. Cada punto representa cada uno de los individuos. Al estar la gran mayoría de ellos en la diagonal, se muestra que se preserva una buena relación de distancias en ambas dimensiones. Los puntos que se encuentran a los extremos muestran mayor variabilidad que la mayoría. Esto puede ser debido a:

- Primera representación $\rightarrow$ Las distancias no estan bien capturadas al no ser tan lineales
- Segunda representación $\rightarrow$ El valor sigma puede habernos afectado ya que si es muy bajo se enfoca mucho en las relaciones locales y si es muy alto en las globales

Si observamos la medida local, vemos que de media en $\sigma = 0.2$ obtenemos un 86% y para $\sigma = 0.8$ no se pueden calcular las métricas por el motivo explicado. Esto nos hace ver la importancia de elegir un valor adecuado para $\sigma$ y la importancia que este parámetro tiene a la hora de hacer la reducción de la dimensionalidad.

Finalmente, merece destacar que para la primera de las representaciones los nuevos individuos parecen pertenecer a la clase roja (puede que a la verde) si reparamos en la distancia euclídea (así a ojo) en la figura 3. Por el contrario, al haber aplicado el Kernel PCA y siguiendo esta misma lógica, viendo la figura 5 uno podría pensar que los nuevos individuos encajan más en la clase cian. La realidad observando las espectroscopías que aparecen en el enunciado de la práctica es que los nuevos datos parecen pertenecer a la clase roja o verde (tampoco está claro). Lo que queda claro es que la aproximación primera parece estar mejor adecuada a la realidad de los datos.