PCA: compañías
Tarea 5.1

Vamos a trabajar con el conjunto de datos consistente en 79 compañías estadounidenses. Llamemos Xf a la tabla.

1. Según lo que hemos visto hasta ahora con estos datos, ¿crees que es mejor estandarizar las variables o no? ¿por qué?

2. Estandariza univariantemente cada columna de Xf y guarda el resultado en $Xest$. Halla la matriz R de correlaciones de los datos originales Xf y la matriz de covarianza de los datos estandarizados $Xest$. ¿Coinciden? ¿Por qué?

3. Halla los autovalores de la matriz R y con ellos la variabilidad explicada acumulada.

4. ¿Cuántos componentes principales habría que elegir si quisiéramos que explicaran más de un 98 % de la variabilidad?

5. Haz un screeplot de estos resultados. Interprétalo.

6. Escoge los primeros 3 componentes principales. Obtén los scores y llama a las nuevas columnas CP1, CP2 y CP3.

7. Todos los valores de CP1 son negativos. Todas las variables están aproximadamente igual representadas de forma negativa. ¿Cómo se podría interpretar este componente en función del tamaño de la empresa?

8. El segundo componente CP2, ¿entre qué tipo de empresas distingue?

9. El tercer componente CP3, ¿qué variables destaca principalmente?

10. Dibuja tres scatterplots de los datos en función de los componentes principales usando los scores. ¿Hay empresas que estén muy alejadas del resto? ¿Tiene sentido?

PCA: Cáncer
Tarea 5.2

Vamos a utilizar el dataset de imágenes de cáncer de mama que vimos anteriormente.

1. Según el estudio exploratorio y descriptivo que hicimos de estos datos, te parece que las variables estén altamente correlacionadas o no?

2. El análisis de componentes principales ayuda cuando tenemos muchas variables y además están altamente correlacionadas (multicolinealidad). Así que vamos a aplicarlo. ¿Te parece mejor usar la matriz de covarianza o la de correlaciones para hallar los componentes? ¿Por qué?

3. Utiliza la matriz R de correlaciones para hallar los PC y haz un screeplot de los resultados. ¿Con cuántos componentes te quedarías si quieres que por lo menos expliquen hasta un 85 % como mínimo de la variabilidad?

4. Selecciona esa cantidad de componentes y mira las correlaciones con respecto a las variables originales. ¿Son altas o bajas? ¿esto es bueno o malo?