Uso de PCA en MATLAB

Ulises Jiménez Guerrero

3 de abril de 2025

Se realizaron pruebas del funcionamiento del análisis de componentes principales, PCA, en una base de datos con una dimensionalidad mayor a 3. Esto con el objetivo de que no se pueda visualizar de manera directa. Para este análisis se utilizo la función `pca(data, options)`, parte del S*tatistics and Machine Learning toolbox*.

Se decidió utilizar una base de datos relativamente sencilla, siendo esta Wine Quality, disponible en el repositorio para aprendizaje automático de UC Irvine [Cortez and Reis, 2009]. Esta describe la calidad de vino, en un valor de 1 a 10, en relación con diferentes características químicas. Estas se describen de la siguiente manera en la documentación:

Variable Name	Role	Type	Description	Units	Missing Values
fixed_acidity	Feature	Continuous			no
volatile_acidity	Feature	Continuous			no
citric_acid	Feature	Continuous			no
residual_sugar	Feature	Continuous			no
chlorides	Feature	Continuous			no
free_sulfur_dioxide	Feature	Continuous			no
total_sulfur_dioxide	Feature	Continuous			no
density	Feature	Continuous			no
pH	Feature	Continuous			no
sulphates	Feature	Continuous			no
alcohol	Feature	Continuous			no
quality	Target	Integer	score between 0 and 10		no
color	Other	Categorical	red or white		no

Por tanto, tenemos un total de 11 *feautures* para explicar la calidad del vino. Se encuentra una base de datos para vino rojo y otra para vino rojo, y en nuestro caso se eligió la segunda. Esta contiene un total de 1599 observaciones.

Se utilizó el siguiente script para realizar el análisis. Se comentarán los resultados pertinentes.

```
wine_quality = readtable("winequality-red.csv");
quality = wine_quality.quality;
data = table2array(wine_quality);
data = data(:, 1:end-1);

[coeff, score, latent, ~, explained] = pca(data);

gscatter(score(:, 1), score(:, 2), quality)
```

Se tiene la siguiente matriz para los 11 componentes principales:

$$\begin{bmatrix}
-0.0061 & -0.0239 & 0.9531 & -0.2651 & 0.0981 & 0.0793 & 0.0111 & 0.0649 & -0.0162 & 0.0112 & -0.0009 \\
0.0004 & -0.0020 & -0.0251 & 0.0073 & -0.0412 & 0.6892 & 0.4644 & -0.3388 & 0.4286 & -0.0916 & -0.0007 \\
0.0002 & -0.0030 & 0.0737 & -0.0098 & 0.0415 & -0.5040 & -0.2055 & -0.3271 & 0.7605 & -0.1055 & -0.0001 \\
0.0086 & 0.0111 & 0.2809 & 0.9432 & -0.1766 & -0.0058 & 0.0025 & 0.0041 & -0.0069 & -0.0015 & -0.0004 \\
0.0001 & -0.0002 & 0.0029 & -0.0006 & -0.0095 & -0.0570 & 0.1139 & -0.1325 & 0.0722 & 0.9802 & -0.0018 \\
0.2189 & 0.9753 & 0.0209 & -0.0212 & -0.0079 & 0.0011 & -0.0001 & -0.0026 & 0.0015 & -0.0003 & 0.0000 \\
0.9757 & -0.2189 & -0.0015 & -0.0040 & 0.0103 & 0.0006 & -0.0002 & 0.0012 & -0.0007 & 0.0003 & 0.0000 \\
0.0000 & -0.0000 & 0.0008 & 0.0001 & -0.0008 & 0.0005 & 0.0016 & 0.0041 & 0.0028 & 0.0020 & 0.9999 \\
-0.0003 & 0.0033 & -0.0586 & 0.0206 & 0.0126 & 0.1420 & -0.0103 & 0.8592 & 0.4788 & 0.0906 & -0.0051 \\
0.0002 & 0.0006 & 0.0175 & -0.0072 & 0.0236 & -0.4884 & 0.8537 & 0.1329 & -0.0536 & -0.1056 & -0.0013 \\
-0.0064 & 0.0146 & -0.0486 & 0.1976 & 0.9771 & 0.0508 & 0.0082 & -0.0218 & -0.0180 & 0.0101 & 0.0009
\end{bmatrix}$$

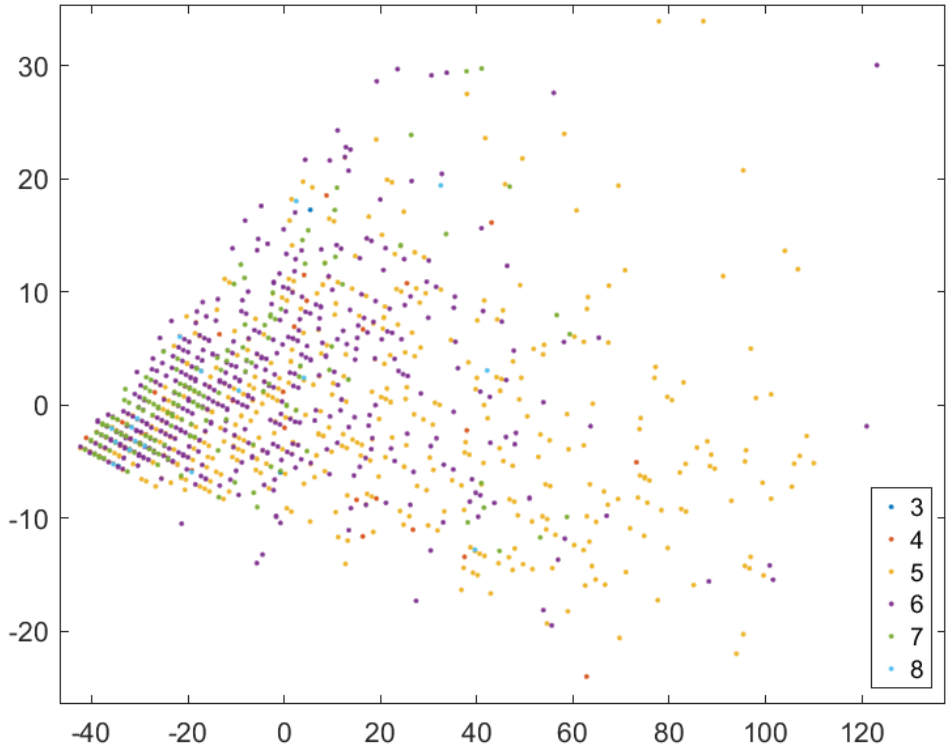
Más importante para el análisis, se tienen los siguientes valores para la varianza correspondiente a cada componente:

$$\begin{bmatrix}
1133.8071 & 57.9354 & 3.1013 & 1.8194 & 1.0463 & 0.0414 & 0.0232 & 0.0113 & 0.0101 & 0.0015 \\
5.61 \times 10^{-7}
\end{bmatrix}$$

Se nota una diferencia importante en cada componente, donde el primero tiene una magnitud significativamente más grande que el segundo, y el resto tienen un valor poco significativo. A partir del sexto componente se tienen valores menores a 1, y el último tiene un valor muy cercano a cero. Esto se destaca más al ver el porcentaje de varianza explicada por cada componente, encontrado en el cuadro 1. Con solamente dos componentes principales se explica más del 98 % de la varianza, cantidad muy representativa de los datos.

Finalmente, se utilizan los primeros dos componentes para realizar una reducción de la dimensionalidad y visualizar los datos, con los resultados mostrados en la figura enseguida. Se destaca el hecho, mencionado en clase, de que esta técnica no realiza una separación de las clases, y esta visualización permite observarlo de manera directa. Los datos de diferentes clases se encuentran mezclados, sin que se observe un clúster o agrupación en particular.

Figura 1: Graficación de dos componentes principales, coloreando en base al valor de la calidad de cada dato.



Cuadro 1: Porcentaje de varianza explicada por cada componente principal

Componente	**Varianza Explicada (%)**
1	94.657 698
2	4.836 830
3	0.258 917
4	0.151 897
5	0.087 355
6	0.003 456
7	0.001 936
8	0.000 947
9	0.000 841
10	0.000 121
11	4.69×10^{-8}

Referencias

[Cortez and Reis, 2009] Cortez, Paulo, C. A. A. F. M. T. and Reis, J. (2009). Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.