TRABAJO DE CURSO

MNC 21/22

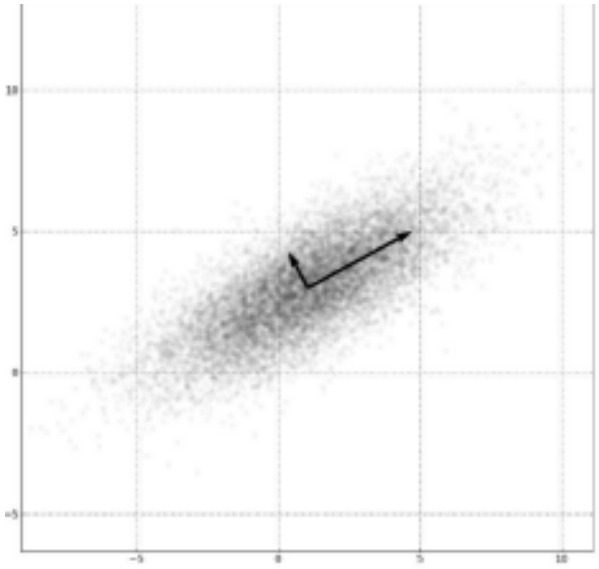
Análisis de Componentes Principales

Brian Palmés Gómez

1. Introducción

La técnica denominada PCA (Principal Component Analysis) es empleada en áreas como la Visión por Computador o la Minería de Datos para simplificar la representación de datos masivos y obtener una descripción más compacta de los mismos.

Dado un conjunto de puntos (instancias o muestras), se desea obtener las coordenadas de esos puntos en un sistema de representación de nuevas coordenadas (dimensiones o características) que están centrada en el conjunto de datos y rotadas en un alineamiento a las direcciones principales del agrupamiento de los datos.



2. Selección del conjunto de datos

Cada grupo deberá seleccionar un conjunto de puntos con el que trabajar distinto, elegido de las bases de datos disponibles en el repositorio de Machine Learning de la Universidad de California en Irvine (UCI): https://archive.ics.uci.edu/ml/datasets

Deberá escogerse una base de datos con al menos cinco características numéricas de tipo real y con más de 200 instancias. Anotarla al crear el grupo para el trabajo de curso, a fin de evitar duplicidades.

Para nuestro trabajo hemos elegido de la web el siguiente data set:
https://archive.ics.uci.edu/ml/datasets/Room+Occupancy+Estimation

Que es una recopilación de datos de estancia en habitaciones de hotel.

3. Solución con Matlab/Octave

Deberán resolverse las siguientes fases:

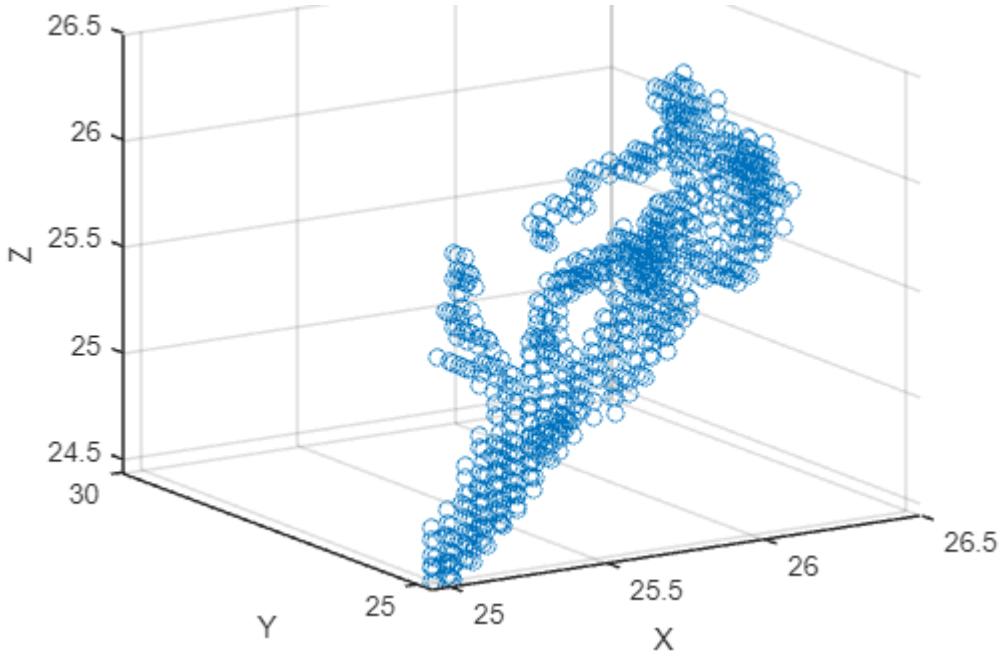
1. Extraer del fichero de datos las características de tipo real.
Se generará una matriz X de m filas (instancias) por n columnas (dimensiones)

```
clc;
clear;

%Extraccion de datos
data = readtable("ocupancy.csv");

%Cogemos solo las columnas que no dan un error de formato
t = data(:,3:16);
tabla = t.Variables;
```

Con estas instrucciones en Matlab leemos los datos del archivo cvs que contiene la información y lo almacenamos en la variable data que ahora es una tabla de matlab.

Para filtrar las columnas que no se pueden convertir a matriz ya que da un error de formato seleccionamos las columnas numéricas y con la instrucción t.Variables convertimos la tabla de matlab en una matriz de datos con la que sí podremos trabajar.

Se nos pide mostrar la nube de puntos que forman los datos antes de aplicar la técnica PCA.



2. Centrar los datos restando la media de cada componente, generando una matriz XC

A continuación mostramos cómo centramos la matriz en matlab:

realizamos iteraciones sobre la matriz XC a la que le vamos restando la media de cada valor para ir formando la matriz XC que contendrá los datos centrado.

```
%Primero vamos a centrar los datos de la matriz
%Para ello generamos la matriz XC

meanTabla = mean(tabla, 1);

XC = zeros(size(tabla)); %%Matriz de ceros
for i=1:size(tabla,2)
    XC(: , i) = tabla(:, i) - meanTabla(i);
end
%Se itera sobre la matriz con los datos, restandoles la media
%calculada de esa columna
```

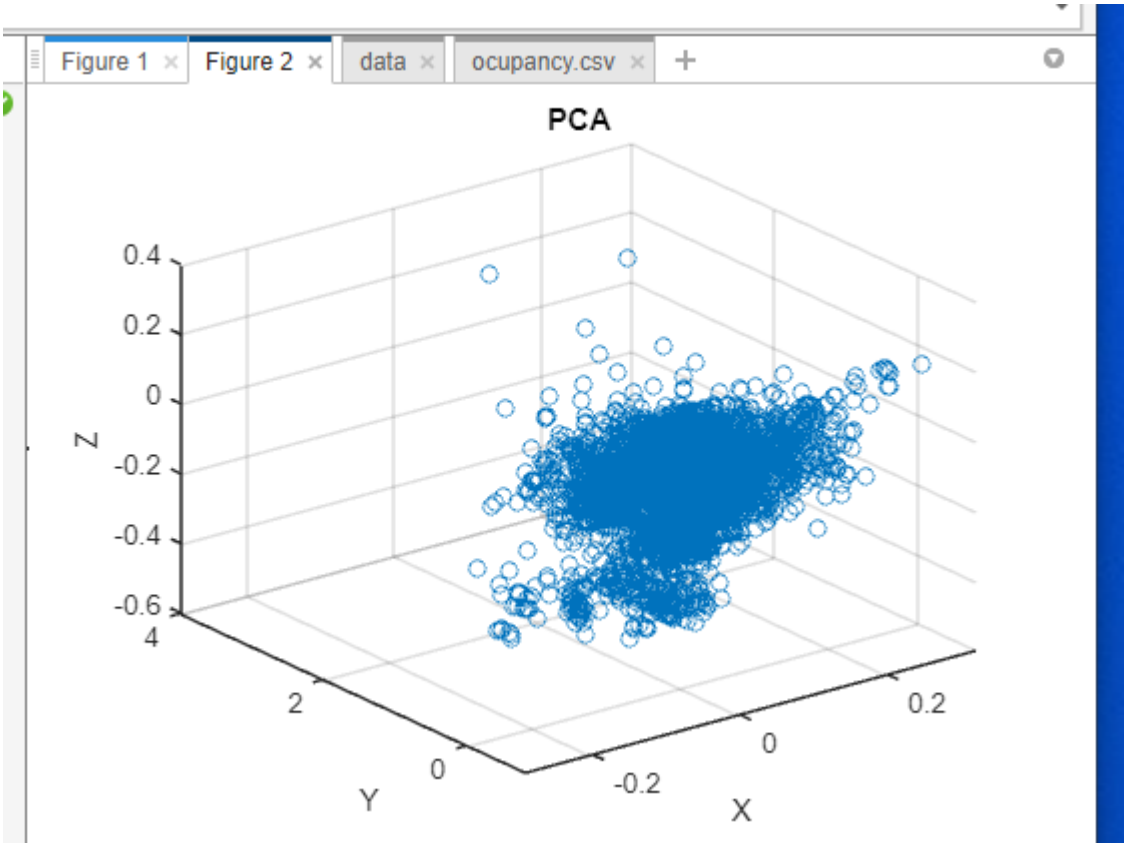
3. Calcular los autovalores y los autovectores de la matriz de covarianza Z = (XC'*XC)/m

Una vez los datos están centrados procedemos a hallar la matriz Z de covarianza.
Una vez la tengamos usaremos la función de matlab eig() para obtener los autovalores y los autovectores. almacenados en D y V respectivamente.

Luego multiplicamos la matriz centrada de datos por los autovectores para obtener los datos PCA.

```
%Pasamos a calcular Z = (XC`*XC)/m donde Z es la matriz de covarianza

Z = (XC'*XC)/(size(XC,1));

[V,D] = eig(Z); %La funcion eig() devuelve los autovalores y autovectores de la matriz de covarianza. V  es los autovectores y D

PCA = XC*V; %Reduccion de datos
```

4. Representar los datos y los autovalores principales



5. ¿Qué ocurre al multiplicar los datos por la matriz de autovectores?

En la teoría cuando multiplicas los autovectores por la matriz de datos centrada obtienes las coordenadas de los puntos de los datos de la matriz XC en un nuevo sistema de coordenadas rotadas hacia donde se agrupan los datos.

Cada columna de la Matriz PCA se corresponde con una combinación lineal de la matriz original extraída del cvs, pero está combinación lineal abarca menos carga de información.

Referencias :

https://www.youtube.com/watch?v=VqjJ5YYt78Y&ab_channel=SteveBrunton
https://www.youtube.com/watch?v=x-7BHjMA15M&ab_channel=C%C3%B3digoM%C3%A1quina
Y documentación de la asignatura.