# Machine Learning

# Seminar 8

1. Explain what Principal component analysis (PCA) is.

2. Consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane below.

- If we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x,y) value.
- The axes don't actually mean anything physical; they're combinations of height and weight called "principal components" that are chosen to give one axes lots of variation.



original data set
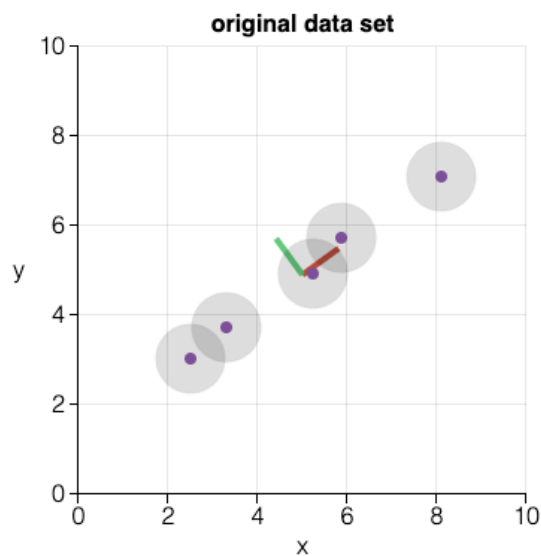
Try to manually draw a new coordinate system that can help us reduce the number of dimension to 1 without losing much information.

3. With three dimensions, PCA is more useful, because it's hard to see through a cloud of data. In the example below, the original data are plotted in 3D, but you can project the data into 2D through a transformation no different than finding a camera angle: rotate the axes to find the best angle. Check the following figures showing the results of PCA transformation.
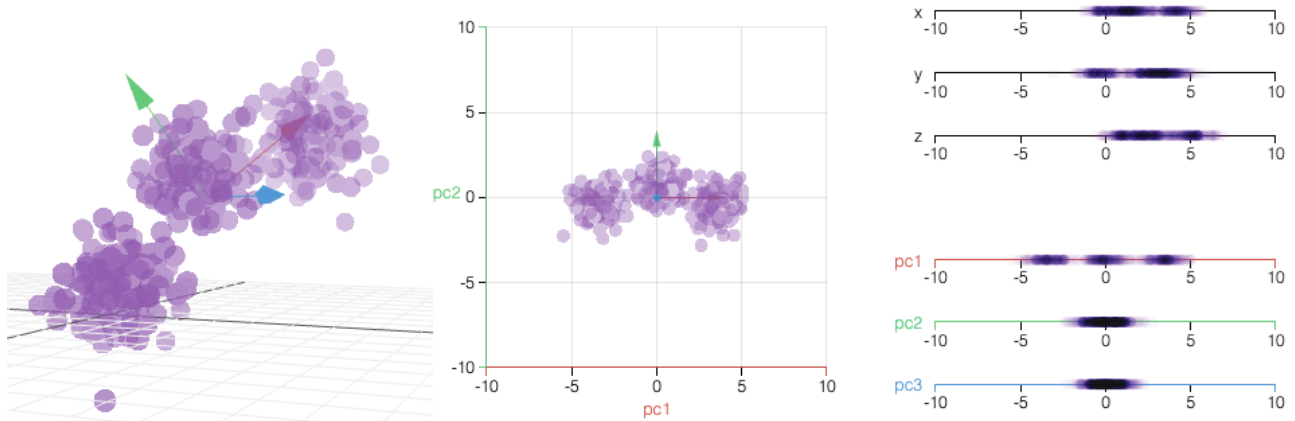


The rightmost figure shows the variations of the original variables and derived three principal components. Which PCs should we drop if we want to reduce the number of dimensions to 2? Which PCs should we drop if we want to reduce the number of dimensions to 1?

4. In the table is the average consumption of 17 types of food in grams per person per week for every country in the UK. The table shows some interesting variations across different food types, but overall differences aren't so notable. Check the PCA results and answer the following questions.

| | England | N Ireland | Scotland | Wales |
|---|---|---|---|---|
| Alcoholic drinks | 375 | 135 | 458 | 475 |
| Beverages | 57 | 47 | 53 | 73 |
| Carcase meat | 245 | 267 | 242 | 227 |
| Cereals | 1472 | 1494 | 1462 | 1582 |
| Cheese | 105 | 66 | 103 | 103 |
| Confectionery | 54 | 41 | 62 | 64 |
| Fats and oils | 193 | 209 | 184 | 235 |
| Fish | 147 | 93 | 122 | 160 |
| Fresh fruit | 1102 | 674 | 957 | 1137 |
| Fresh potatoes | 720 | 1033 | 566 | 874 |
| Fresh Veg | 253 | 143 | 171 | 265 |
| Other meat | 685 | 586 | 750 | 803 |
| Other Veg | 488 | 355 | 418 | 570 |
| Processed potatoes | 198 | 187 | 220 | 203 |
| Processed Veg | 360 | 334 | 337 | 365 |
| Soft drinks | 1374 | 1506 | 1572 | 1256 |
| Sugars | 156 | 139 | 147 | 175 |
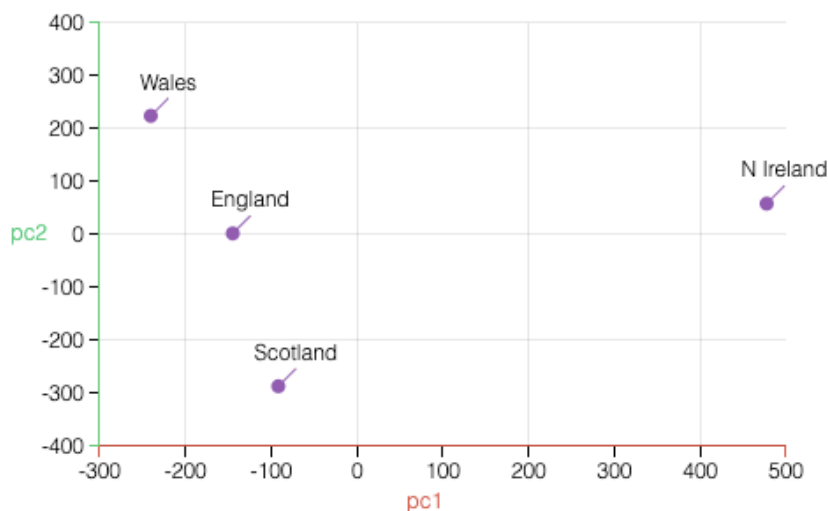


- How should we cluster the 4 countries based on the PCA results?
- How many groups? For each group, which countries should be included?
- Explain the reason for your clustering results.

5. The steps involved in PCA Algorithm are as follows-

- Step-01: Get data.
- Step-02: Compute the mean vector (μ).
- Step-03: Subtract mean from the given data.
- Step-04: Calculate the covariance matrix.
- Step-05: Calculate the eigen vectors and eigen values of the covariance matrix.
- Step-06: Choosing components and forming a feature vector.
- Step-07: Deriving the new data set.

Consider the two-dimensional patterns (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8), compute the principal component using PCA Algorithm.


6. Use PCA Algorithm to transform the pattern (2, 1) onto the eigen vector in the previous question.