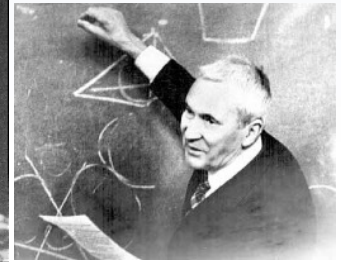
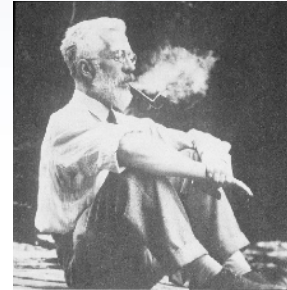
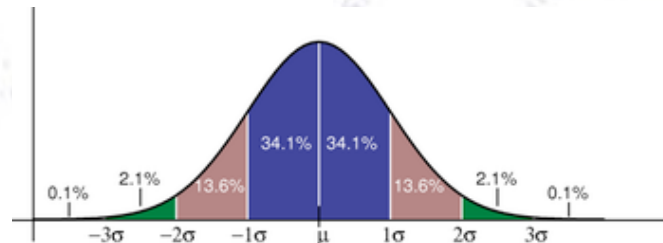


Applied ML

(Kernel) Principle Component Analysis



Troels C. Petersen (NBI)

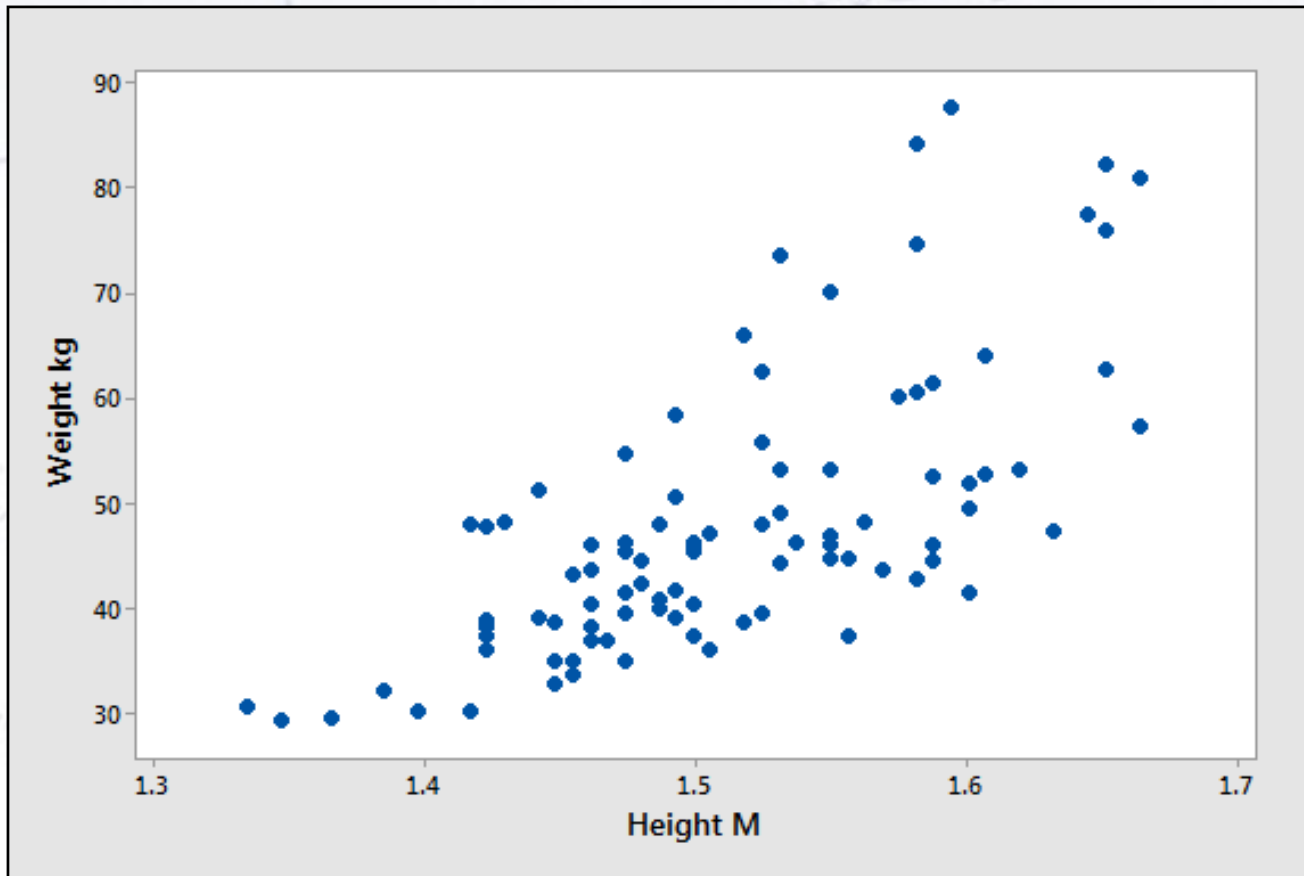


"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

PCA overview

Consider data which have correlations, here in 2D (for visualisation), but potentially in (very) high dimension.

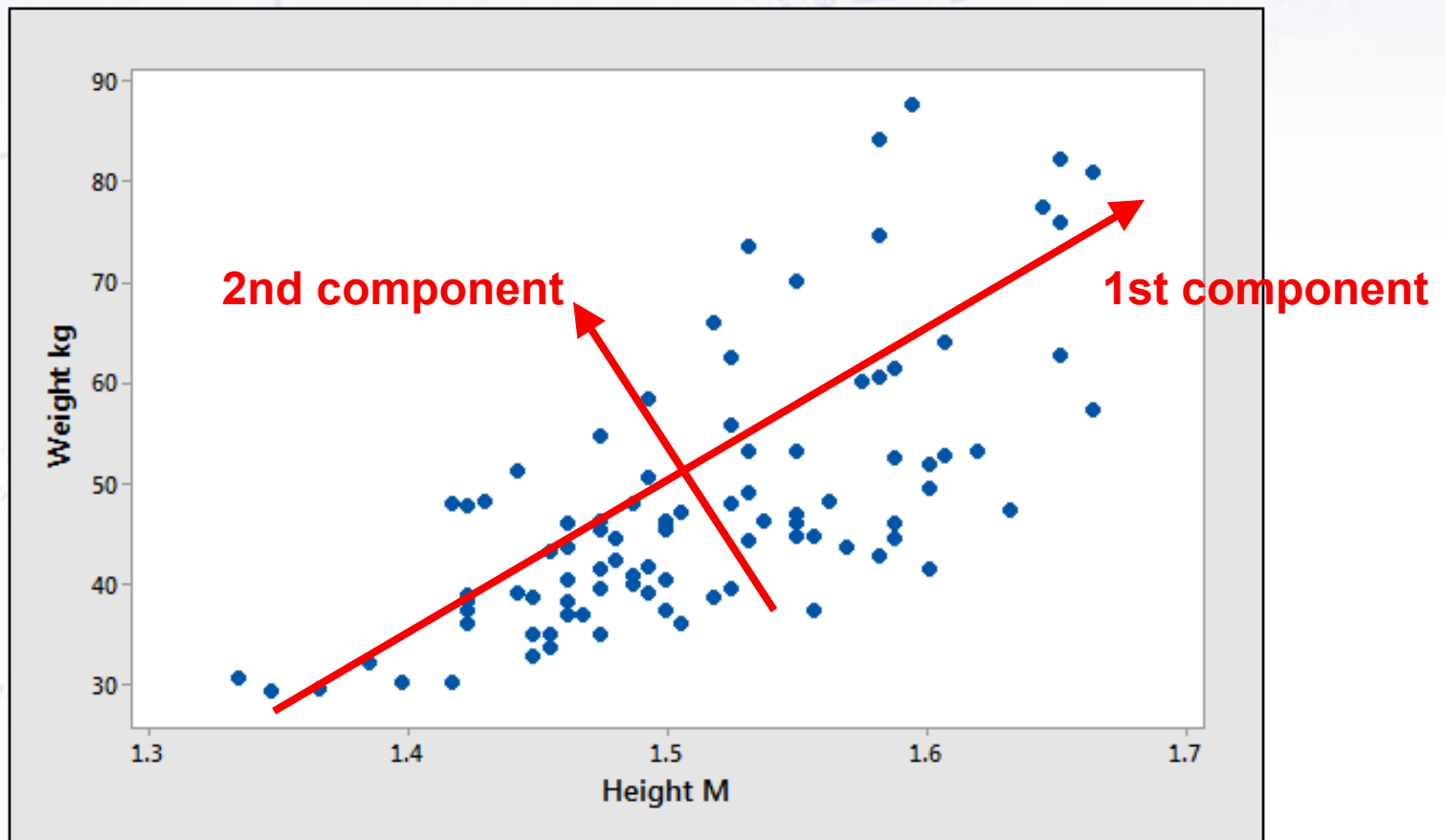
We want to apply a PCA to this data, to reduce dimensionality!



PCA overview

Find the direction, which has the maximum variance, i.e. “best along the direction of the data”.

The effective way to do this, is to find the eigenvectors and eigenvalues, and rank the eigenvectors (i.e. directions) according to eigenvalues.



PCA overview

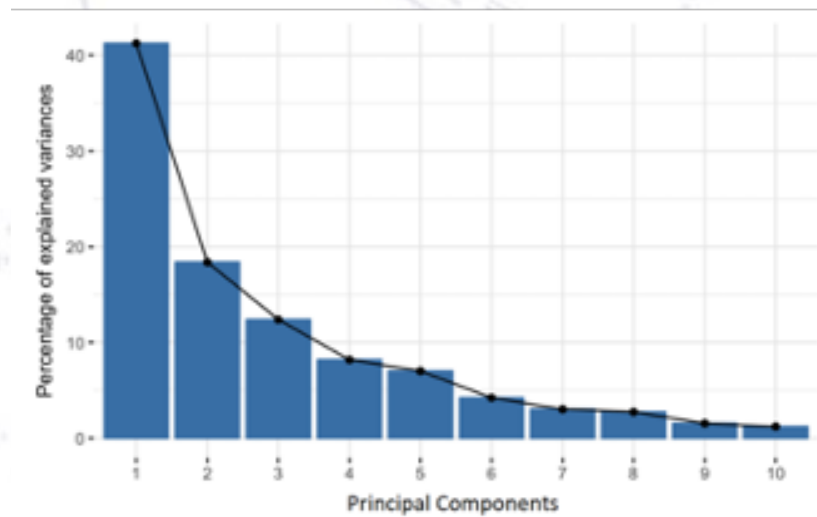
Find the direction, which has the maximum variance, i.e. “best along the direction of the data”.

The effective way to do this, is to find the eigenvectors and eigenvalues, and rank the eigenvectors (i.e. directions) according to eigenvalues.

PCA recipe:

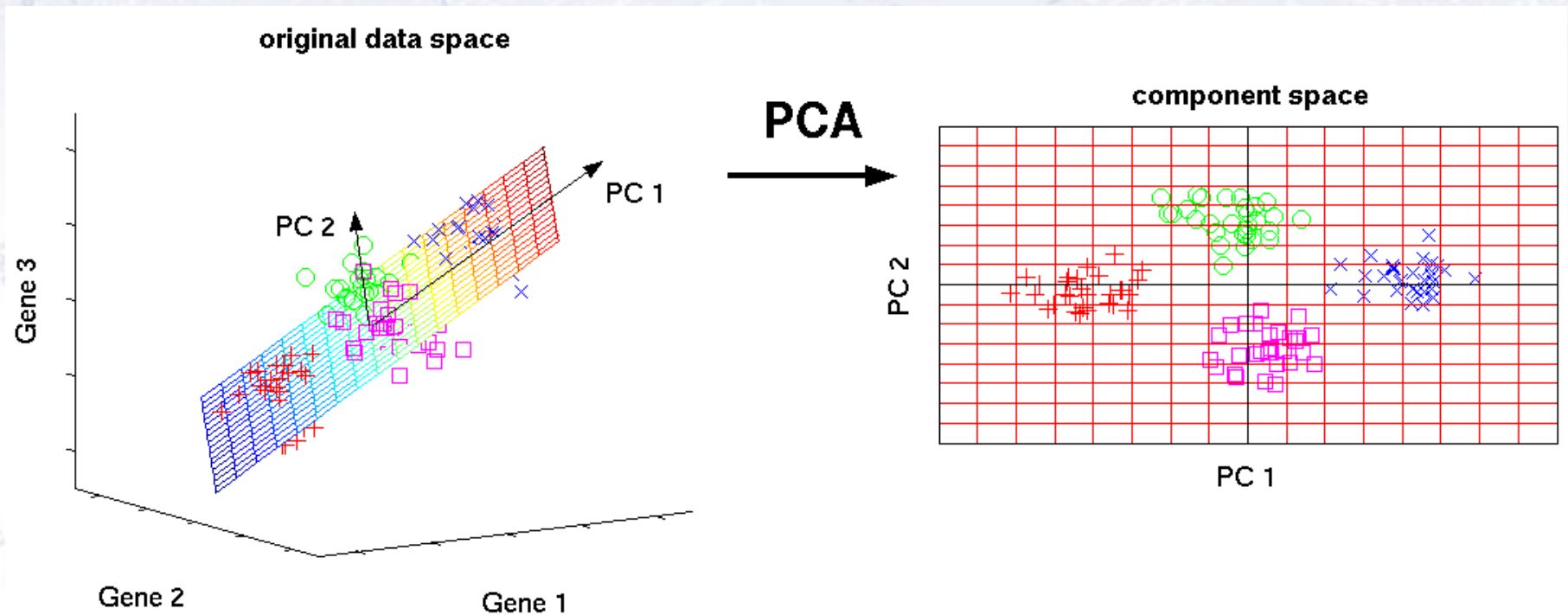
1. Standardize all variables
2. Compute covariance matrix
3. Compute eigenvectors and eigenvalues
4. Discard vectors/ directions that are not important enough

$$\begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Cov}(y, y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Cov}(z, z) \end{bmatrix}$$



PCA overview II

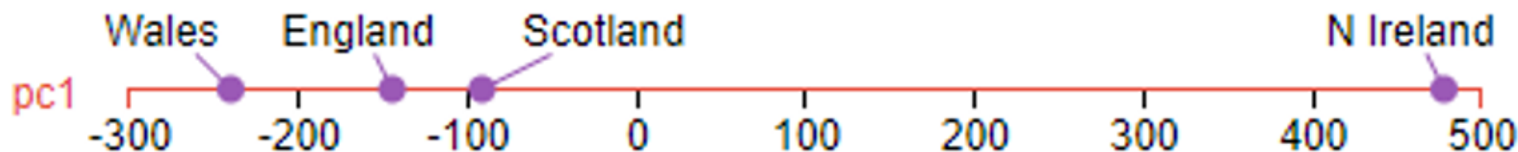
It is hard to illustrate the high dimensional cases, but here is an attempt at seeing 3D points reduced to 2D points by PCA.



Essentially, one finds the two (orthogonal) directions, which approximates the data best, and “throw away” all the other dimensions in this new space.

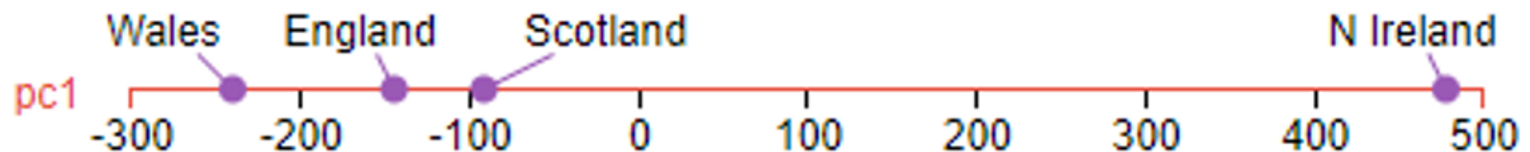
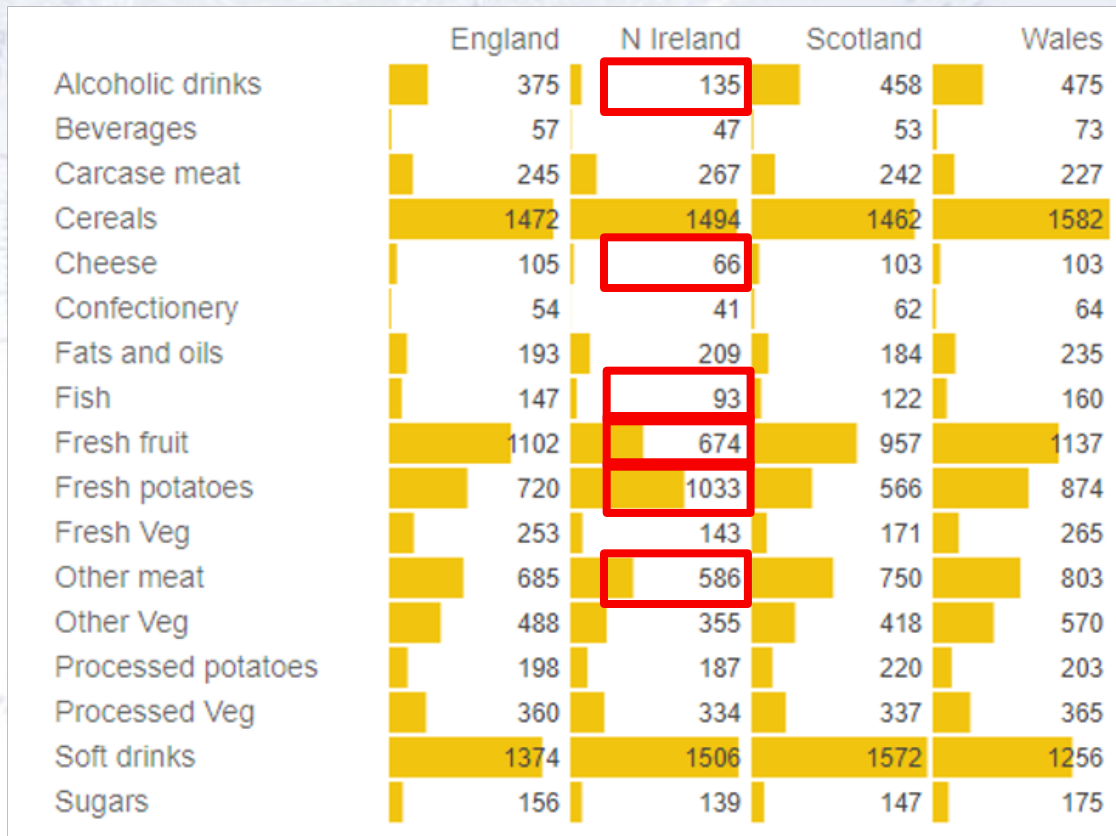
17-dimensional PCA

Considering the diet of UK parts, the PCA approach (again) clearly shows structure:



17-dimensional PCA

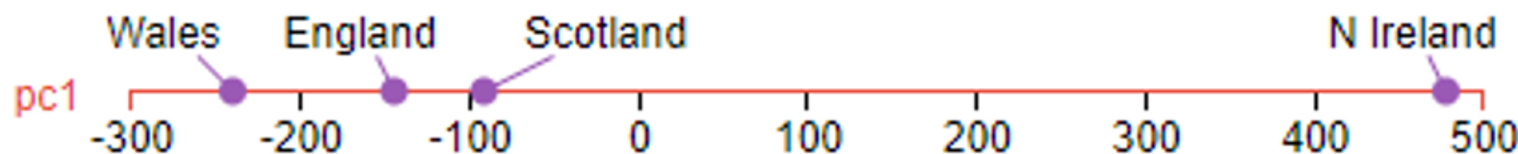
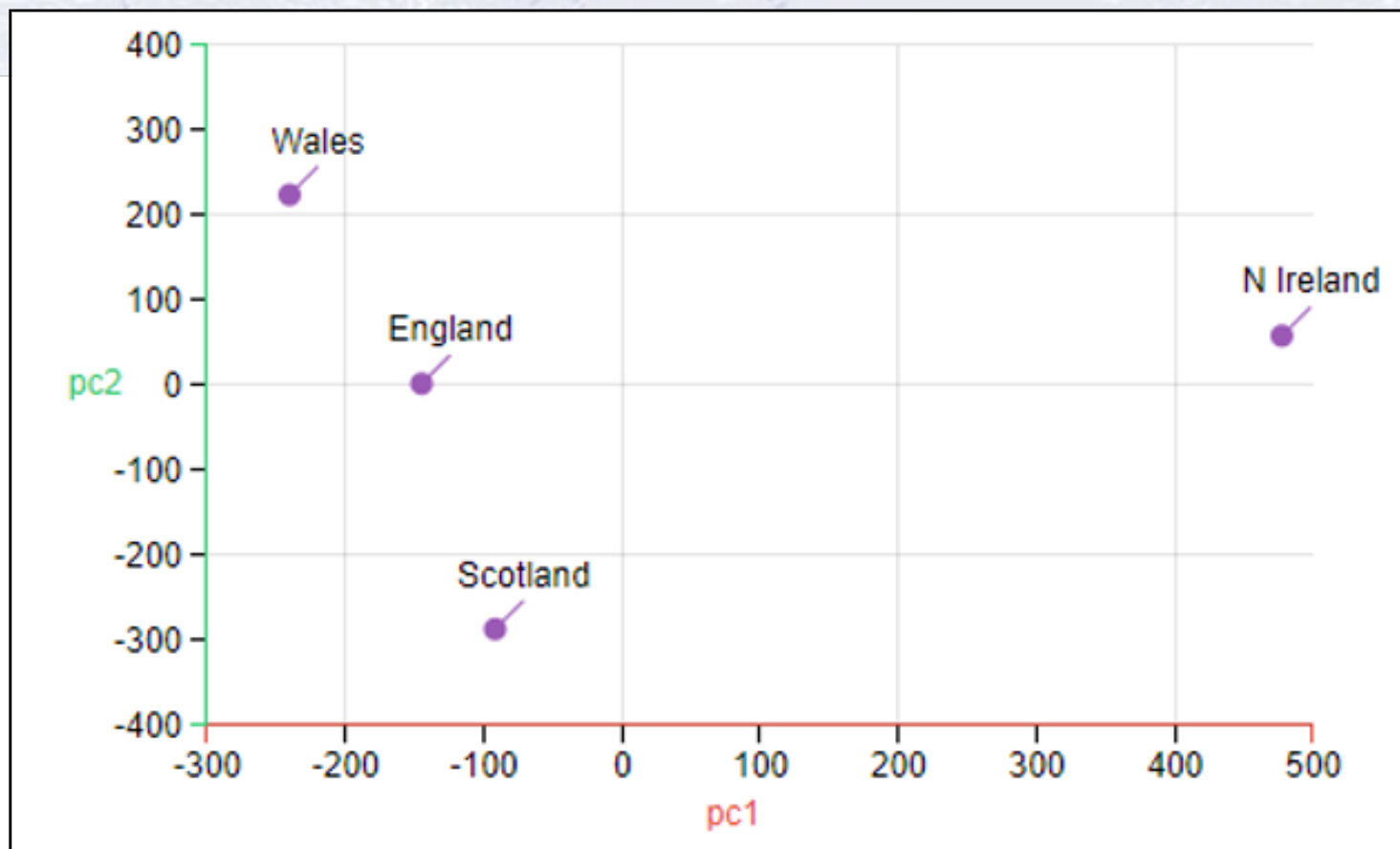
Considering the diet of UK parts, the PCA approach (again) clearly shows structure:



17-dimensional PCA

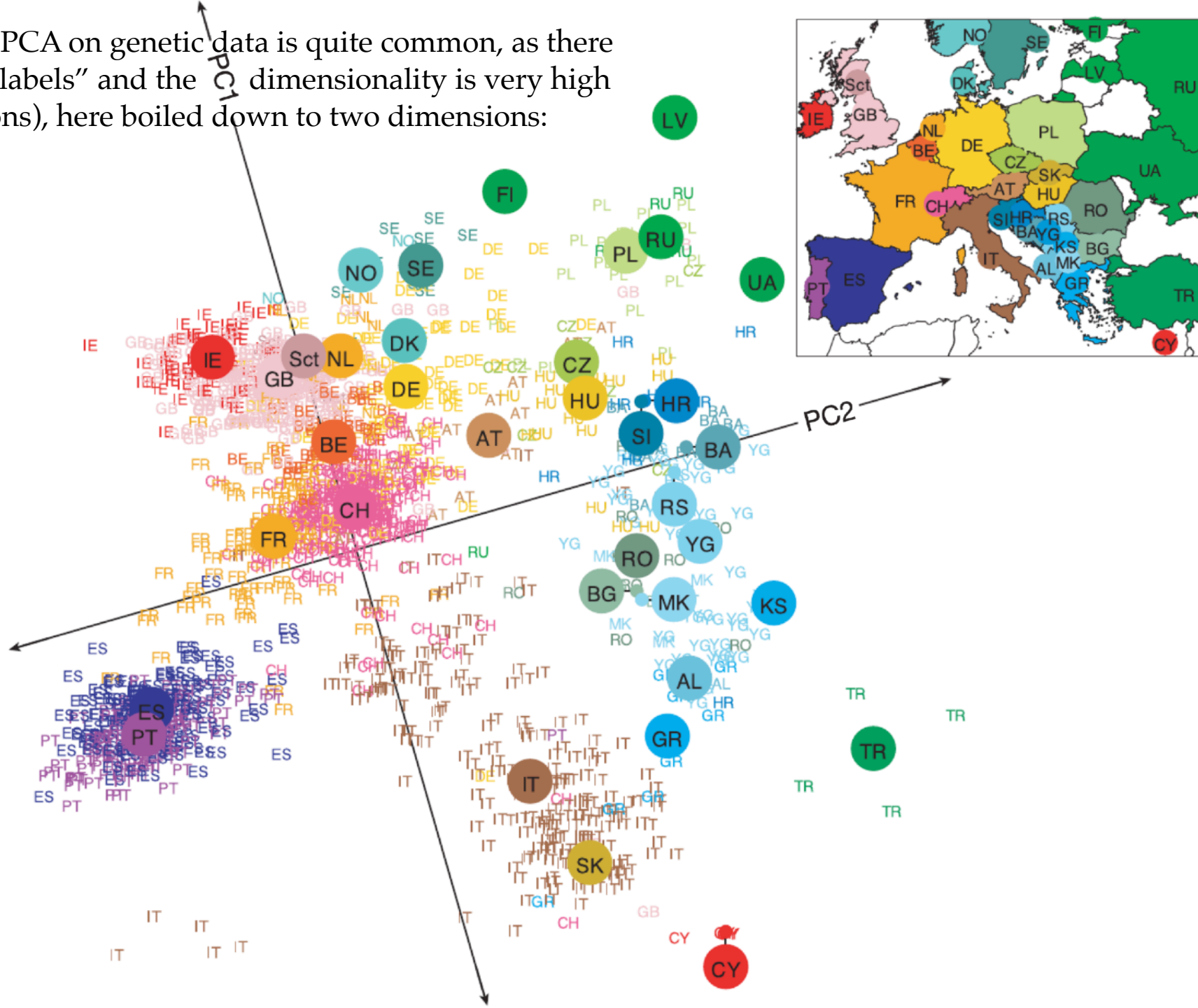
Considering the diet of UK parts, the PCA approach (again) clearly shows structure:

Alcoholic drinks
Beverages
Carcase meat
Cereals
Cheese
Confectionery
Fats and oils
Fish
Fresh fruit
Fresh potatoes
Fresh Veg
Other meat
Other Veg
Processed potatoes
Processed Veg
Soft drinks
Sugars



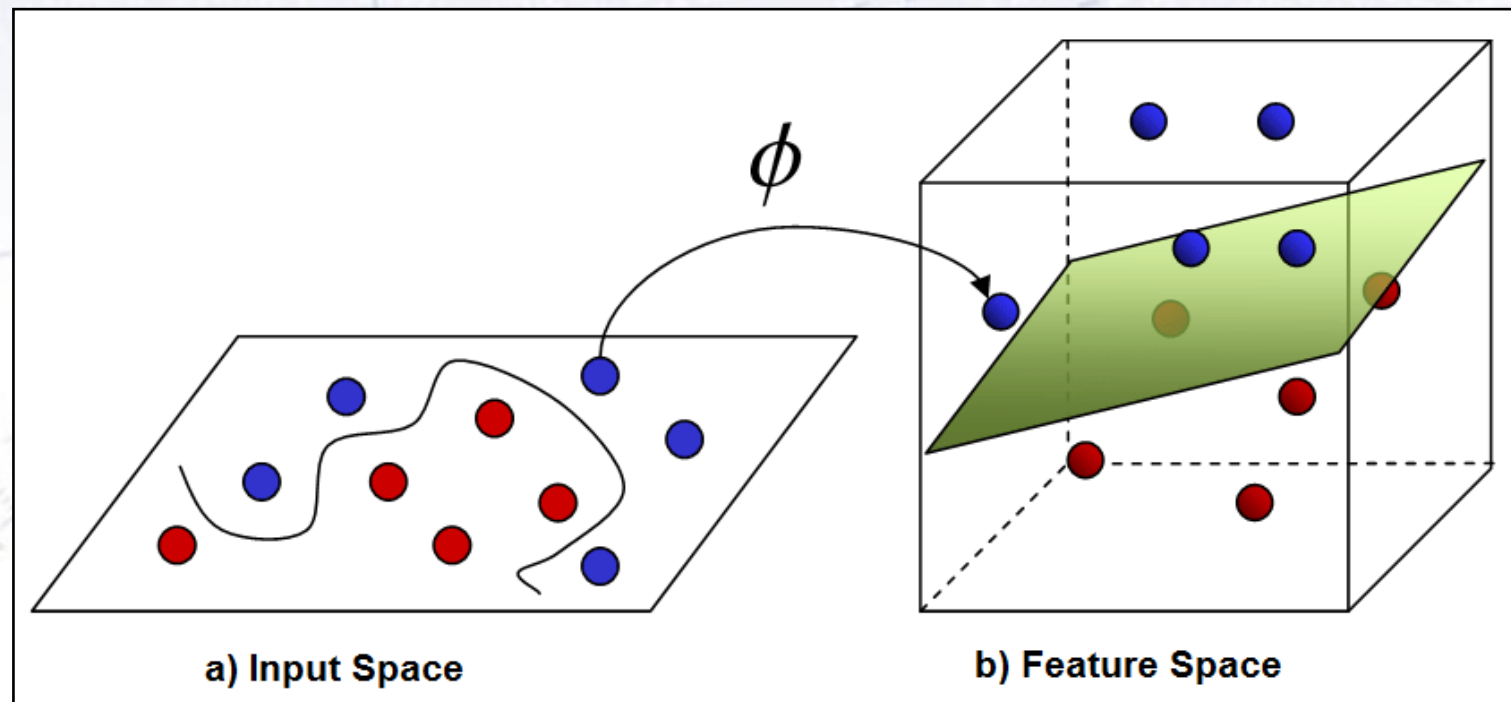
PCA example

Using PCA on genetic data is quite common, as there is no “labels” and the dimensionality is very high (millions), here boiled down to two dimensions:



Kernel PCA

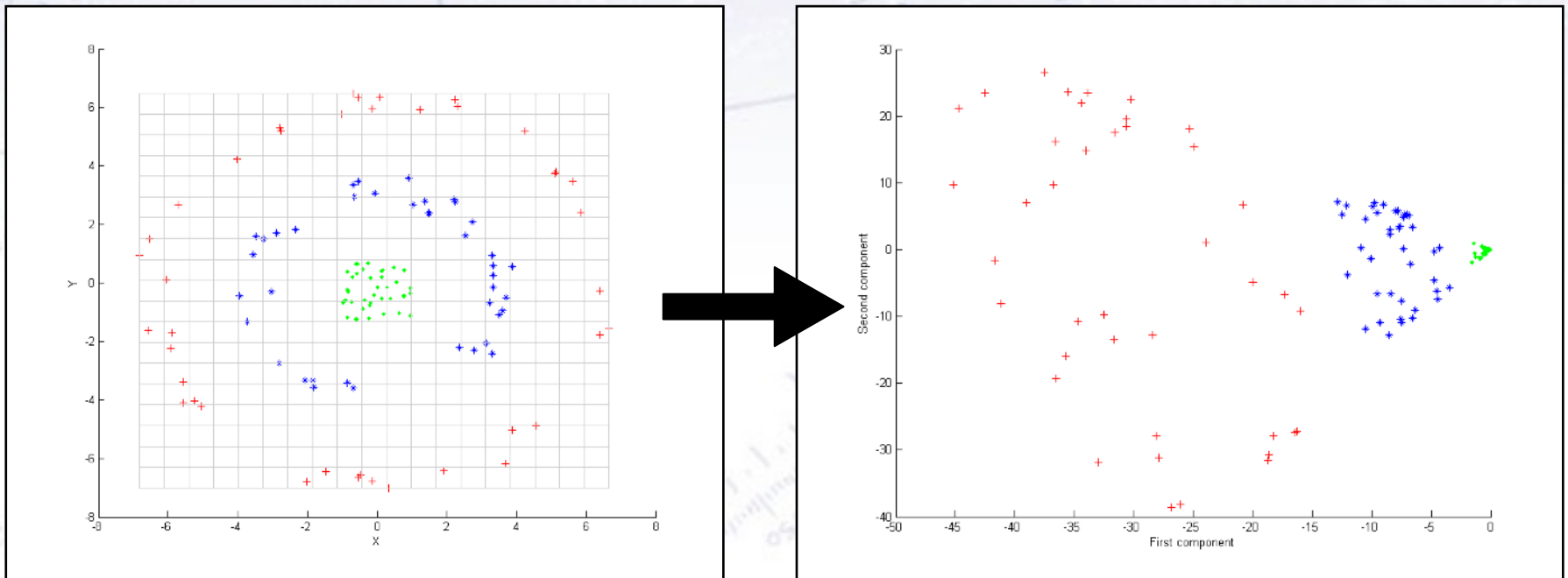
For non-linear problems, the kernel PCA might be the solution. Here, a (non-linear) kernel is applied before the PCA transformation. This is computationally heavy, but often works well as shown below:



There are other non-linear unsupervised methods, in particular t-SNE and UMAP have gained popularity from their performance.

Kernel PCA

For non-linear problems, the kernel PCA might be the solution. Here, a (non-linear) kernel is applied before the PCA transformation. This is computationally heavy, but often works well as shown below:



There are other non-linear unsupervised methods, in particular t-SNE and UMAP have gained popularity from their performance.