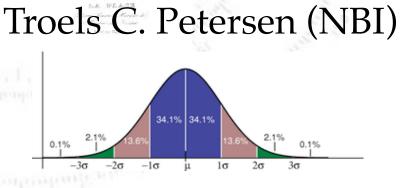
Applied ML (Kernel) Principle Component Analysis



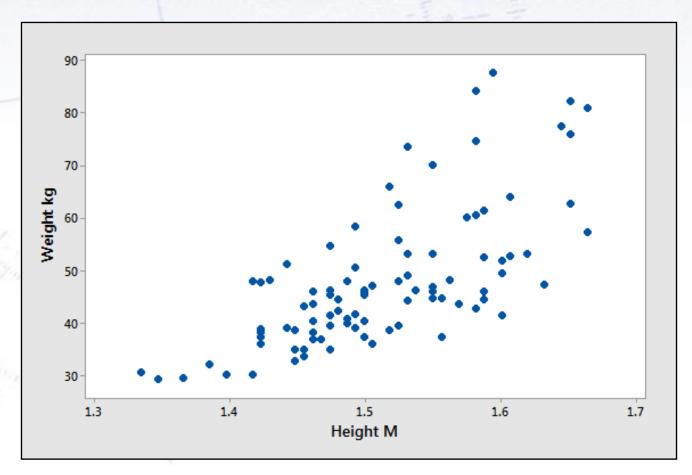


"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

PCA overview

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

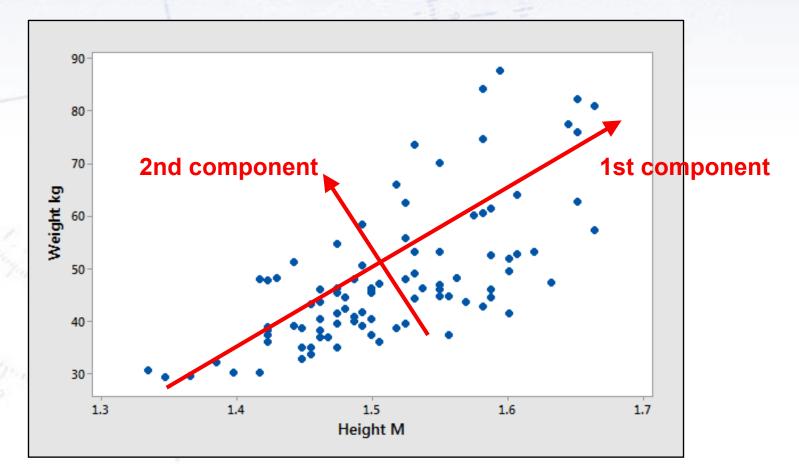
We want to apply a PCA to this data, to reduce dimensionality!



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.



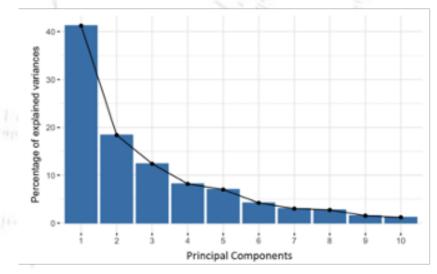
PCA overview

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:

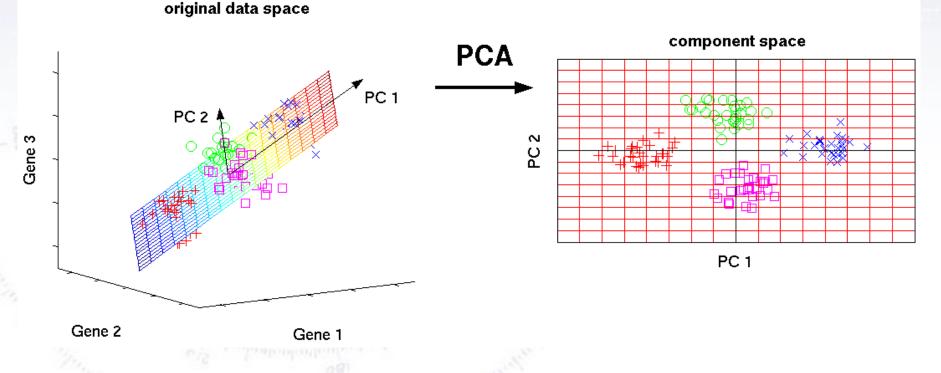
- 1. Standardize all variables
- 2. Compute covariance matrix
- 3. Compute eigenvectors and eigenvalues
- 4. Discard vectors/directions that are not important enough



$$\begin{array}{cccc} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{array}$$

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.

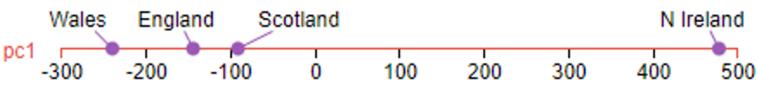


Essentially, one finds the two (orthogonal) directions, which approximates the data best, and "throw away" all the other dimensions in this new space.

17-dimensional PCA

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

	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	3
Other Veg	488	355	418	570	12
Processed potatoes	198	187	220	203	
Processed Veg	360	334	337	365	
Soft drinks	1374	1506	1572	1256	
Sugars	156	139	147	175	
Wales	England	Scotla	and		



17-dimensional PCA

100

0

200

300

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

	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	
Wales	England	Scotlan	d		

DC1

-300

-200

-100

7

N Ireland

500

400

17-dimensional PCA

Considering the diet of UK parts, the PCA approach (again) clearly shows structure: 400 300 -Wales Alcoholic drinks Beverages 200 Carcase meat Cereals N Ireland 100 -Cheese England Confectionery pc2 0 Fats and oils Fish -100 -Fresh fruit Fresh potatoes -200 -Scotland Fresh Veg Other meat -300 -Other Veg Processed potatoes -400 Processed Veg -200 -100 100 200 300 -300 400 500 0 Soft drinks pc1 Sugars England Scotland N Ireland Wales DC1

100

200

300

400

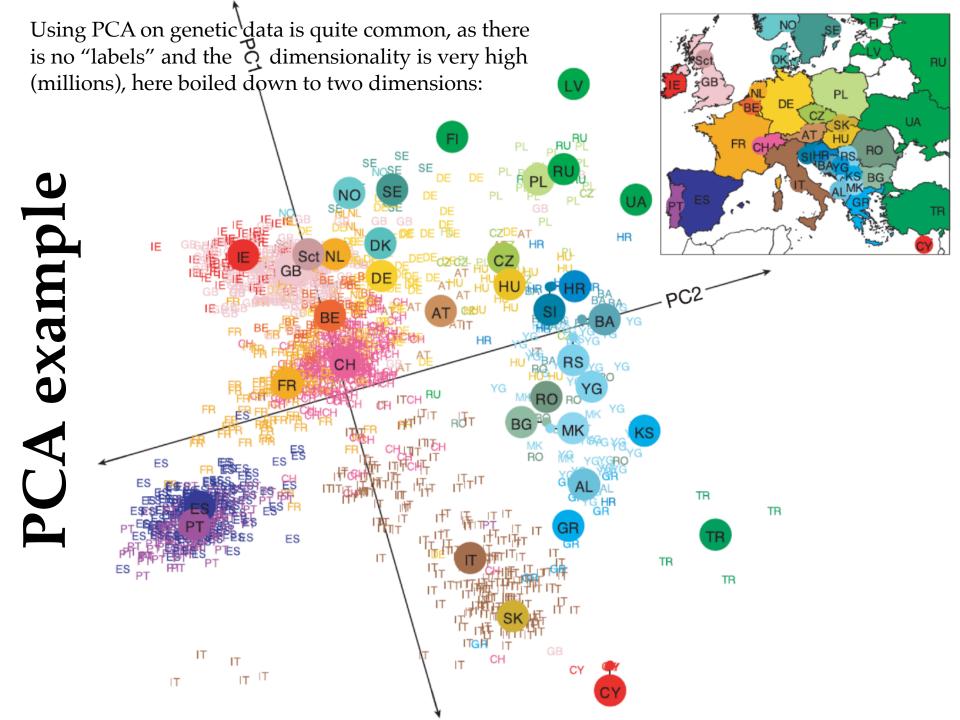
500

-200

-300

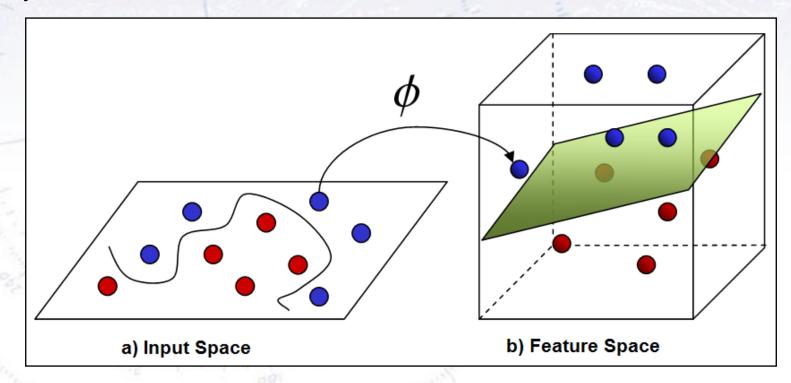
-100

0



Kernel PCA

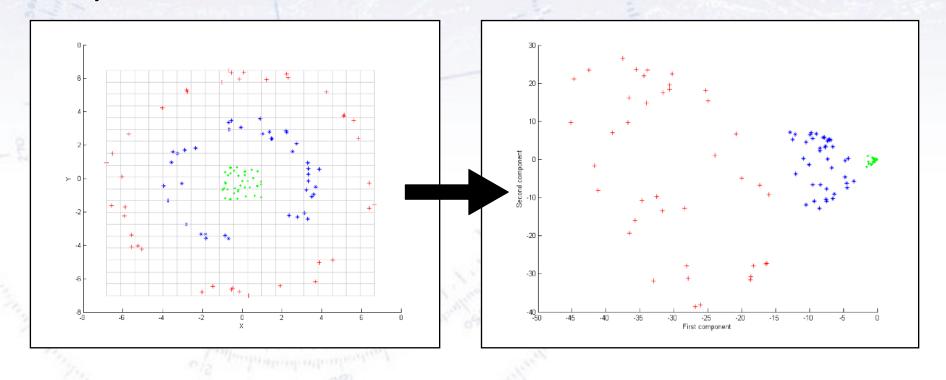
For non-linear problems, the kernel PCA might be the solution. Here, a (nonlinear) kernel is applied before the PCA transformation. This is computationally heavy, but often works well as shown below:



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 (nonlinear) 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.