"by spreading PCA, we're more likety to see patterns"

BIMM143: Introduction to Principal Component Analysis (Part 1)

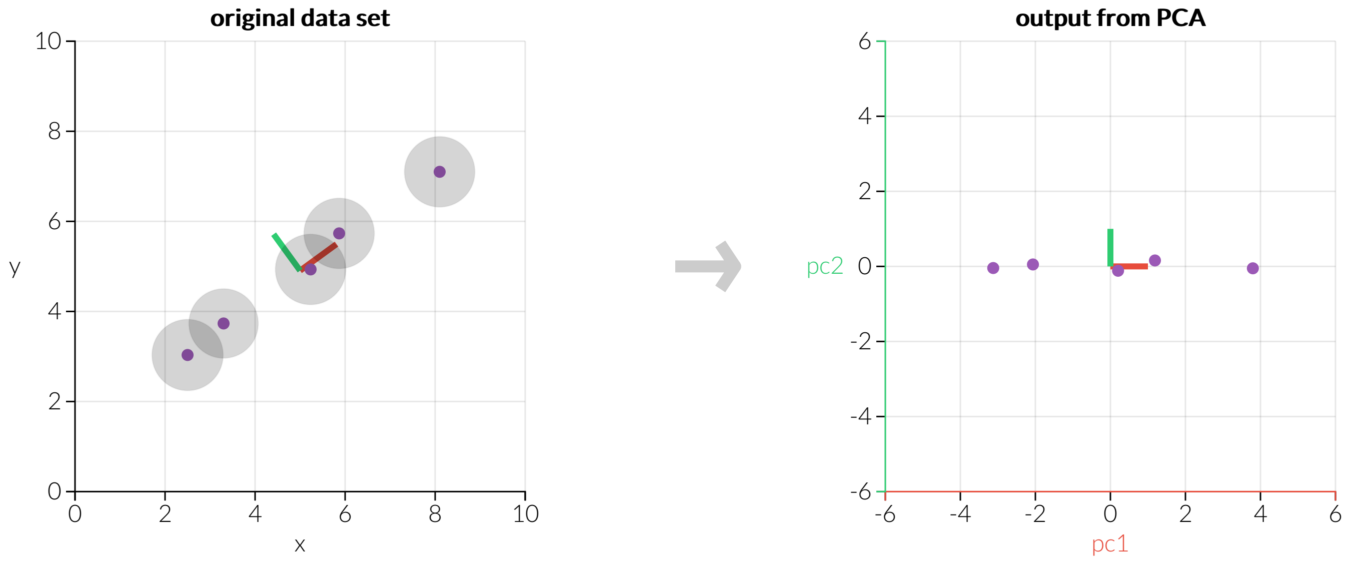
By Barry Grant

Principal component analysis (PCA) is a well established "multivariate statistical technique" used to reduce the dimensionality of a complex data set to a more manageable number (typically 2D or 3D). This method is particularly useful for highlighting strong paterns and relationships in large datasets (i.e. revealing major similarities and diferences) that are otherwise hard to visualize. As we will see again and again in this course PCA is often used to make all sorts of bioinformatics data easy to explore and visualize.

2D example

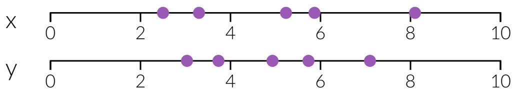
First, consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane. But if we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The axes don't actually mean anything physical; they're combinations of height and weight called "principal components" that are chosen to give one axes lots of variation.

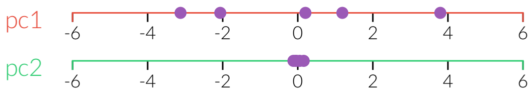
Drag the points around in the following visualization to see how the PC coordinate system (our new axis) adjusts.



PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

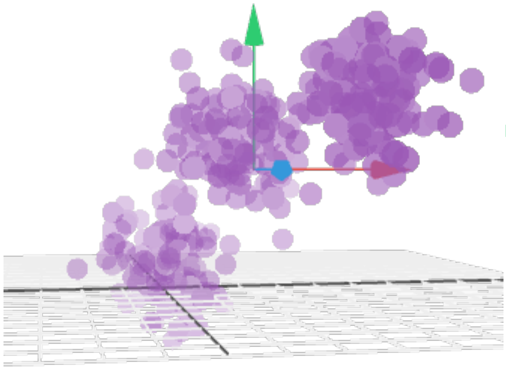
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.

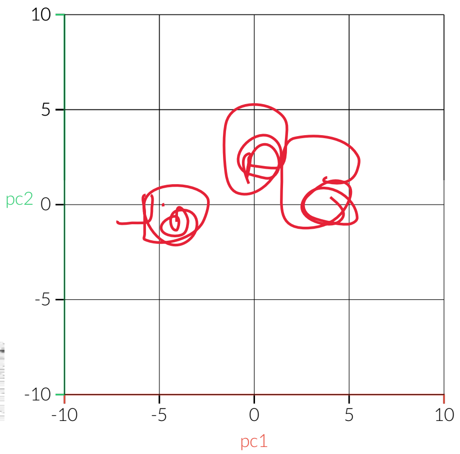


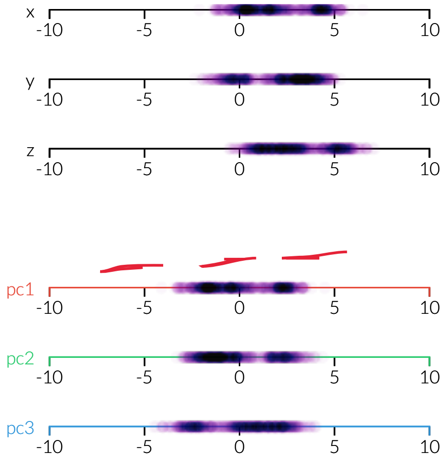


3D example

With more dimensions, PCA is more useful. Even with 3 dimensions it can be 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. To see the "official" PCA transformation, click the "Show PCA" button. The PCA transformation ensures that the horizontal axis PC1 has the most variation, the vertical axis PC2 the second-most, and a third axis PC3 the least. Obviously, PC3 is the one we drop.







show PCA reset

Eating in the UK (a 17D example)

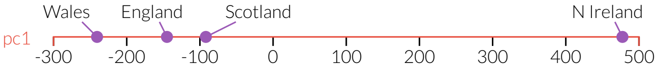
Original example from Mark Richardson's class notes Principal Component Analysis

What if our data have a few more dimensions? Like, **17** dimensions?! In the table is the average consumption of 17 types of food in grams per person per week for every country in the UK. We will learn more about this data and analyze it ourselves in R in Part 2.

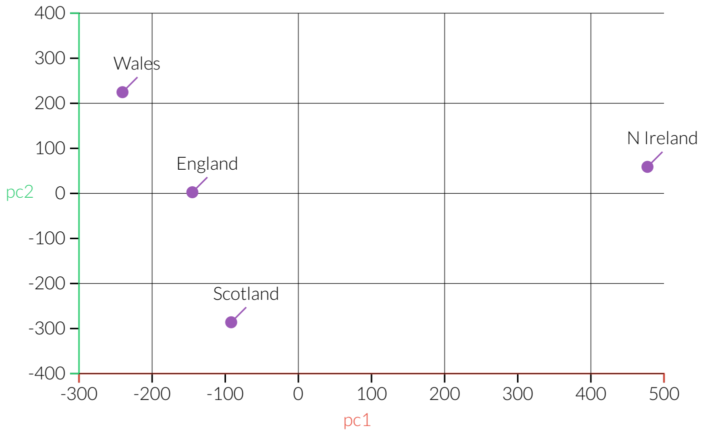
The table shows some interesting variations across different food types, but overall differences aren't so notable even for this relatively small 17 dimensional dataset. Let's see if PCA can eliminate dimensions to more clearly highlight how countries differ.

	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

Here's the plot of the data along the first principal component. Already we can see something is different about Northern Ireland.



Now, see the first and second principal components, we see Northern Ireland a major outlier. Once we go back and look at the data in the table, this makes sense: the Northern Irish eat way more grams of fresh potatoes and way fewer of fresh fruits, cheese, fish and alcoholic drinks. It's a good sign that structure we've visualized reflects a big fact of real-world geography: Northern Ireland is the only of the four countries not on the island of Great Britain. (If you're confused about the differences among England, the UK and Great Britain, see: this video.)



This is the end of Part 1. Next we will continue our introduction to PCA with some lecture material from Barry and then in Part 2 we will examine some real life multivariate data in order to explain, in simple terms, what PCA achieves. We will perform a principal component analysis of several different data sets including RNA-Seq and proteomics data and examine the results.