Dimensionality Reduction - Principal Component Analysis

In Machine learning, we are always interested in finding major patterns in the dataset. There are several methods to do it, one of which is using Principal Component Analysis (PCA).

Consider a situation where you have a large number of variables (measurements) in the dataset, not all variables are useful in finding the patterns in the data, and maybe some variables are correlated (a lot of covariance between the variables) to each other and capture the same pattern, in such situations PCA comes very handy, as it tries to find major patterns in the dataset using the correlation between the variables and reduce the number of variables required to get insights from the data without losing much information.

This method of reducing the number of variables to get insights without losing much information is called **Dimensionality Reduction**.

Let's start with an example:
See Figure 1, this image is the matrix of ratings given by 6 people to 4 different holiday destinations.

From the matrix, we can observe that Anne likes spas in the Alps, Bill likes Hawaii, Maggie Loves Himalayas, and Jen enjoys Scuba diving.
Here, people Anne, Bill, Chris... are the observations, while holiday destination is our features.

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.

	APLPINE SPA	HIMALAYAS	HAWAII	SCUBA
ANNE	20	4	16	0
BILL	10	2	18	10
CHRIS	13	1	19	7
JEN	1	13	7	19
JOE	18	10	10	2
MAGGIE	4	20	0	16

Figure 1

Now, as we have the data, let's find out if there is any hidden pattern in the dataset, and can we represent each person(observation) as a combination of these patterns?

	ALPINE SPA	HIMALAYAS	HAWAII	SCUBA
PATTERN 1	-1	+1	-1	+1
PATTERN 2	+1	+1	-1	-1

Figure 2

Consider Figure 2, we can see that 2 patterns are emerging from it.

Pattern 1: -1, +1, -1, +1

In pattern 1 we have +1 for Himalayas and Scuba, maybe this pattern represents the *amount of adventure* for each place.

Pattern 2: +1, +1, -1, -1

In pattern 2 we have +1 for Alpine Spa and Himalayas, maybe this pattern represents the liking for the mountains.

These patterns are called **principal components** and can be represented as a vector. In the above example, we have 2 principal components, ordered in terms of importance

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.

and are slightly rounded. This importance is decided based on the eigenvectors, which we will cover in the next video.

Now, we can represent the rating for each person using these 2 principal components and a mean rating.

Mean rating for 4 holiday destinations are:

1. Alpine spa: 20 + 10 + 13+ 1+ 18 + 4 = 66
    a. Average rating= 66/6 = 11.

Similarly, the Average ratings for Hawaii, Himalayas, and Scuba are 8.3,11.7, and 9 respectively.

We can represent Anne's rating as:

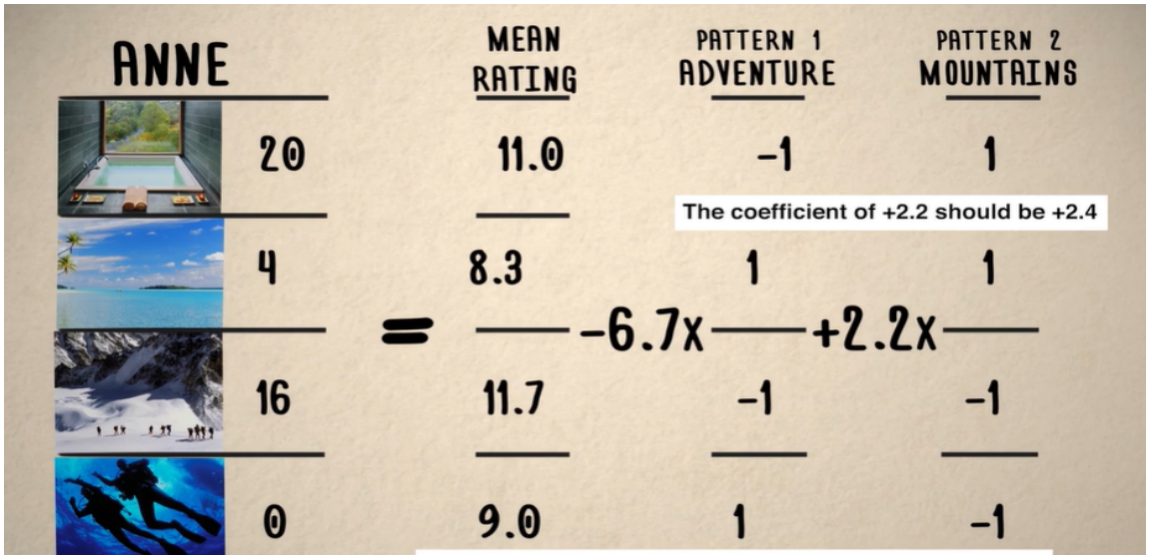


Figure 3

To find the values -6.7 and 2.4,
We have to solve for equations:

20 = 11 + a*(-1) + b*(1)
4 = 8.3 + a*(1) + b*(1)
16 = 11.7 + a*(-1) + b*(-1)
20 = 9 + a*(1) + b*(-1)

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Solving for a and b, we will get, a= -6.7 and b=2.4.

We can do this for each person and get the values. The conclusion from this is, that we are now able to represent the rating given by each person using only 2 factors instead of 4 and we can also comment on the destination preference of the people.
This is called **Principal component analysis.**

One of the other applications for PCA is, as now we have found *hidden patterns* in the dataset and reduced the dimensions, we can now be able to *visualize the results.*
See the Figure 4:

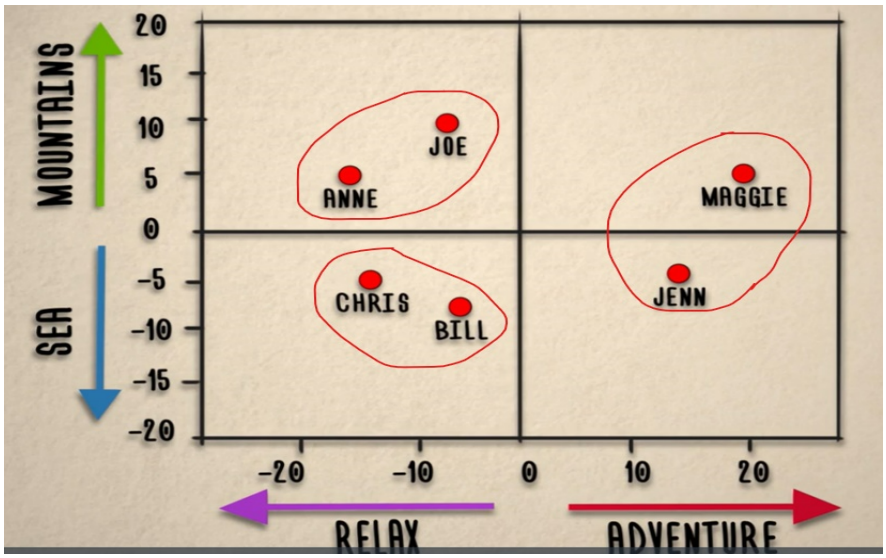


Figure 4

We can see natural clusters emerge when we apply PCA to the dataset and reduce the dimensions to 2. Anne and Joe can be in 1 cluster and like mountains, whereas Jenn and Maggie like Adventure more.

Therefore, **PCA can be used for 2 things:**

1. Reduce Dimensions of the data without losing much information.
2. To visualize the hidden patterns in the dataset.

Let's consider a few more examples in brief:

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.

As we saw in the previous example, how can you set the rating of people as a linear combination of 2 principal components, in a similar fashion you can represent images using a linear combination of eigenfaces.

We can represent the Face images using vector pixels (see the images below), where each column is stacked over one another.

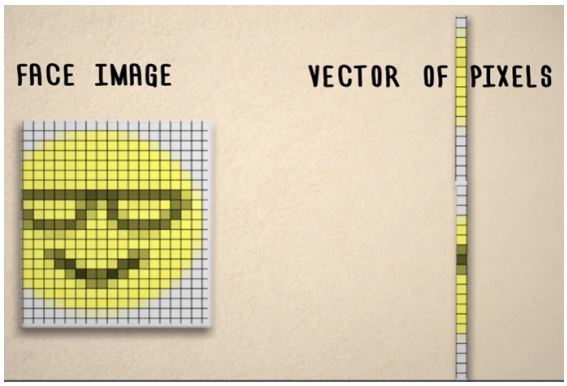
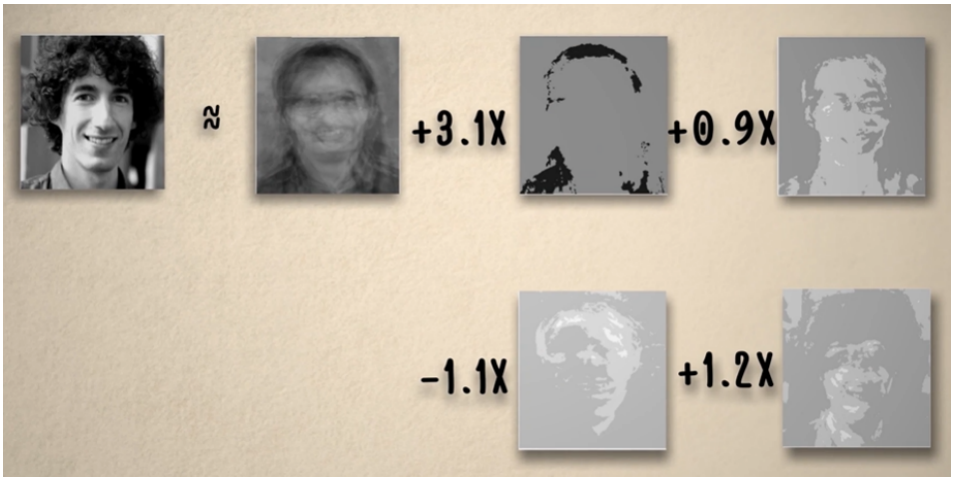


Figure 6

Figure 6

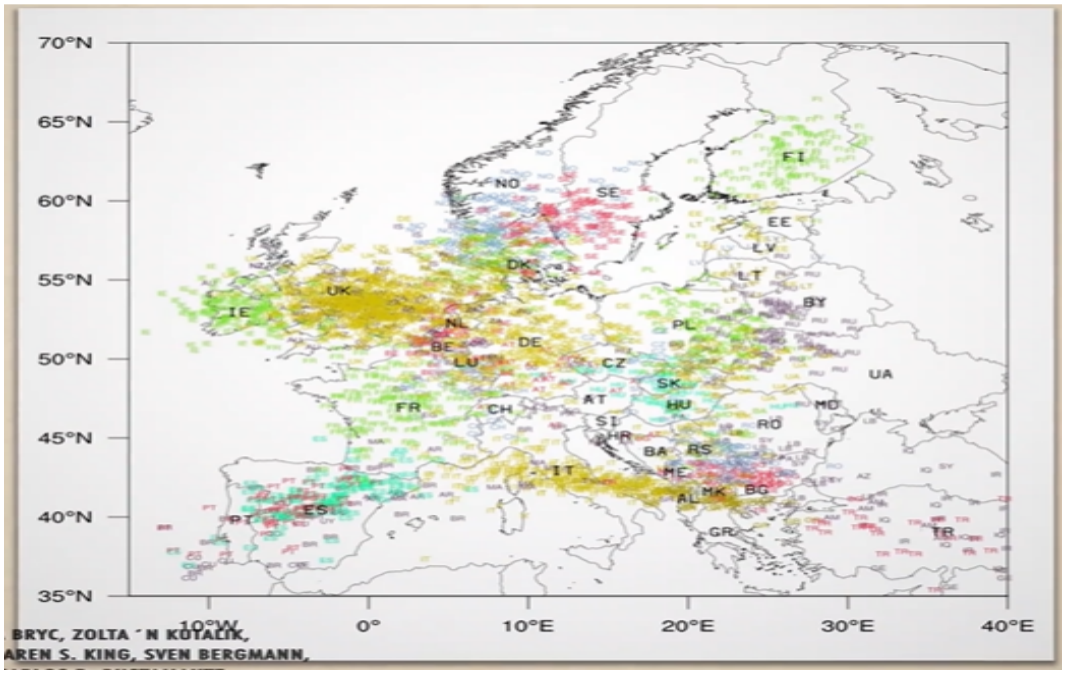
Using this vector we can create **eigenfaces**, which are a form of ghostly representation of the original image. You can think of these eigenfaces as principal components, this means we can represent these images using a linear combination of these eigenfaces.



Our second example is the study of genetics, here we try to answer the question, can we see someone's origin from the DNA. For this study, researchers collected data from

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.

1400 Europeans, each person is described by his or her genetic variations, which means we have around 200,000 features. Then researchers used PCA and reduced these dimensions to only 2 and plotted the results of 2 Principal components on the graph of Europe.



Each data point here is a person, and we can clearly observe clusters emerging in the data, with only 2 principal components.

We can conclude that the principal genetic variations in Europe are highly correlated to geography, and we found this just using PCA.

This file is meant for personal use by rswetland@outlook.com only.
Sharing or publishing the contents in part or full is liable for legal action.