Introduction to Data Science - Assignment 3

Shahriyar Mahdi Robbani

February 29, 2020

Exercise 1

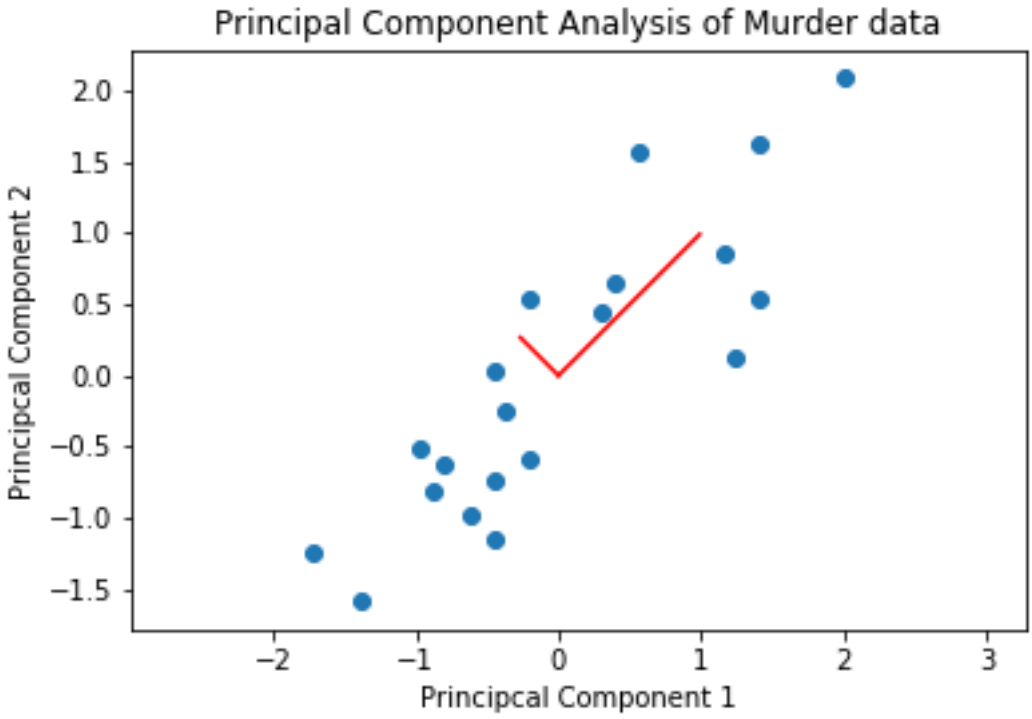

Figure 1: Principal Component Analysis of Murder data

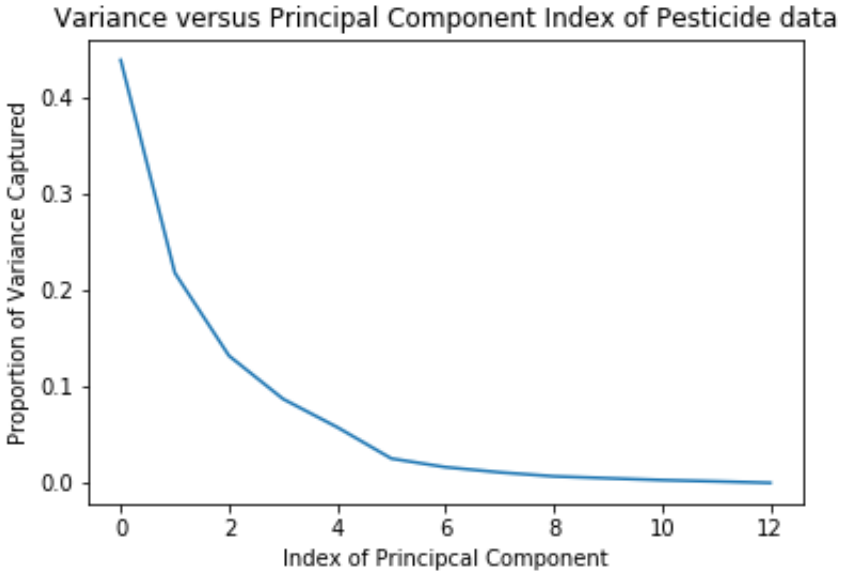

Figure 2: Variance versus Principal Component Index of Pesticide data

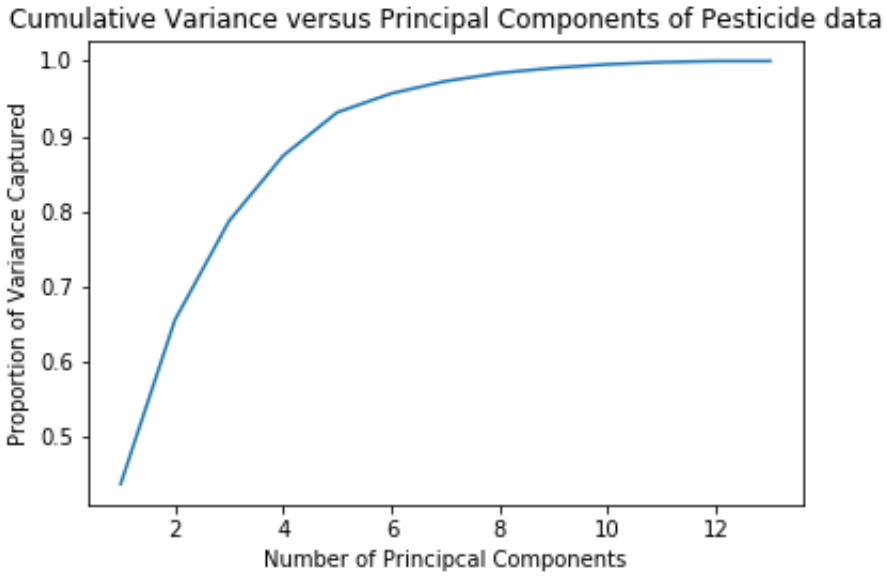

Figure 3: Cumulative Variance versus Principal Components of Pesticide data

```
5 PCs are needed to capture 90.0% of the variance
6 PCs are needed to capture 95.0% of the variance
```

Exercise 2

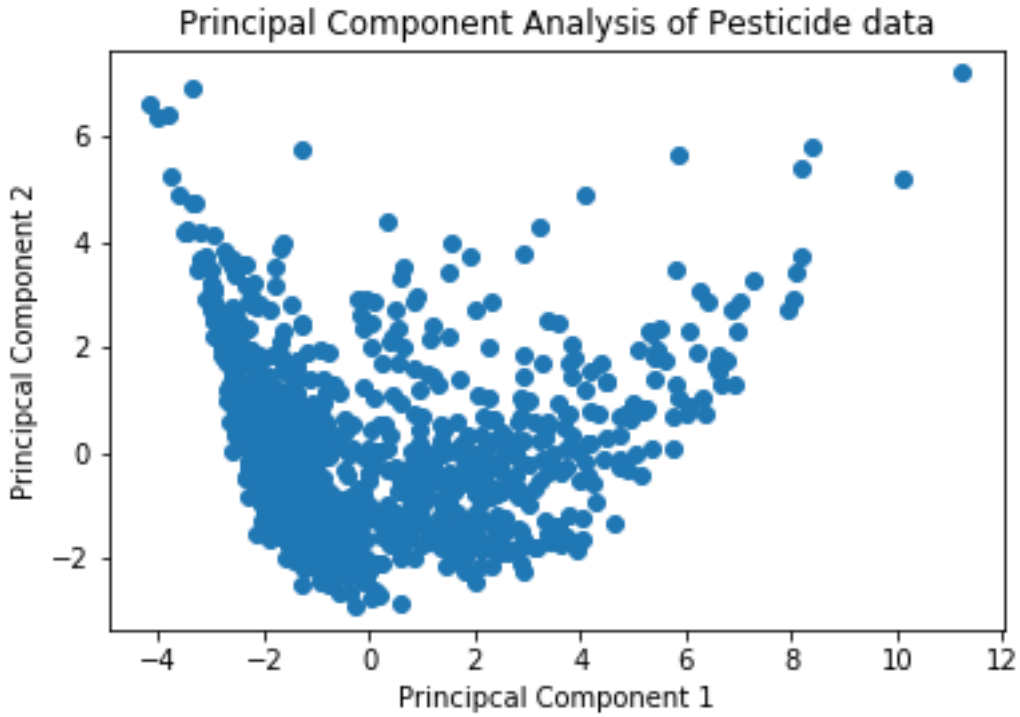

Figure 4: Principal Component Analysis of Pesticide data

Exercise 3

```
Centroid 1:
[ 0.10697804  0.15743643  0.24443175  0.42630232  0.36461238 -0.28733807
 -0.48392646 -0.52535925 -0.49566457 -0.43337993 -0.37703005 -0.26288248
 -0.17762998]
Centroid 2:
[-0.26064409 -0.38358222 -0.59553991 -1.03865412 -0.88835112  0.70007797
  1.17905107  1.27999901  1.20765011  1.05589819  0.91860586  0.64049373
  0.43278231]
```

The function creates a list of distances from each data point to a centroid for k centroids and converts this to a matrix. It then finds the lowest distance for each centroid, and assigns the data point for the corresponding distance to a cluster to make k clusters. The new centroids are then obtained by finding the mean of each cluster. The loss is also calculated for all clusters. This cycle is repeated until the loss in one cycle is the same as the previous cycle, indicating no change in centroids.

Exercise 4

How is probability interpreted differently in the frequentist and Bayesian views?

In the frequentist view, the probability of an event is essentially the frequency of that event occuring after many trials. The Bayesian view of probability is a measure of the belief of that event occurring.

Cheap, efficient computers played a major role in making Bayesian methods mainstream. Why?

Bayesian methods require large amounts of sampling which was difficult to do before cheap and efficient computers became mainstream.

What is the difference between a Bayesian credible interval and a frequentist confidence interval?

A 95% frequentist confidence interval means the confidence interval will contain the true value 95 out of 100 times. The Bayesian credible interval indicates that the data will be in the interval 95% of the time

How does a maximum likelihood estimate approximate full Bayesian inference?

skagsdhlsadjk

When will point estimates be a good approximation of full Bayesian inference?

Point estimates are a good approximation when the distribution is unimodal and has a narrow peak.