Introduction to Data Science - Assignment 3

Shahriyar Mahdi Robbani

March 4, 2020

Exercise 1

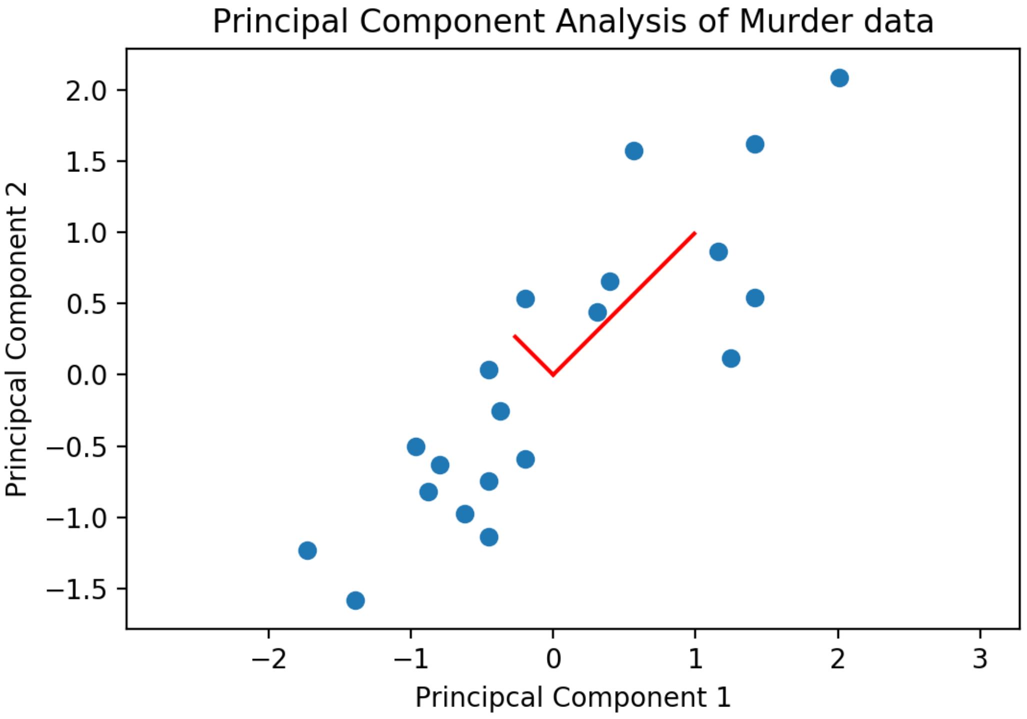

Figure 1: Principal Component Analysis of Murder data

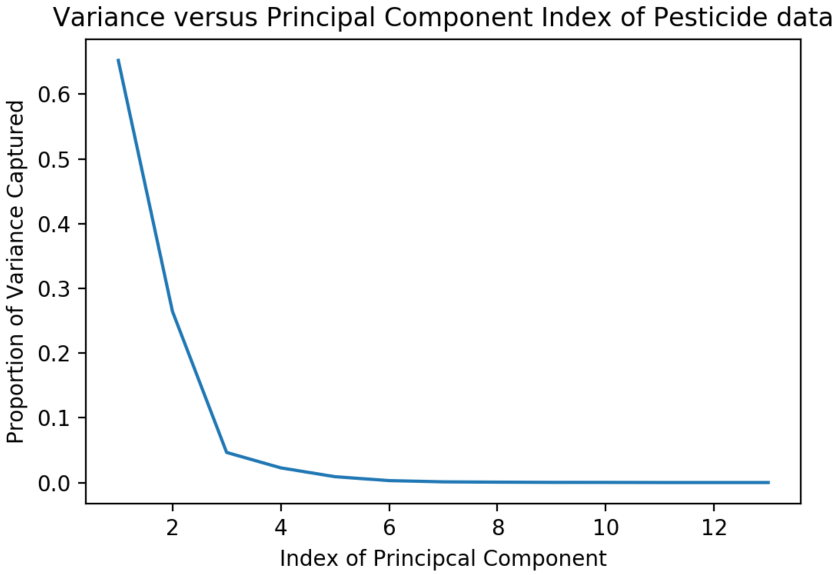

Figure 2: Variance versus Principal Component Index of Pesticide data

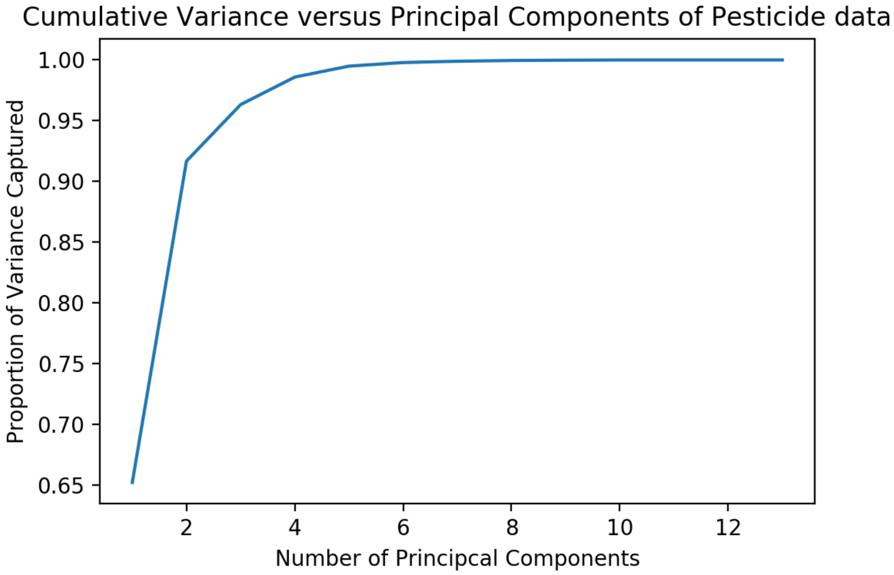

Figure 3: Cumulative Variance versus Principal Components of Pesticide data

```
2 PCs are needed to capture 90.0% of the variance
3 PCs are needed to capture 95.0% of the variance
```

Exercise 2

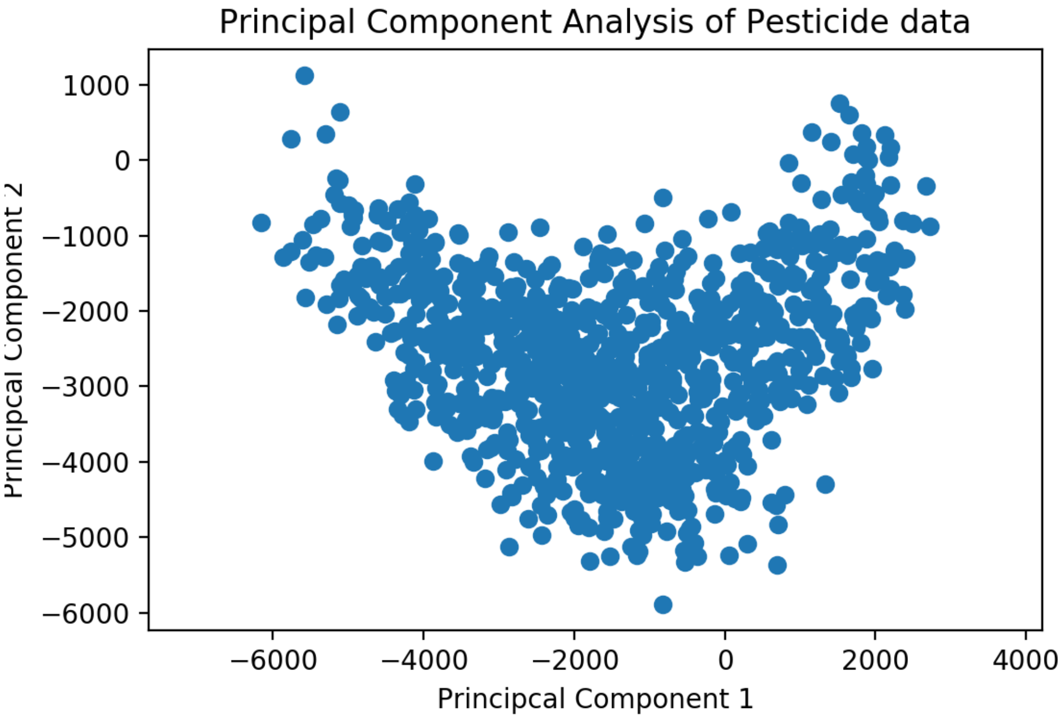

Figure 4: Principal Component Analysis of Pesticide data

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 calculates a list of distances between each data point and the centroid for k centroids, and converts these lists into a matrix. It then finds which data point has the smallest distance to a centroid, and then assigns that data point to a cluster. This results in k clusters. The new centroids are then calculated by finding the mean position of each cluster. The loss is defined as the sum of squared distances between each point and the corresponding centroids and is used as a measure of how good the centroids are. The algorithm continues looping until the loss converges, indicating no change in the centroids.

Exercise 4

1. **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 occurring after many trials. The Bayesian view of probability is a measure of the belief of that event occurring.

2. **Cheap, efficient computers played a major role in making Bayesian methods mainstream. Why?** Bayesian methods involve calculating the posterior probability, and this usually does not have a closed form equation. Therefore large amounts of sampling is required to approximate it which was difficult to do before cheap and efficient computers became mainstream.

3. **What is the difference between a Bayesian credible interval and a frequentist confidence interval?** A 95% frequentist confidence interval means the confidence interval will capture the true value 95 out of 100 times. The Bayesian credible interval indicates that the parameter has a 95% chance of being in the interval.

4. **How does a maximum likelihood estimate approximate full Bayesian inference?** The Bayesian posterior is obtained from the product of the likelihood and the prior. If a uniform prior is used, the posterior then only depends on the likelihood. So given a large enough dataset, the maximum likelihood estimate corresponds to an estimate of the posterior.

5. **When will point estimates be a good approximation of full Bayesian inference?** Point estimates are a good approximation when the posterior distribution is unimodal and has a narrow peak.