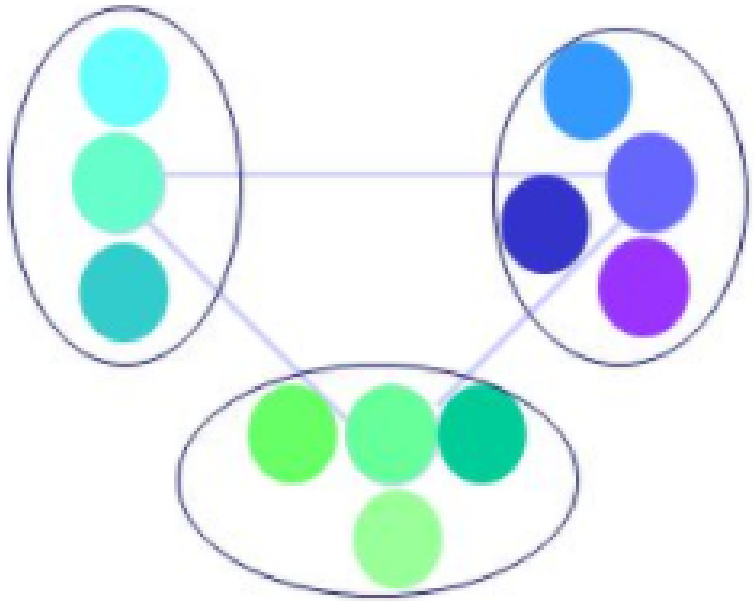
STAT3612 Statistical Machine Learning (Lecture 12)

Unsupervised Learning

Dr. Aijun Zhang
The University of Hong Kong

24 November 2020

Table of Contents



Why Unsupervised Learning?

- When there are no explicit target outputs/responses ...
- To learn the input patterns that reflect the statistical structure, e.g.
  - grouping or clustering
  - mixture distribution
  - association
- Applications of unsupervised learning in
  - customer segmentation
  - anomaly detection
  - dimension reduction

1. Clustering

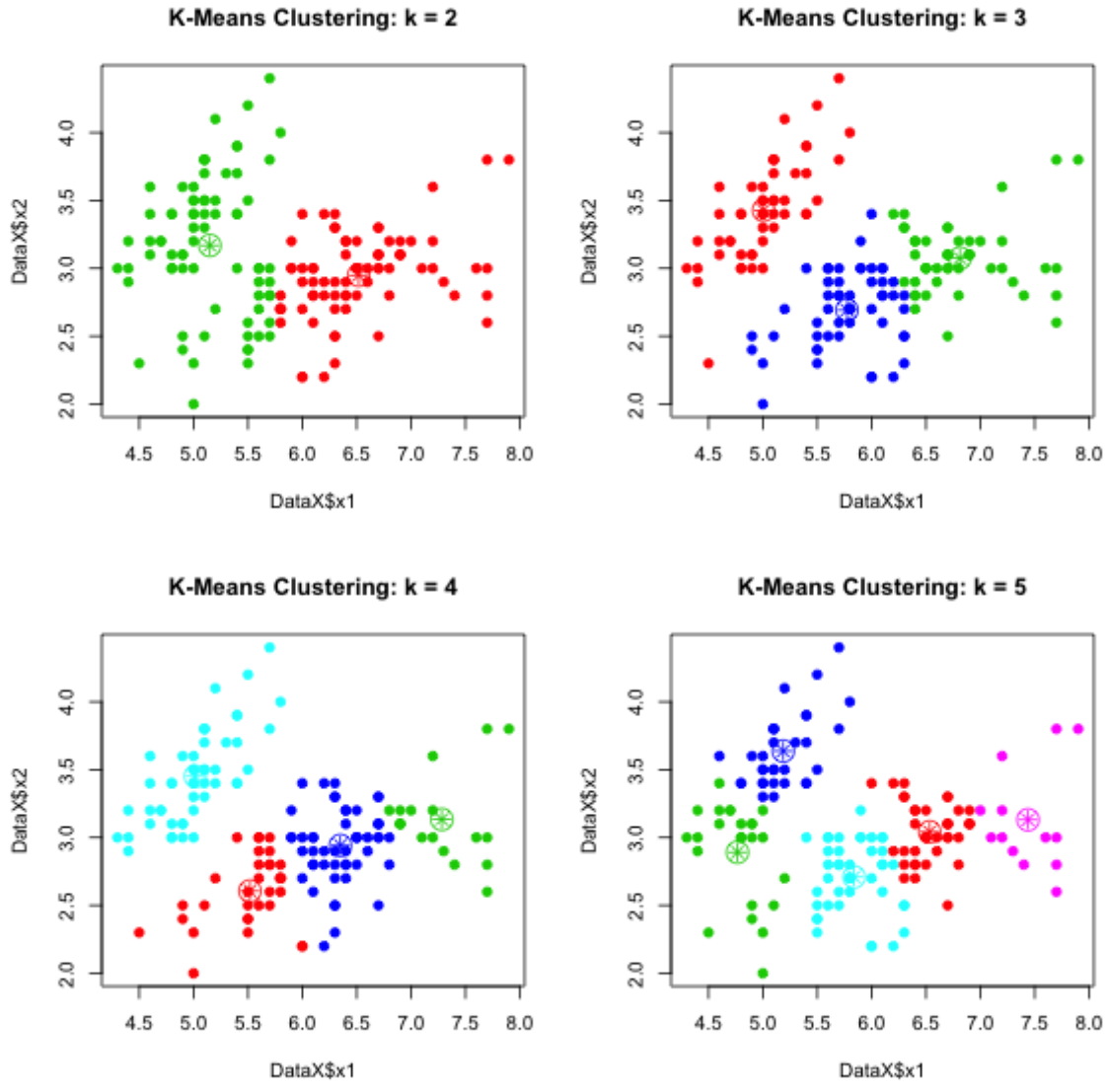
K-means clustering

1. Define the k centroids. Initialize them at random.

2. Find the closest centroid and update cluster assignments. Assign each data point to one of the k clusters. Each data point is assigned to the nearest centroid's cluster. (often Euclidean distance)

3. Move the centroids to the center of their clusters. The new position of each centroid is calculated as the average position of all the points in its cluster.

Keep repeating steps 2 and 3 until the centroid stop moving a lot at each iteration (i.e., until the algorithm converges).

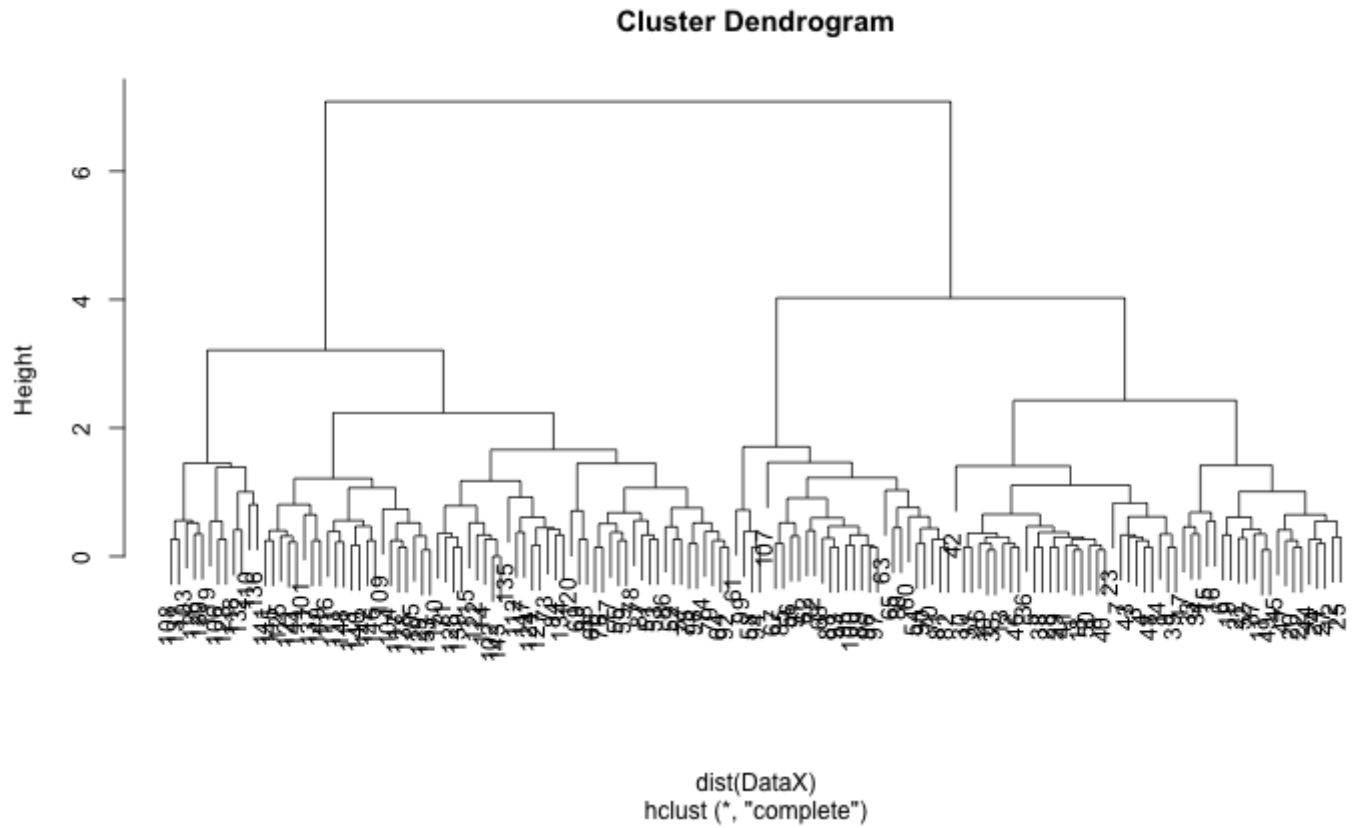

K-Means Clustering: k = 2
K-Means Clustering: k = 3
K-Means Clustering: k = 4
K-Means Clustering: k = 5
DataX$x1
DataX$x2
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

R code:

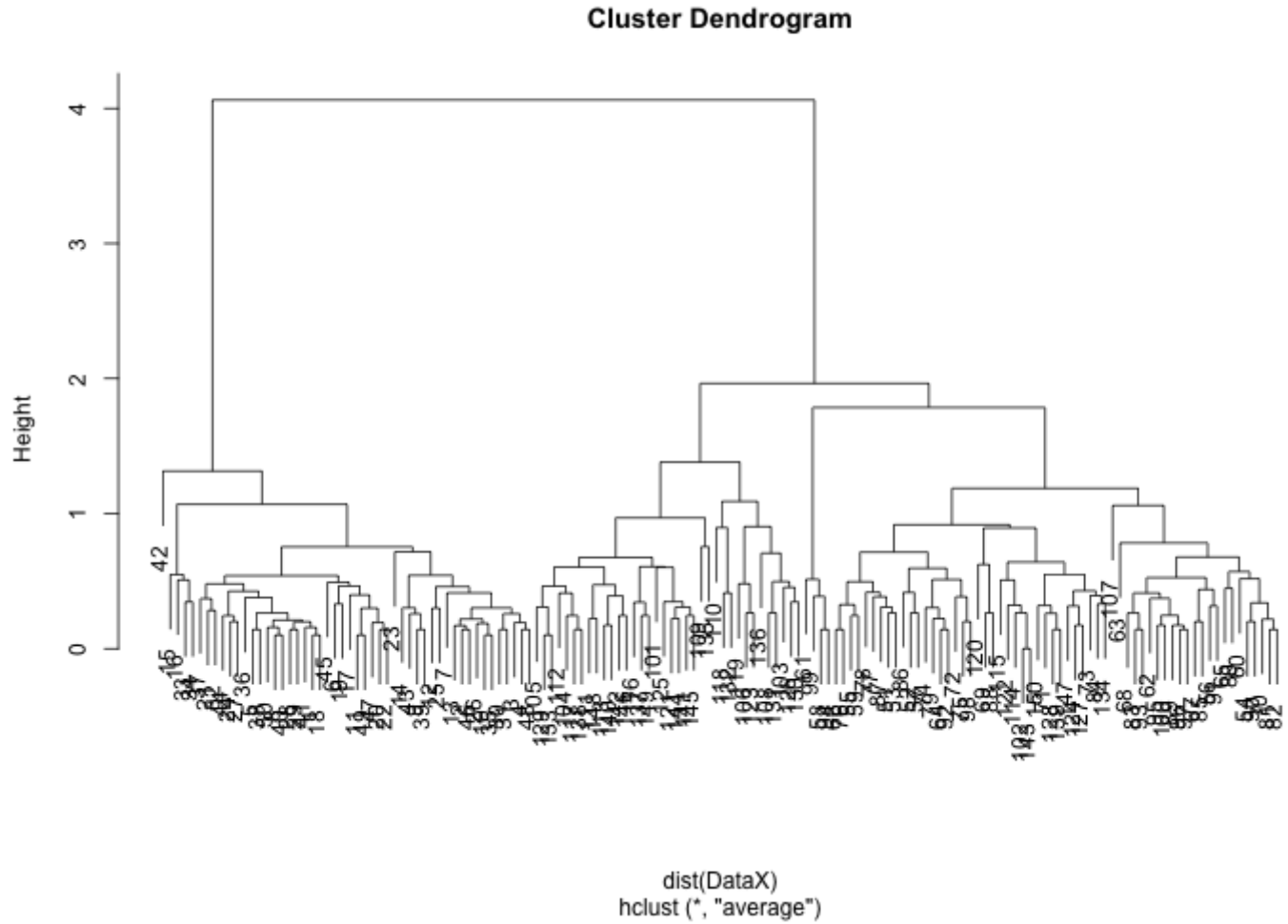
```
DataX = data.frame(x1 = iris$Sepal.Length, x2 = iris$Sepal.Width)
set.seed(9999)
par(mfrow=c(2,2))
for (kk in 2:5) {
  yfit_kmeans = kmeans(DataX, kk)
  plot(DataX$x1, DataX$x2, pch=19, cex=1, col=seq(2,1+kk)[yfit_kmeans$cluster])
  points(yfit_kmeans$centers, pch=8, cex=1.4, col=seq(2,1+kk))
  points(yfit_kmeans$centers, pch=21, cex=2.5, col=seq(2,1+kk))
  title(main=paste("K-Means Clustering: k =", kk))
  }
```

Hierarchical clustering

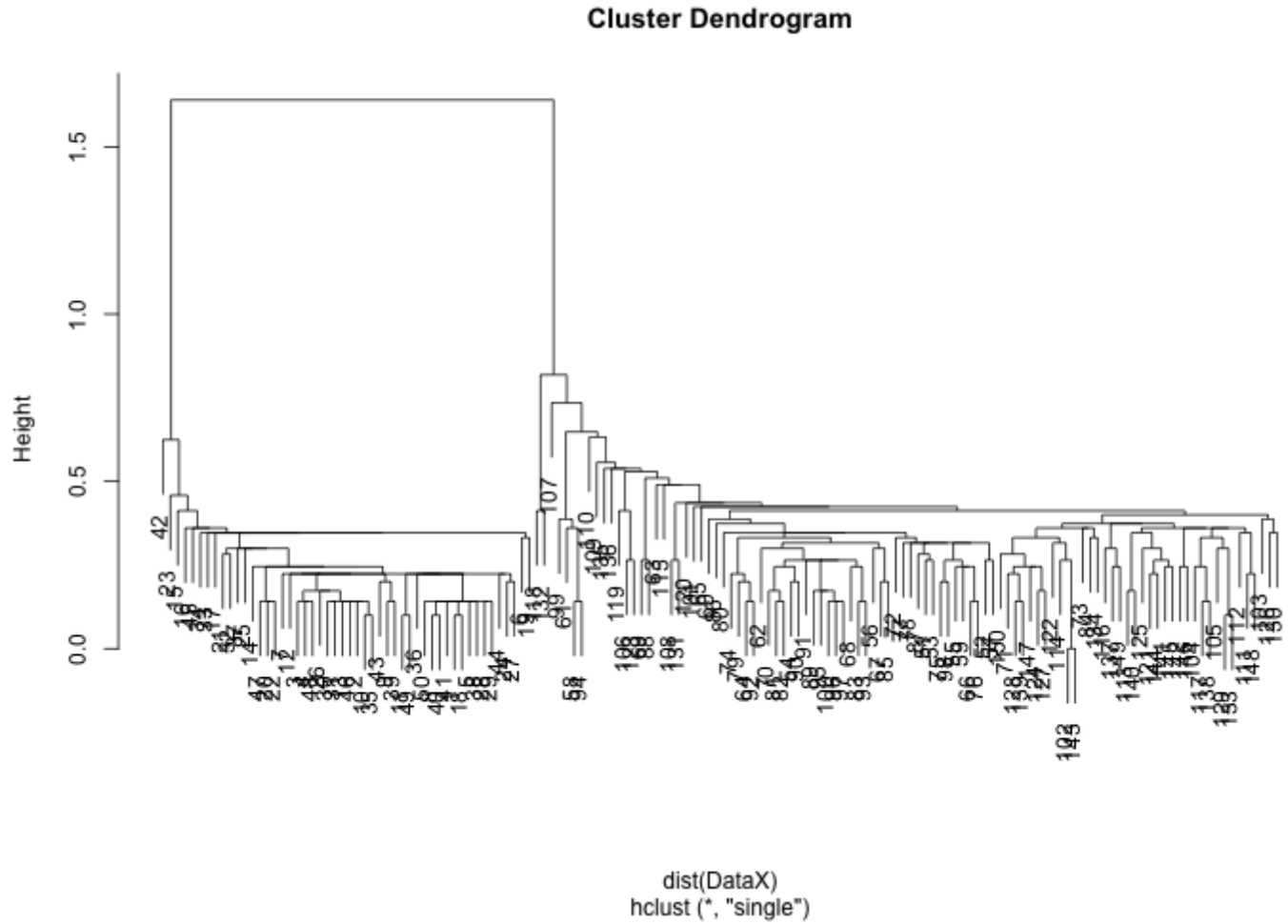
```
DataX = iris[,-5]
hc = hclust(dist(DataX), method="complete")
plot(hc)
```



```
hc = hclust(dist(DataX), method="average")
plot(hc)
```



```
hc = hclust(dist(DataX), method="single")
plot(hc)
```



2. Principal Component Analysis

Principal Component Analysis

PCA is to project the data to a new coordinate system such that the greatest variance lies on the first coordinate (i.e. the first principal component), the second greatest variance on the second principal component, and so so.

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\} = \arg\max_{\|\mathbf{w}\|=1} \left\{\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}\right\}$$

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^{\mathrm{T}}$$

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{\|\hat{\mathbf{X}}_k\mathbf{w}\|^2\right\} = \arg\max \left\{\frac{\mathbf{w}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{X}}_k\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right\}$$

Eigenvector and Eigenvalues

- Denote the $p \times p$ covariance matrix $\mathbf{C}$ of the centered data $\mathbf{X}$, i.e. $\mathbf{C} = \mathbf{X}^T\mathbf{X}/(n-1)$.
- Eigen-decomposition:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

  where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots)$ consisting of the eigenvalues in the decreasing order
- Then, the principal directions are given by the eigenvectors, i.e., columns of $\mathbf{V}$
- The principal components (also know PC scores) are given by the transformed variables, i.e. columns of $\mathbf{XV}$

Singular Value Decomposition

- Perform the SVD of the centered data $\mathbf{X}$, i.e.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

  where $\mathbf{S}$ consists of singular values $\text{diag}(s_1, s_2, \ldots)$

- It is easy to verify that

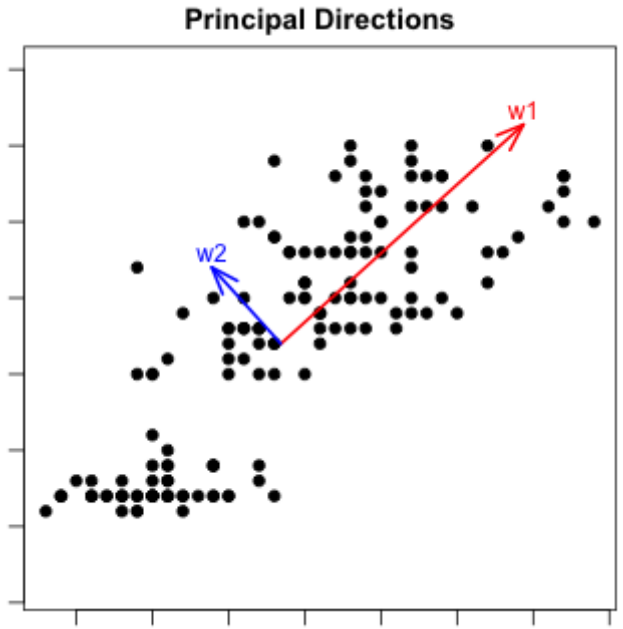
$$\mathbf{C} = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T$$

- Thus, $\lambda_i = s_i^2/(n-1)$. Principal directions are given by $\mathbf{V}$. The principal components are given by

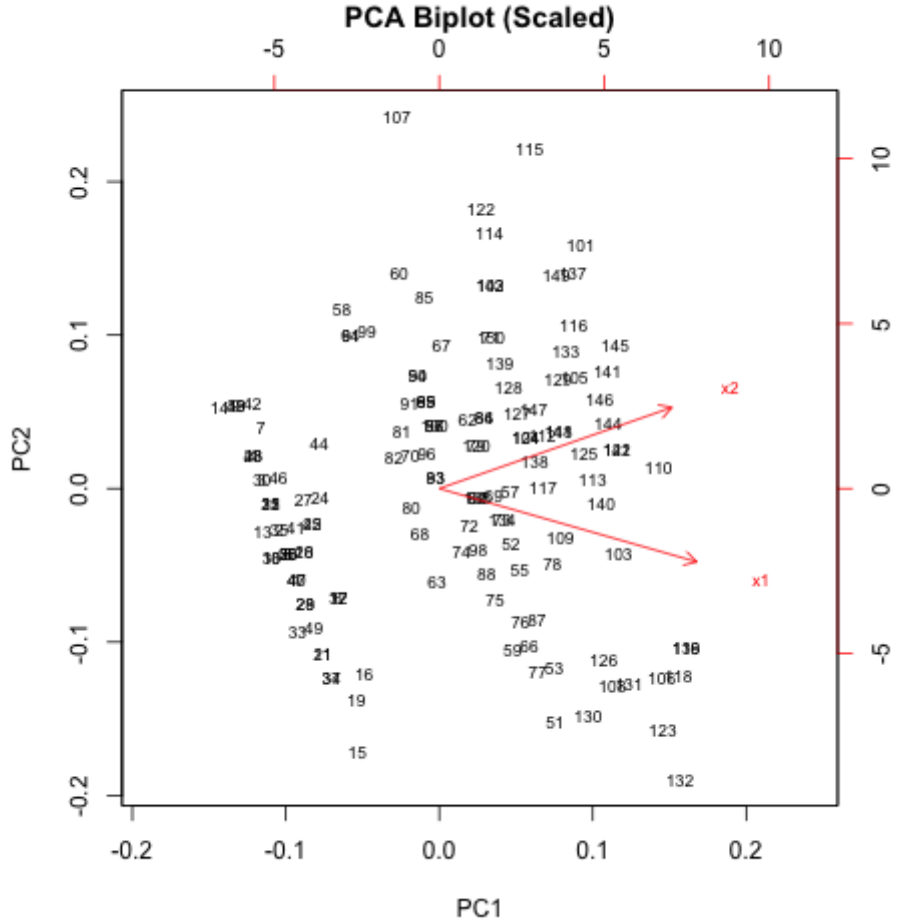
$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{S}\mathbf{V}\mathbf{V}^T = \mathbf{U}\mathbf{S}$$

```
DataX = data.frame(x1 = iris$Sepal.Length, x2 = iris$Petal.Width)
(tmp = prcomp(DataX))
```

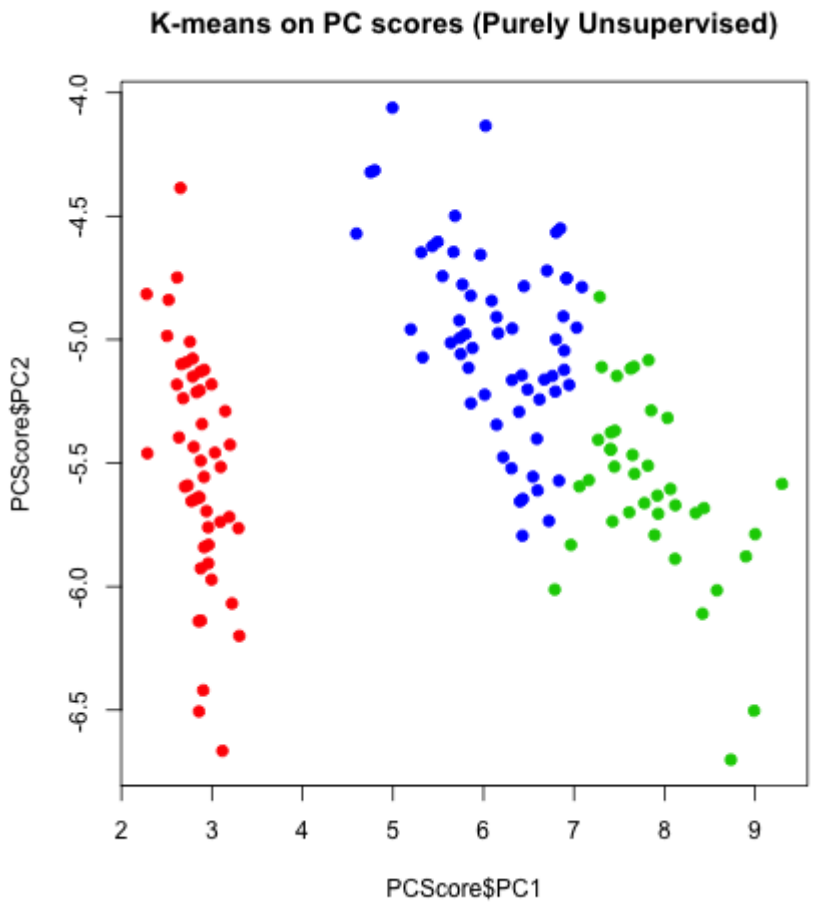
```
## Standard deviations (1, .., p=2):
## [1] 1.0734371 0.3382787
##
## Rotation (n x k) = (2 x 2):
##           PC1        PC2
## x1 0.7419133 -0.6704958
## x2 0.6704958  0.7419133
```

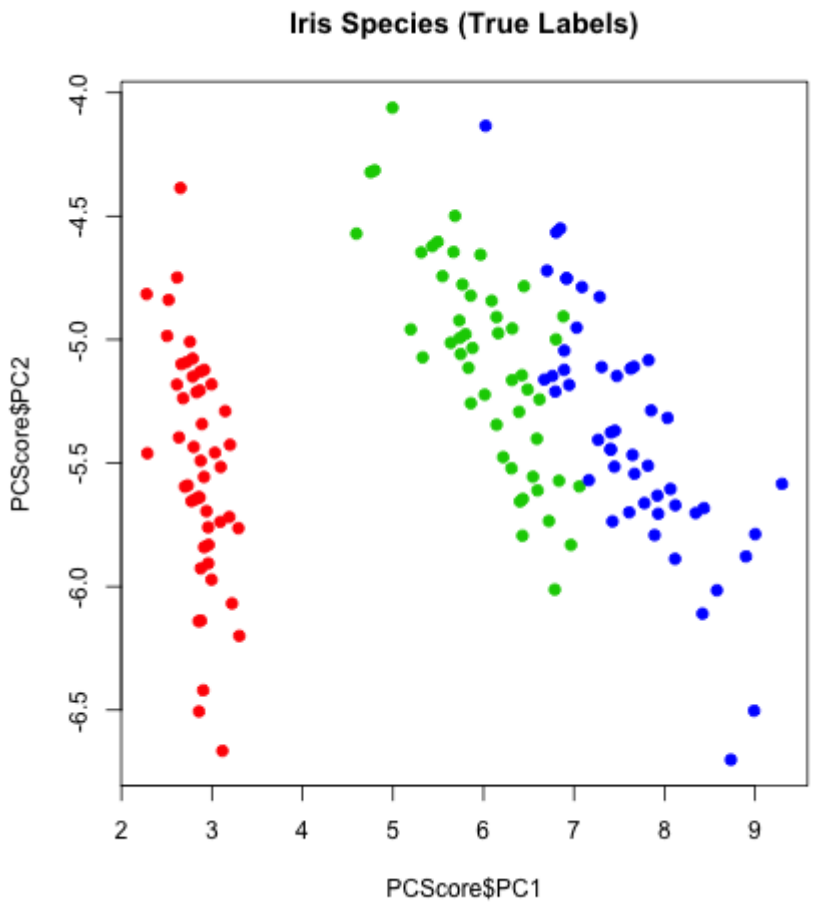


```
biplot(tmp, cex=0.7, main="PCA Biplot (Scaled)", scale=T)
```



K-means on PC1, PC2





Thank you!

Q&A or Email ajzhang@umich.edu。