Unsupervised Learning

Thomas Lonon

Division of Financial Engineering
Stevens Institute of Technology

April 24, 2017

Unsupervised Learning

Unlike in *supervised learning*, we do not look to make predictions using our data sets

Instead, we are interested in whether there are relationships or groups within a data set. This is a part of *exploratory data analysis*

The techniques we concentrate on are:

- Principal Components Analysis (PCA)
- Clustering

Principal Component Analysis refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data.[1]

PCA is an unsupervised approach, as we have a data set $X_1, X_2, \ldots, X_p$ and no associated response

The *first principal component* of a set of features is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

that has the largest variance.[1]

By *normalized* we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

where the ϕ_{j1}'s are the *loadings* of the first principle component

Given a $n \times p$ data set X, we have the first principal component loading vector as the one that solves

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

We could also view the principal components as the *closest*. We seek a dimension of data that is closest to all of the n observations.

The first dimension would give a line, the second a plane and so on.

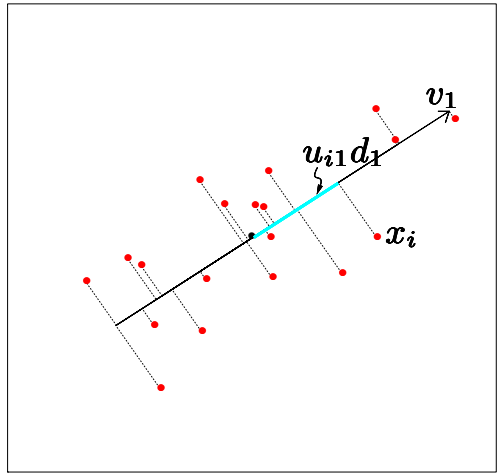

FIGURE 14.20. *The first linear principal component of a set of data. The line minimizes the total squared distance from each point to its orthogonal projection onto the line.*

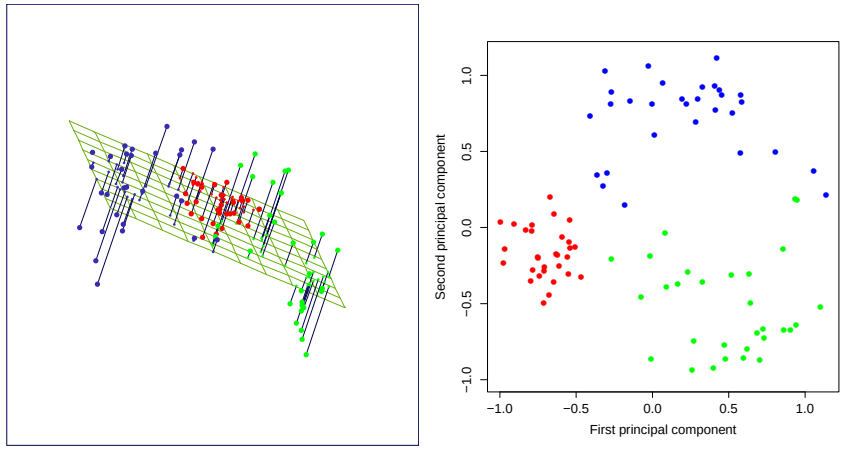

FIGURE 14.21. *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by* $\mathbf{U}_2\mathbf{D}_2$*, the first two principal components of the data.*

Based on this approach, the first M principal component score vectors and loading vectors provide the best M-dimensional approximation:

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}$$

When $M = \min(n-1, p)$, the representation is exact.

- Scaling the Vectors
- Uniqueness of the Principal Components
- The Proportion of Variance Explained (PVE)

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

- Deciding How Many Components to Use

We decide on the number of principal components by using a *scree plot*

The idea is to choose the smallest number of components that explains a "suitable" amount of the variance.

We tend to look for the point of drop-off or the *elbow*

Clustering

Broad set of techniques for finding *subgroups* or *clusters*.

- PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance.
- Clustering looks to find homogeneous subgroups among the observations

K-Means Clustering

Technique involving choosing a number of clusters (K) and then assigning each element in the data to one and only one of these clusters.

Define sets $C_1, \ldots, C_K$ containing the indices or the observations, then:

1. $C_1 \bigcup C_2 \bigcup \cdots \bigcup C_K = \{1, \ldots, n\}$
2. $C_k \bigcap C_{k'} = \emptyset$ for all $k \neq k'$

A *good* cluster is one for which the *within-cluster variation* is as small as possible. If we denote this variation as $W(C_k)$ then the goal is:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

The most common choice of $W(C_k)$ is the *squared Euclidean distance* defined as:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

where $|C_k|$ is the number of elements in the set.[1]

Algorithm 10.1: K-Means Clustering

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   - 2.1 For each of the K clusters, compute the *centroid*. The k^{th} cluster centroid is the vector of the p feature means for the observations in the k^{th} cluster.
   - 2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance)

Hierarchical Clustering

Does not require that we commit to a choice of K.

Results in a tree-based representation of the observations called a *dendrogram*

Start with the leaves and fuse the branches together as you move up in the tree.

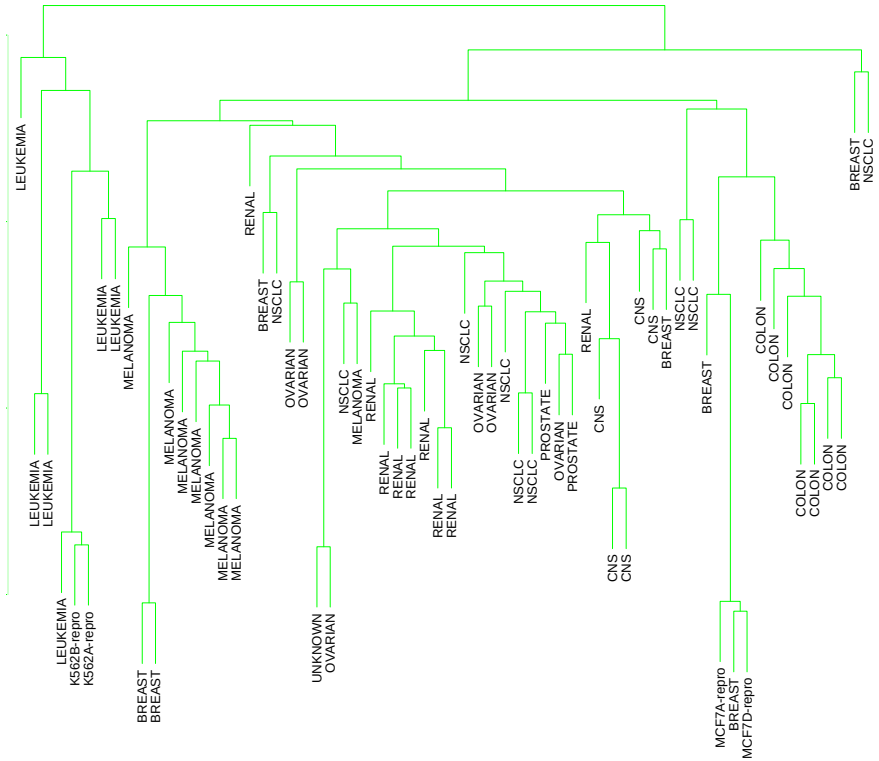

FIGURE 14.12. *Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.*

Algorithm 10.2: Hierarchical Clustering

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
   - 2.1 Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
   - 2.2 Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

Linkage	Description
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster *A* and the observations in cluster *B*, and record the largest of these dissimilarities
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster *A* and the observations in cluster *B*, and record the smallest of these dissimilarities Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster *A* and the observations in cluster *B* and record the average of these dissimilarities
Centroid	Dissimilarity between the centroid for cluster *A* (a mean vector of length *p*) and the centroid for cluster *B*. Centroid linkage can result in undesirable inversions

[1]

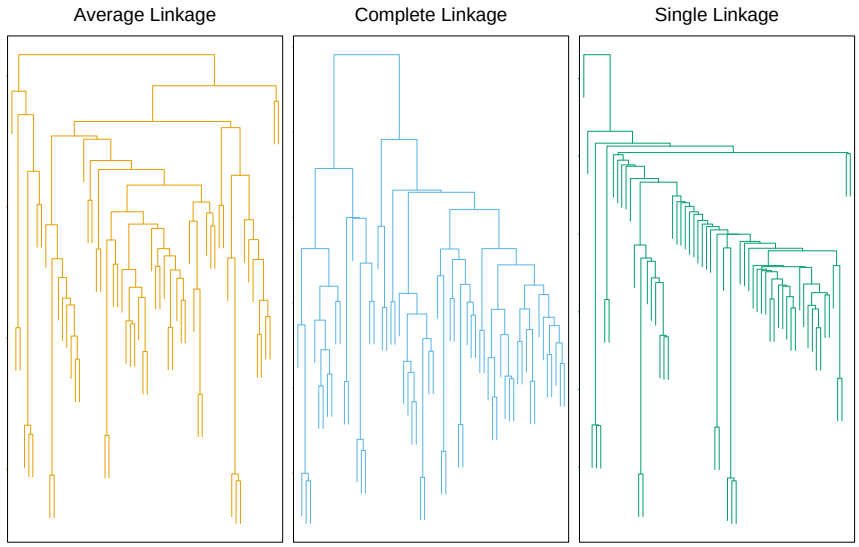

FIGURE 14.13. *Dendrograms from agglomerative hierarchical clustering of human tumor microarray data.*

[1] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Number v. 6. Springer, 2013.