Statistics and Data Analysis

Clustering

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- Recall the wholesale data set:

Channel	Label	Fresh	Milk	Grocery	Frozen	D. & P.	Deli.
1	1	30624	7209	4897	18711	763	2876
1	1	11686	2154	6824	3527	592	697
				⋮			
2	3	14531	15488	30243	437	14841	1867

- The wholesaler records the annual amount each customer spends on six product categories:
  - Fresh, milk, grocery, frozen, detergents and paper, and delicatessen.
  - Amounts have been scaled to be based on "monetary unit."
- Channel: hotel/restaurant/café = 1, retailer = 2.
- Region: Lisbon = 1, Oporto = 2, others = 3.

Dividing customers into groups

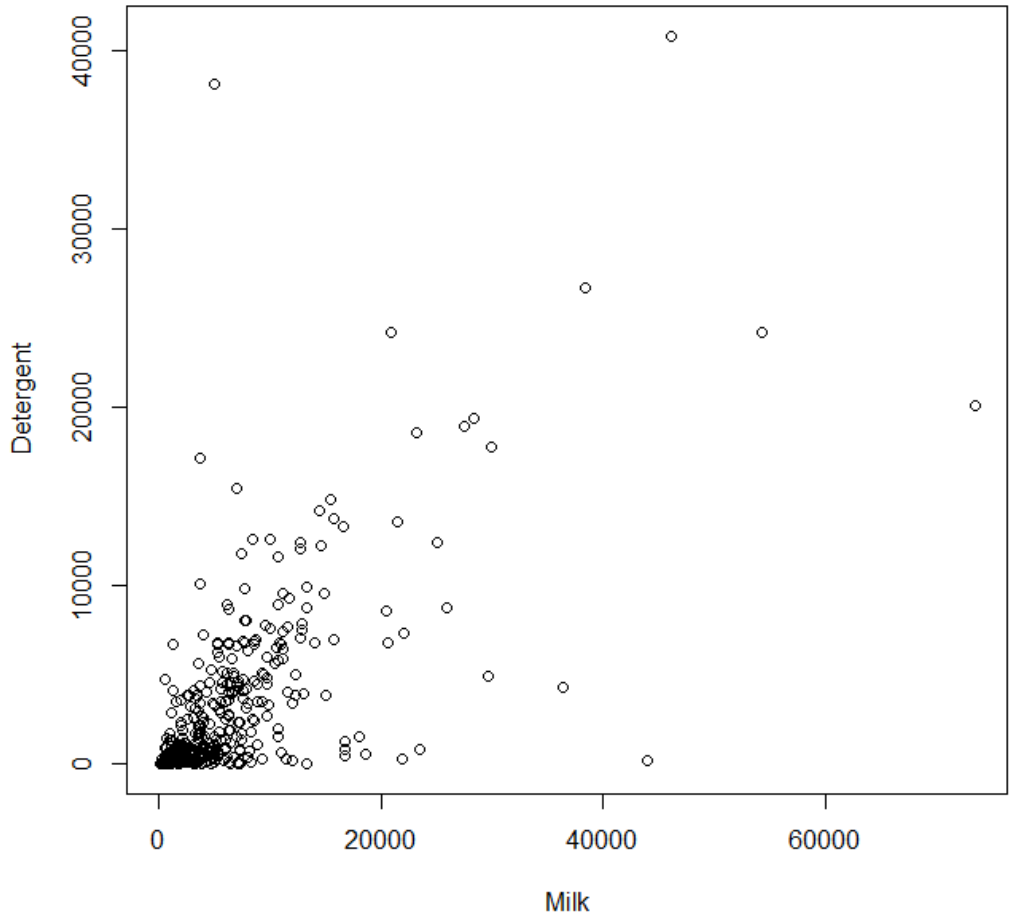
- In many cases, we would like to **customize** the advertising, service, and selling plans for different customers.
  - E.g., the price for milk may be different from customer to customer.
  - E.g., we may assign special agents for big customers.
- While there are 440 customers, we do not want to have 440 ways.
  - We want to **divide** customers to **groups**.
  - According to channel, region, a kind of sales, or what?
- This task is called **clustering**.

Clustering vs. classification

- Both **clustering** and **classification** are grouping data points (e.g., customers) into groups.
- However, they are different.
- Classification: Group information is **known** for existing data points.
  - Each existing data point is known to be in a group,
  - E.g., survival or death of a person, purchasing or not of a customer.
  - We use existing data points to identify critical factors leading to the grouping outcomes.
  - For future data whose groups are unknown, we classify them into groups.
- Clustering: Group information is **unknown** for existing data points.
  - We divide data points to clusters to make points within a class as **similar** as possible.
  - A future data point is put into the cluster that is "closest" to it.

Example

- How to create 6 clusters based on the milk and Detergent sales?



Cluster centers and distances

- Let $x^i = (x_1^i, x_2^i)$ be data point i, $i = 1, ..., 440$, where x_1^i and x_2^i are its milk and detergent sales, respectively.
- We want to create 6 clusters.
  - Let C_j be the set of points in cluster j, $j = 1, ..., 6$.
  - For cluster j, there is a **cluster center** $c^j = (c_1^j, c_2^j)$, $j = 1, ..., 6$.
  - If a point is in cluster j (i.e., $x^i \in C_j$), its **distance** to cluster center c^j is no longer than that to cluster c^k for all $k \neq j$.
  - The (Euclidean) distance between two points x^i and c^j is

$$d(x^i, c_j) = \sqrt{(x_1^i - c_1^i)^2 + (x_2^i - c_2^i)^2}.$$

- Therefore, the task of making 6 clusters is equivalent to choosing 6 points to be cluster centers.
  - A cluster center needs not to be an existing data point.

Quality of a set of clusters

- How to measure the quality of a set of 6 clusters?
- In cluster j, we want

$$\sum_{i \in C_j} d(x^i, c_j)^2 = \sum_{i \in C_j} \left[(x_1^i - c_1^i)^2 + (x_2^i - c_2^i)^2 \right]$$

  to be small, i.e., the points in the cluster are close to the center.
- We want to find 6 centers to minimize the **within-cluster sum of squared errors**

$$\text{WSSE} = \sum_{j=1}^{6} \sum_{i \in C_j} d(x^i, c_j)^2 = \sum_{j=1}^{6} \sum_{i \in C_j} \left[(x_1^i - c_1^i)^2 + (x_2^i - c_2^i)^2 \right].$$

Quality of a set of clusters

- If we only have one cluster, the within-cluster sum of squared errors can be minimized by setting the cluster center at $\bar{x}$, where

$$\bar{x}_p = \frac{\sum_{i=1}^{440} x_p^i}{440}.$$

- Let

$$\text{TSSE} = \sum_{i=1}^{440} d(x^i, \bar{x})^2 = \sum_{i=1}^{440} \left[(x_1^i - \bar{x}_1)^2 + (x_2^i - \bar{x}_2)^2 \right],$$

- Hopefully the fraction $\frac{WSSE}{TSSE}$ is small.

Finding cluster centers

- To find cluster centers, we may use the R function `kmeans()`.

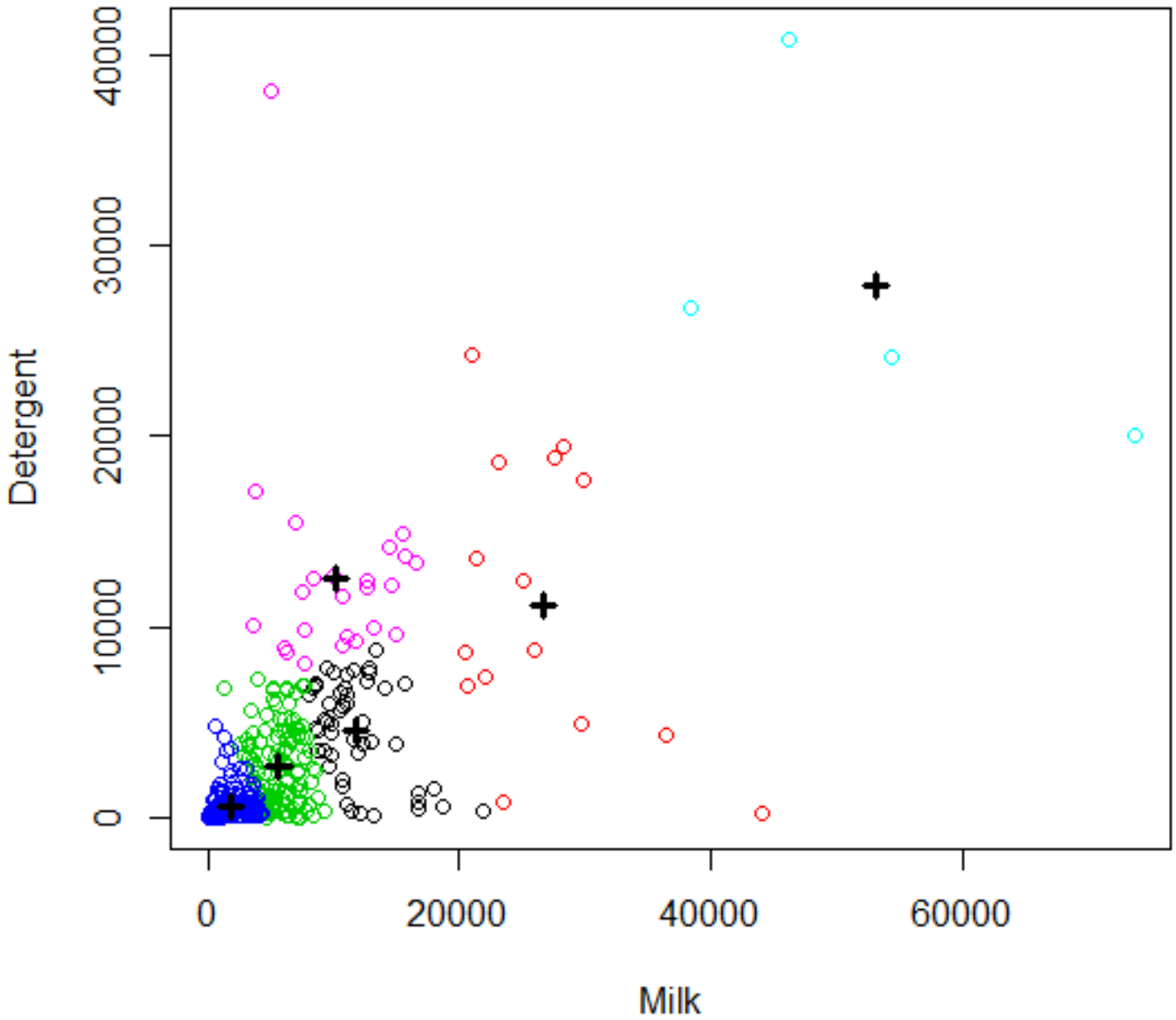
```
W <- read.table("wholesale.txt", header = TRUE)
w <- W[, c(4, 7)]
km <- kmeans(w, centers = 6)
```

- The object `km` contains information about clusters.
  - `km$cluster` indicates the cluster each point belongs to.
  - `km$center` contains the coordinates of the cluster centers.
  - `km$totss` is TSSE.
  - `km$withinss` is WSSE.

Finding cluster centers

- Let's visualize the clustering outcome.

```
plot(w[, ], xlab = "Milk", ylab = "Detergent")
for(i in 1:6)
  points(w[which(km$cluster == i), ], col = i)
points(km$centers, col = 9, lwd = 3, pch = 3)
```


Detergent
0
10000
20000
30000
40000
0
20000
40000
60000
Milk

Five remaining questions

- The scales of milk and detergent sales are different.
- How to decide the number of clusters to build?
- May we use more than two variables?
- May we use categorical variables?
- How to choose variables for the clustering process to be based on?

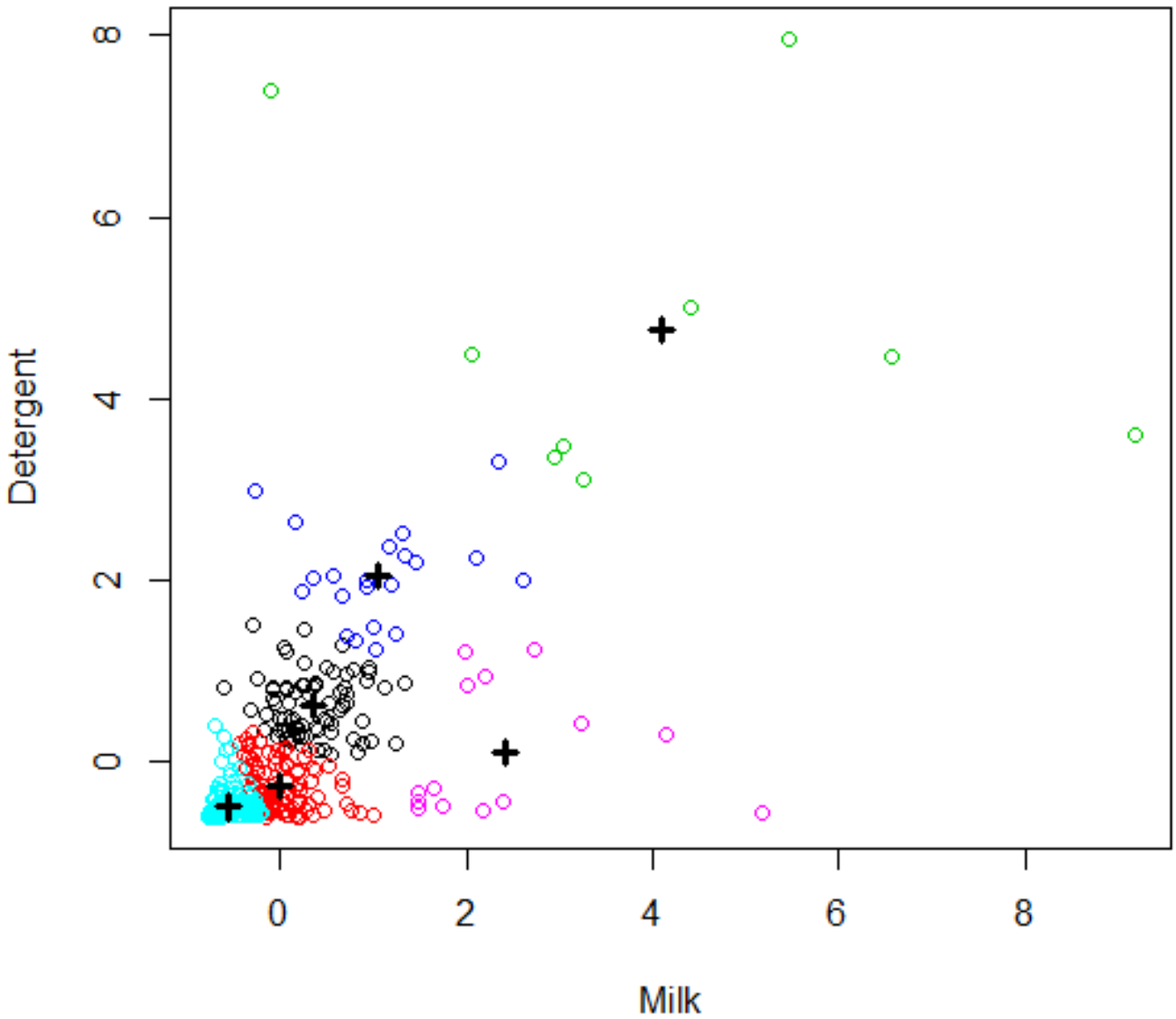
Scaling variables before clustering

- The scales of milk and detergent sales are different.
- In this case, we may scale them first.
- The most common way is to **standardize** each of them into ***z*-scores**:

$$z_p^i = \frac{x_p^i - \bar{x}_p}{s_p}, \text{ where } s_p = \sqrt{\frac{\sum_{i=1}^{440}(x_p^i - \bar{x}_p)^2}{440}}.$$

- In R:

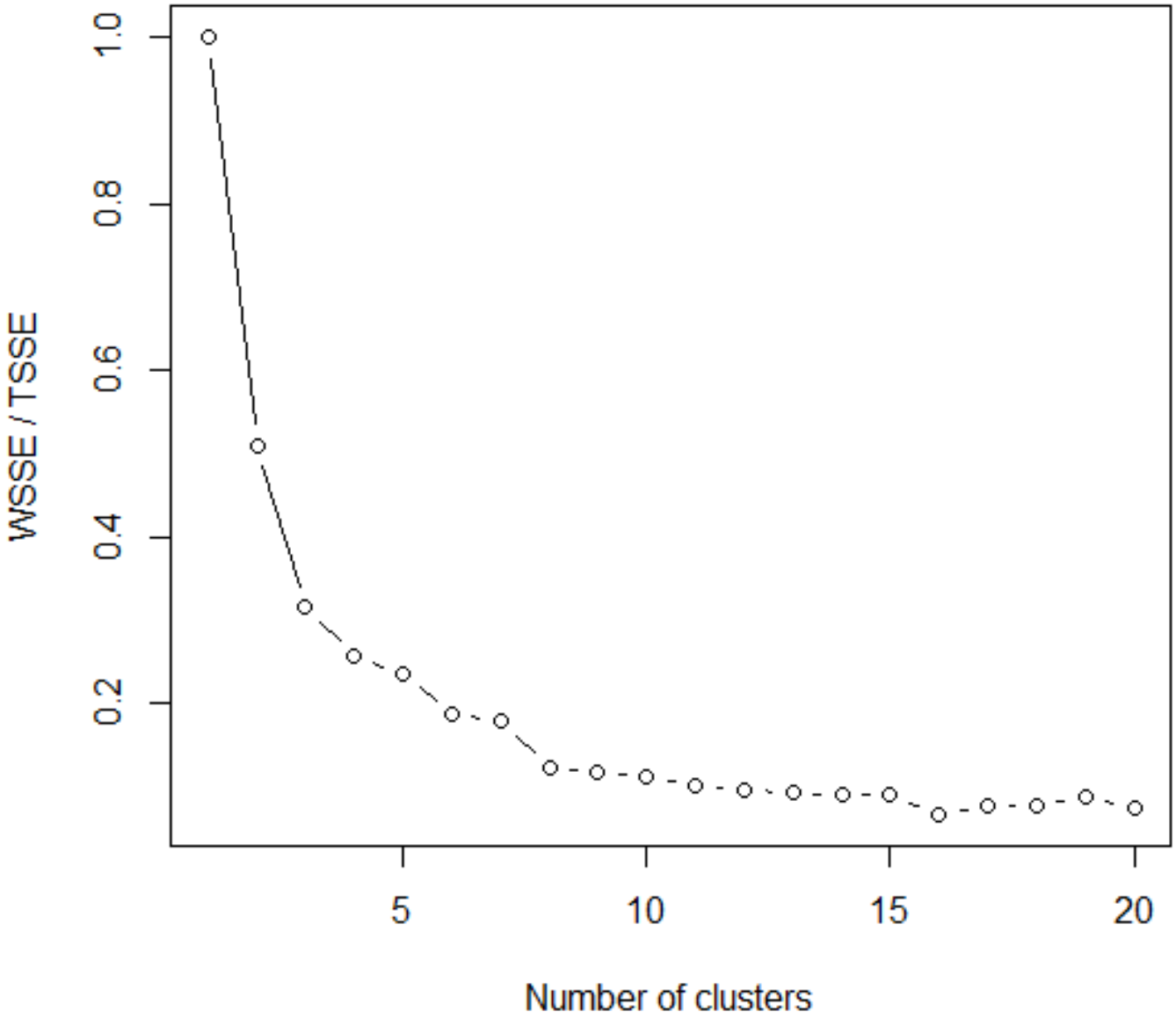
```
w[, 1] <- (w[, 1] - mean(w[, 1])) / sd(w[, 1])
w[, 2] <- (w[, 2] - mean(w[, 2])) / sd(w[, 2])
```


Detergent
Milk
0
2
4
6
8
0
2
4
6
8

Number of clusters

- The more clusters, the smaller WSSE.
  - However, each cluster also becomes less informative.
- We typically stop increasing the number of clusters when the **marginal improvement on WSSE** becomes too small.
- In R:

```
z <- rep(0, 20)
for(k in 1:20)
{
  km <- kmeans(w, centers = k)
  z[k] <- km$tot.withinss / km$totss
}
plot(z, type = "b", xlab = "Number of clusters",
     ylab = "WSSE / TSSE")
```


WSSE / TSSE
0.2
0.4
0.6
0.8
1.0
5
10
15
20
Number of clusters

Using more than two variables

- We may include as many variables as we want.
  - As long as they are **quantitative**.
- In R:

```
w <- W[, 3:8]
for(i in 1:6)
  w[, i] <- (w[, i] - mean(w[, i])) / sd(w[, i])
km <- kmeans(w, centers = 6)
```

Categorical variables

- May we include a categorical variable in the clustering process?
- Unfortunately, no!
  - Because there is no way to calculate distances.

How to choose variables?

- How to choose variables for the clustering process to be based on?
  - Milk and detergent?
  - Milk, fresh food, and detergent?
  - All variables?
- It depends on what you want to do.
  - The decision maker makes her own judgment.
  - Some other methods (e.g., regression) can be applied.