군집분석

이재용,임요한

서 울 대 학 교
통 계 학 과

2017년 8월

노트. 이 장에서 다룰 내용

1. 자율학습과 지도학습의 소개
2. K평균 군집분석
3. 계층군집화

군집분석 I

자율학습(unsupervised Learning)

1. 반응변수 Y는 없고 변수들 $X_1, \ldots, X_p$만 있는 경우이다.
2. 예측에는 관심이 없고, 변수들 사이의 특별한 관계가 있는가? 관측치들에 그룹이 있는가 등의 질문에 관심이 있다.
   노트.
3. 보통 탐색적 자료분석의 일부분이다.
4. 자료분석의 결론이 성능이 좋은지 안좋은지 판단하기가 어렵고, 주관적인 측면이 강하다.

지도학습(supervised learning)

1. 자료가 반응변수 Y와 설명변수 $X_1, \ldots, X_p$로 구성되어 있다.
2. 설명변수 $X_1, \ldots, X_p$가 주어져 있을 때, 반응변수 Y의 예측에 목적이 있다.

군집분석 II

군집분석

1. 자료를 동질적인 군집 혹은 부분으로 나누는 분석.
2. 각 군집을 쉽게 설명하므로 전체를 설명하고자 한다.

노트. 군집분석의 예

1. 유방암 환자들의 자료를 이용해 알려지지 않은 암의 subtype을 찾으려고 한다.
2. 사람들의 자료를 이용하여 시장을 분할하고, 특정 광고에 잘 반응할 subgroup을 찾으려 한다.

노트. 여기서 다룰 내용

K 평균 군집분석과 계층군집분석

K 평균(K means) 군집분석 I

군집

$C_1, \ldots, C_K$를 군집들이라 한다. 이는 다음의 조건을 만족한다.

$$\cup_{k=1}^{K} C_k = \{1, 2, \ldots, n\}$$
$$C_i \cap C_j = \emptyset, \ i \neq j.$$

목표

$$\sum_{k=1}^{K} W(C_k)$$

를 최소화하는 $C_1, \ldots, C_K$를 찾고자 한다.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} ||x_i - x_j||^2$$

를 많이 쓴다.

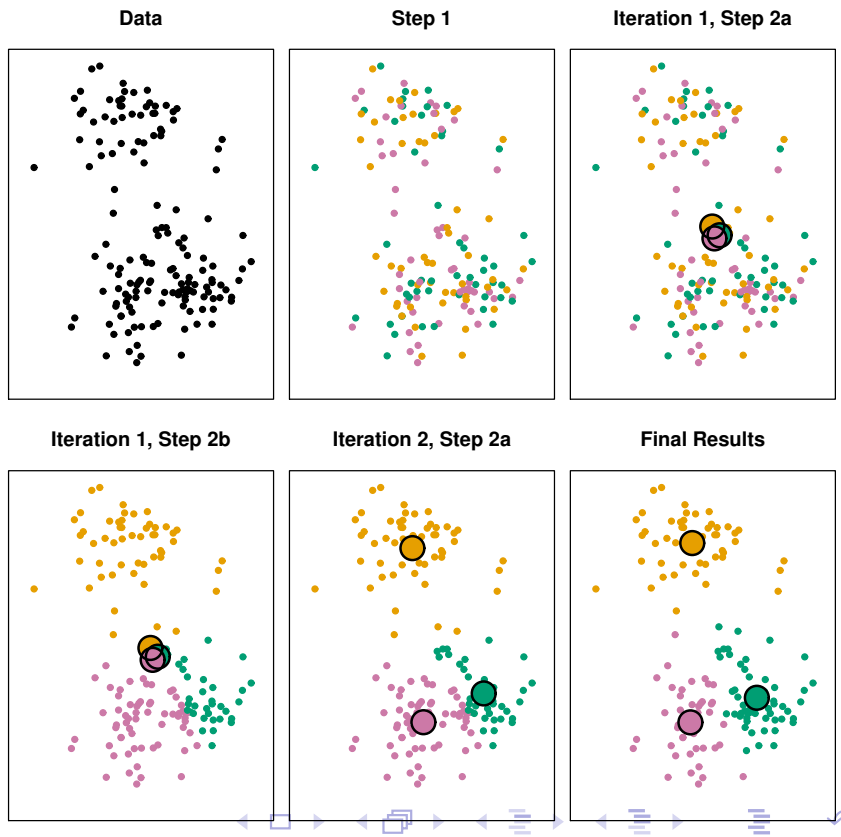
K 평균(K means) 군집분석 II

노트.

$W(C_k)$는 군집 C_k의 변동을 측정하는 측도이다. 즉, C_k의 동질성이 커지면 작아지는 값이다.

알고리듬

1. 관측치들을 $C_1, \ldots, C_K$에 랜덤하게 할당한다.
2. 군집이 바뀌지 않을 때까지 다음을 반복한다.
   - 2.1 각 군집마다 군집의 중앙(centroid)을 계산한다.
   - 2.2 모든 관측치를 가장 가까운 중심의 군집에 할당한다.



K 평균(K means) 군집분석 III

노트. 알고리듬의 수렴

위의 알고리듬은 반복이 진행될 때마다 $\sum_{k=1}^{K} W(C_k)$를 감소시킨다.

따라서, 이 알고리듬은 지역최적값(local optimum)으로 수렴한다. 지역최적값으로 수렴하기 때문에 초기 조건을 달리하여 여러 번 돌리고 그 중 최소의 변동을 갖는 군집을 최종 결론으로 한다.

노트. 그림

$K = 3$인 K 평균 군집화 과정을 보여준다.

노트. 군집의 개수의 결정 F-통계량=군집사이의 평균변동/ 군집안의 평균변동

1. 교차검증이 적용가능하다. 자료가 훈련자료와 시험자료로 나누어져 있을 때, 먼저 훈련자료에 군집분석을 적용한다. 시험자료를 추정된 군집으로 나누고, 시험자료의 군집중심까지 거리 제곱의 합으로 비교를 한다.

K 평균(K means) 군집분석 IV

2. 혼합모형 가정이 있다고 생각하고 AIC, BIC 등의 기준을 적용할 수 있다.

K 평균 군집화 R 코드 I

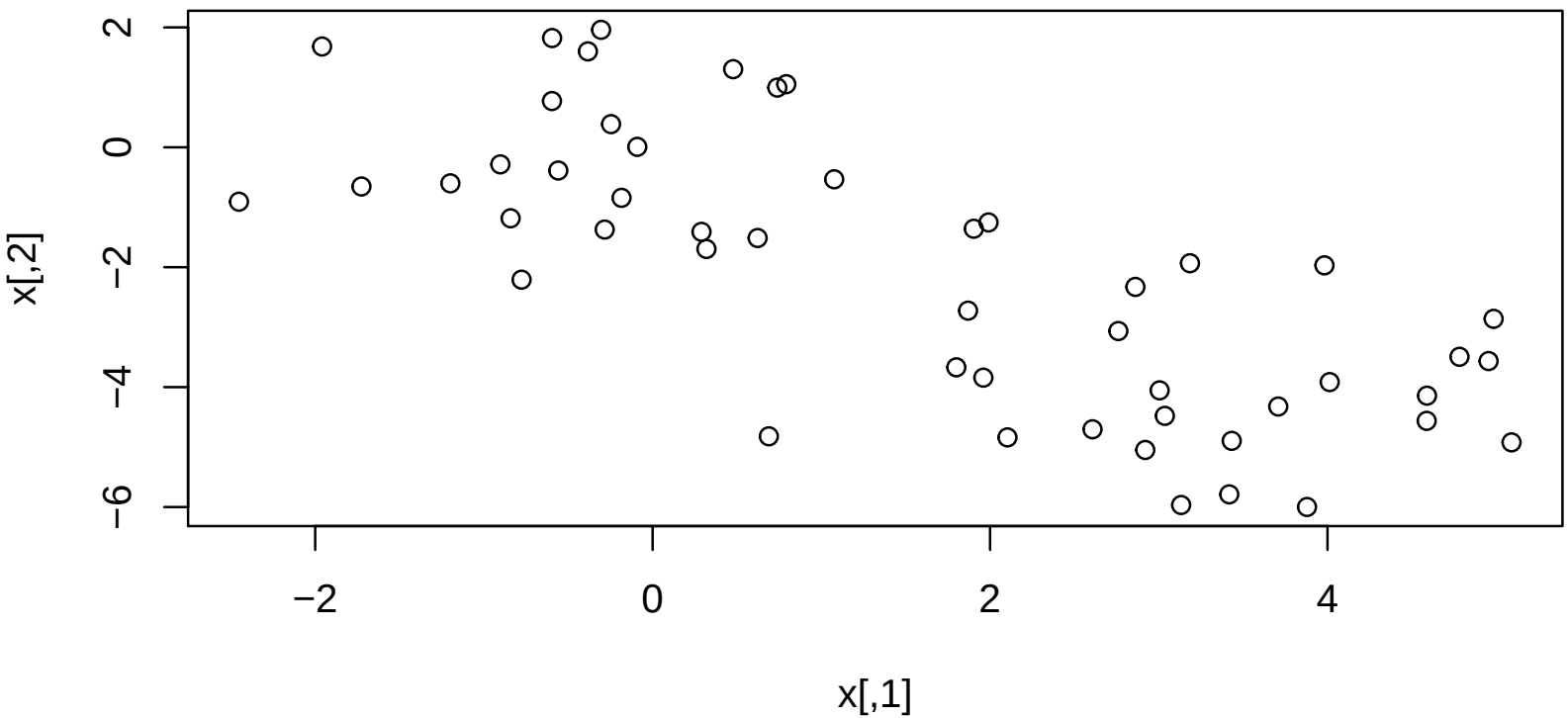
자료의 생성

```r
set.seed(2)
x=matrix(rnorm(50*2), ncol=2)
x[1:25,1]=x[1:25,1]+3
x[1:25,2]=x[1:25,2]-4
```

50개의 자료를 생성하고, 첫 25개의 x_1, x_2 값을 바꾸었다.

K 평균 군집화 R 코드 II

```
plot(x)
```



K 평균 군집화 R 코드 III

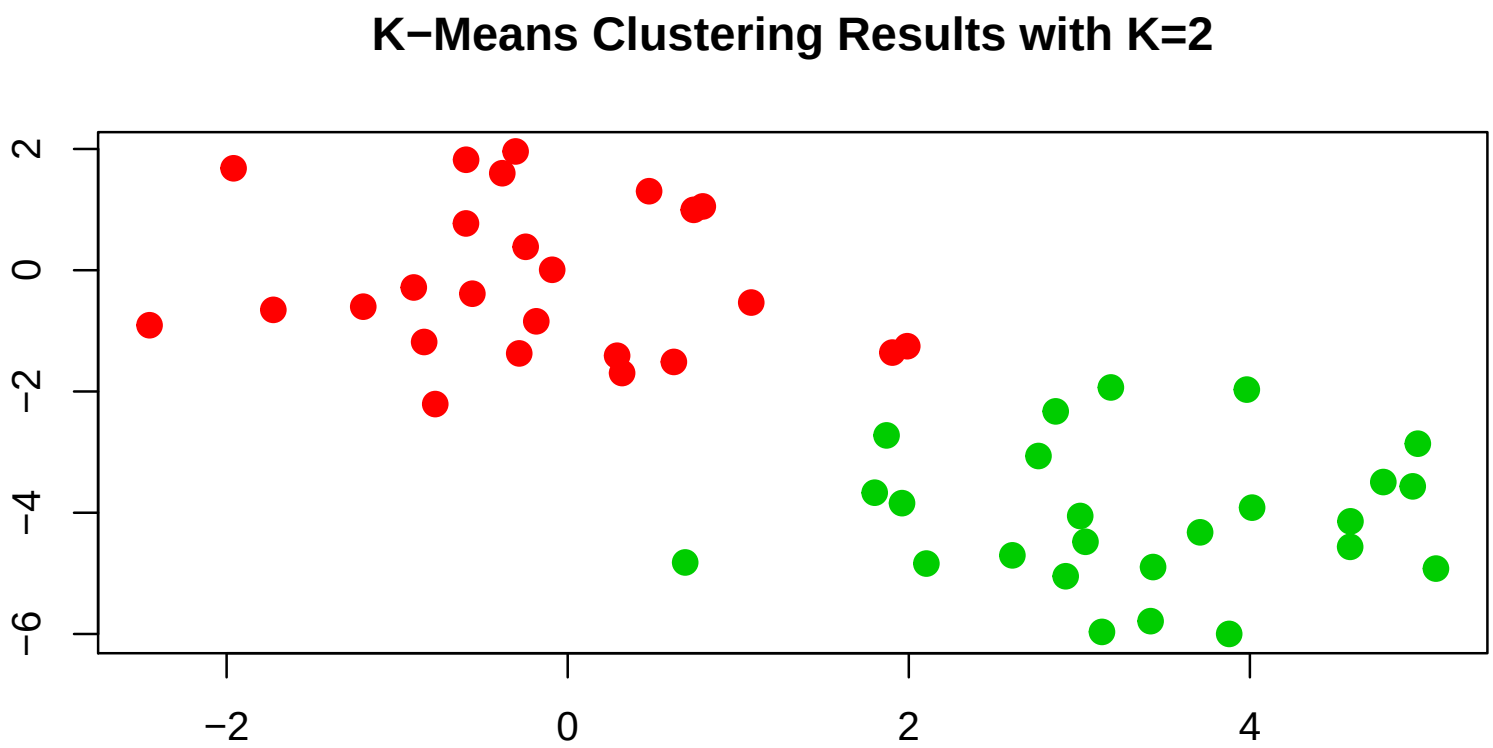
K 평균 군집화의 수행

```
km.out=kmeans(x,2,nstart=20)
```

1. $K = 2$인 군집화이다. 2는 군집의 개수를 의미한다.
2. 종종 초기값으로 쓰인 군집의 중앙값에 군집화의 결과가 다를 수 있다. nstart=20은 20개의 초기값으로 군집화를 한 후에 가장 좋은 결과를 보고하라는 뜻이다.
3. 초기 중앙값을 랜덤하게 선택하므로 수행할 때마다 결과가 다를 수 있다. 이를 방지하기 위해 set.seed 함수를 사용하는 것이 좋다.

K 평균 군집화 R 코드 IV

```
plot(x, col=(km.out$cluster+1), main="K-Means Clustering Results with K=2", xlab="",
 ylab="", pch=20, cex=2)
```



K 평균 군집화 R 코드 V

```
km.out

## K-means clustering with 2 clusters of sizes 25, 25
##
## Cluster means:
##         [,1]       [,2]
## 1 -0.1956978 -0.1848774
## 2  3.3339737 -4.0761910
##
## Clustering vector:
##  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 65.40068 63.20595
##  (between_SS / total_SS =  72.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

클러스터 벡터를 보면 정확하게 첫 25개와 후반 25개로 군집으로 나눈것을 알 수 있다.

K 평균 군집화 R 코드 VI

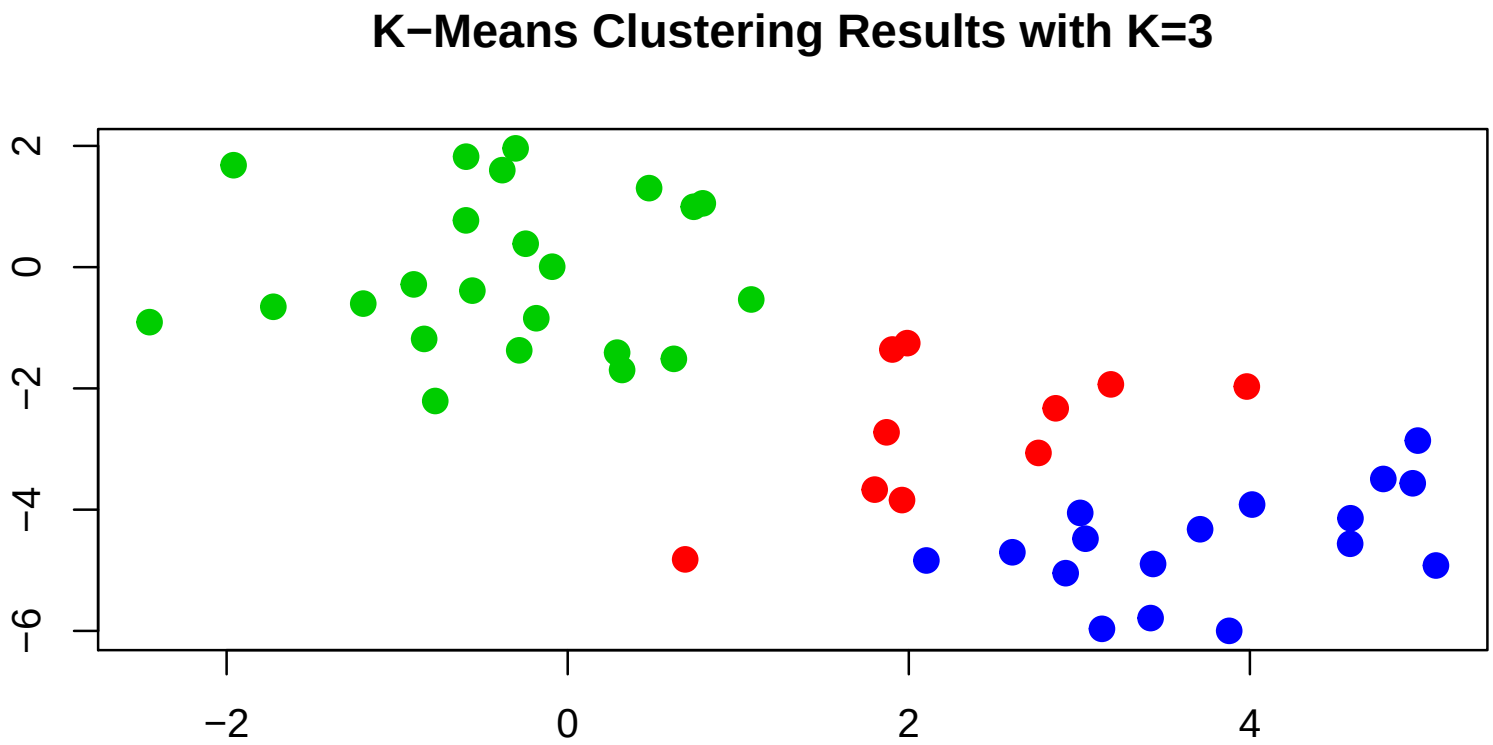
아래는 세 개의 군집으로 K 평균 알고리듬을 수행한 결과이다.

```
set.seed(4)
km.out=kmeans(x,3,nstart=20)
```

K 평균 군집화 R 코드 VII

```
km.out

## K-means clustering with 3 clusters of sizes 10, 23, 17
##
## Cluster means:
##         [,1]        [,2]
## 1  2.3001545 -2.69622023
## 2 -0.3820397 -0.08740753
## 3  3.7789567 -4.56200798
##
## Clustering vector:
##  [1] 3 1 3 1 3 3 3 1 3 1 3 1 3 1 3 1 3 3 3 3 3 1 3 3 3 2 2 2 2 2 2 2 2 2 2
## [36] 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 19.56137 52.67700 25.74089
##  (between_SS / total_SS =  79.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"

plot(x, col=(km.out$cluster+1), main="K-Means Clustering Results with K=3", xlab="",

ylab="", pch=20, cex=2)
```

K 평균 군집화 R 코드 VIII



K 평균 군집화 R 코드 IX

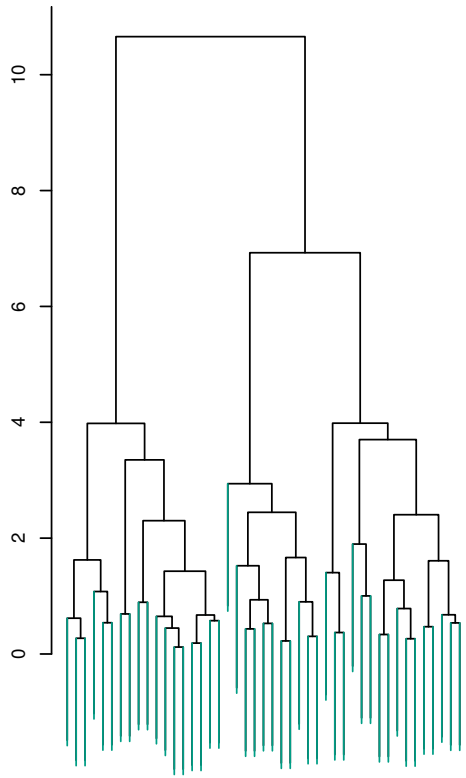
```
set.seed(3)
km.out=kmeans(x,3,nstart
=1)
km.out$tot.withinss

## [1] 104.3319

km.out=kmeans(x,3,nstart
=20)
km.out$tot.withinss

## [1] 97.97927
```

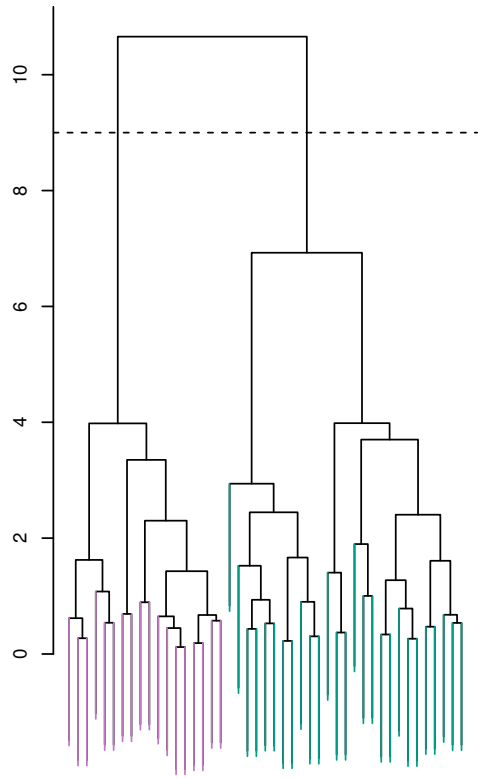
초기값을 1개 사용한 것과 20개 사용한 것을 비교하였다. 20개의 초기값을 사용한 것이 내부제곱합이 더 작다. 항상 초기값을 20 혹은 50 같이 큰 값을 사용하는 것을 추천한다.

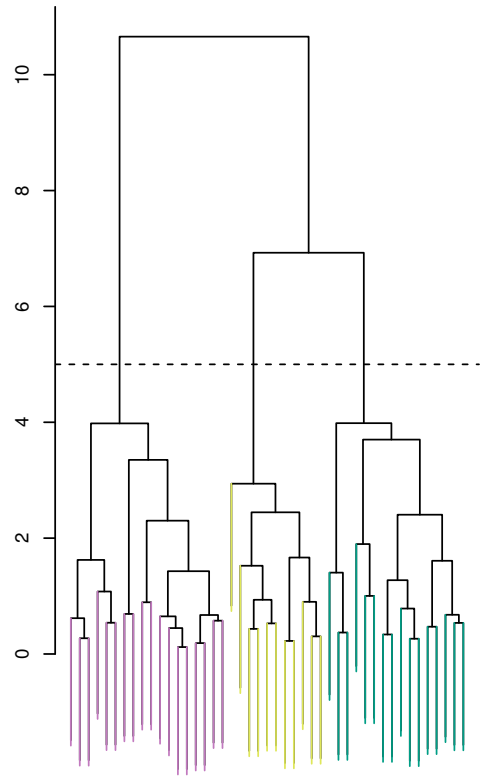
계층군집화(hierarchical clustering) I

덴도그램(dendorgram)

1. 계층 군집화는 top-down과 bottom-up 두 종류가 있다.
2. 본 강의록에서는 bottom-up을 설명한다.
3. 집합(군집)과 집합(군집)사이의 거리를 정의하여야 한다.







계층군집화(hierarchical clustering) II

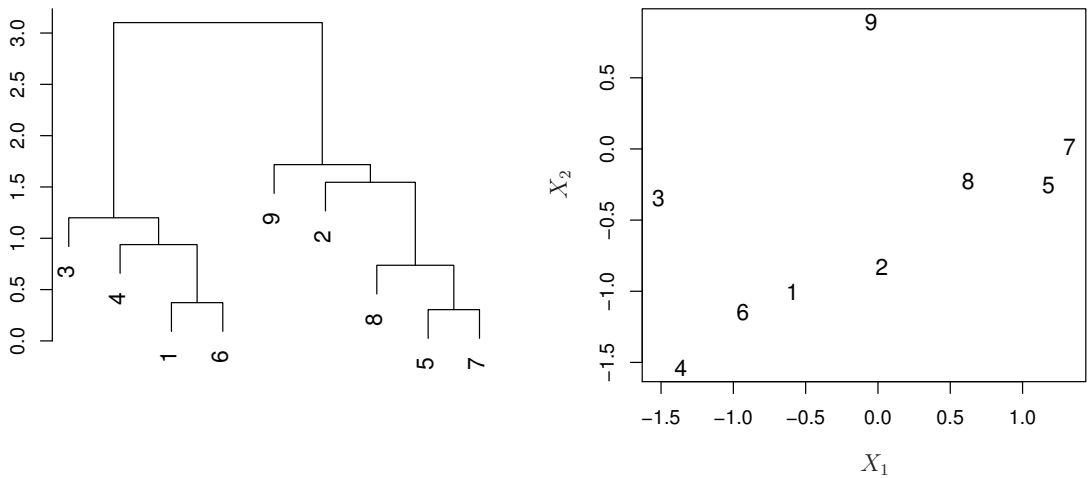
노트. 계층군집화

계층군집화의 결론으로 덴도그램이 생성된다.

1. 중간그림. $y = 9$에서 자르면 2개의 군집이 나타난다.
2. 오른쪽 그림 $y = 5$에서 자르면 3개의 군집이 나타난다.
3. 바닥에서 합쳐진 관측치들은 가깝고, 상부에서 합쳐진 관측치들은 거리가 멀다.

계층군집화(hierarchical clustering) III

노트. 주의점 : 거리의 해석



계층군집화(hierarchical clustering) IV

그림을 보면 9가 2와 제일 가깝고 5,7,8과는 멀다고 보이는데, 그렇게 해석하면 안된다. 2, 5, 7,8 부분의 덴도그림을 표현하는 방법은 매우 많다. 덴도그램의 x축은 가까움을 나타내지 않는다.

노트. 계층군집화가 적당하지 않은 자료

계층군집화 방법의 계층구조는 매우 매력적이지만 어떤 자료에는 적합하지 않다. 예를 들면 남성과 여성과 함께 백인 흑인 황인으로 나누어진 자료가 있다고 한다. 이 그룹은 네스티드되어 있지 않기 때문에 계층군집은 자료를 잘 나타내지 못하고 오히려 K 평균 방법이 더 나을 수 있다.

계층군집화(hierarchical clustering) V

계층군집화의 알고리듬

1. 한 개의 관측치가 포함된 n개의 군집으로 시작한다. n개의 군집간 거리를 계산한다.
2. $i = n, n-1, \ldots, 2$
   - 2.1 i개의 군집 간의 거리를 재서 가장 거리가 작은 군집 2개를 합친다.
   - 2.2 $i-1$개의 군집간 거리를 계산한다.

연결법(linkage) : 군집간의 거리를 계산하는 방법

1. complete

$$d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(x_i, x_j)$$

2. single

$$d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(x_i, x_j)$$

3. average

$$d(C_1, C_2) = ave_{i \in C_1, j \in C_2} d(x_i, x_j)$$

4. centroid

$$d(C_1, C_2) = d(\bar{x}_1, \bar{x}_2), \ \bar{x}_i = ave_{j \in C_i}(x_j).$$

계층군집화(hierarchical clustering) VI

노트

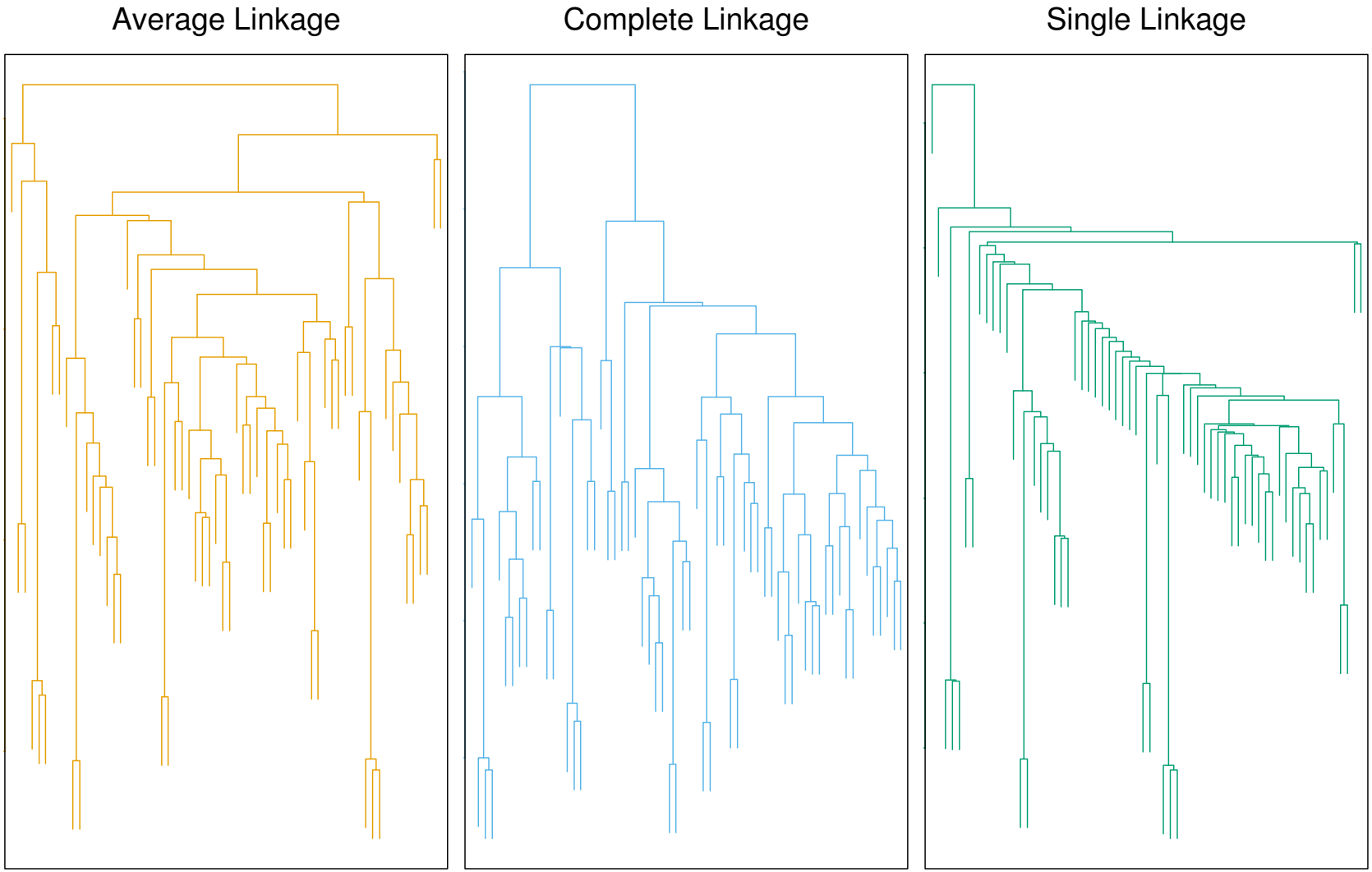
1. 보통 complete이나 average가 선호된다. 왜냐하면 balanced dendogram을 만든다고 한다. 이유는 정확히 이해 못했다.
2. centroid는 genomics에서 종종 사용되는데 inversion이 생길 수 있는게 문제이다. inversion이 무엇인가?
3. 덴도그램은 연결법에 따라 매우 다르게 나타난다. 아래의 그림을 참조
4. 거리 측도도 군집의 형성에 많은 영향을 미친다. 유클리디언 거리외에 두 벡터 사이의 상관계수도 많이 쓰인다.

계층군집화(hierarchical clustering) VII

연결법에 따른 덴도그램의 차이

덴도그램은 연결법에 따라 매우 다르게 나타난다.

계층군집화(hierarchical clustering) VIII



계층군집화(hierarchical clustering) IX

노트.

군집은 purtabation에 robust하지 않다.

노트. 군집분석에서 결정해야할 문제들

1. 변수들을 표준화해야 하는가?
2. 계층 군집의 경우
   - 2.1 거리는 어떤 것을 사용해야하는가?
   - 2.2 연결법은?
   - 2.3 덴도그램은 어디서 짤라야 하는가?
3. K 평균 방법에서 K는 몇 개를 해야하나?

답변

위의 질문들에 대해 명확한 답은 없다. 보통 여러 개를 시도해보고 이 중 가장 해석이 좋은 것을 선택한다.

계층 군집화 R 코드 I

```
hc.complete=hclust(dist(x), method="complete")
hc.average=hclust(dist(x), method="average")
hc.single=hclust(dist(x), method="single")
```

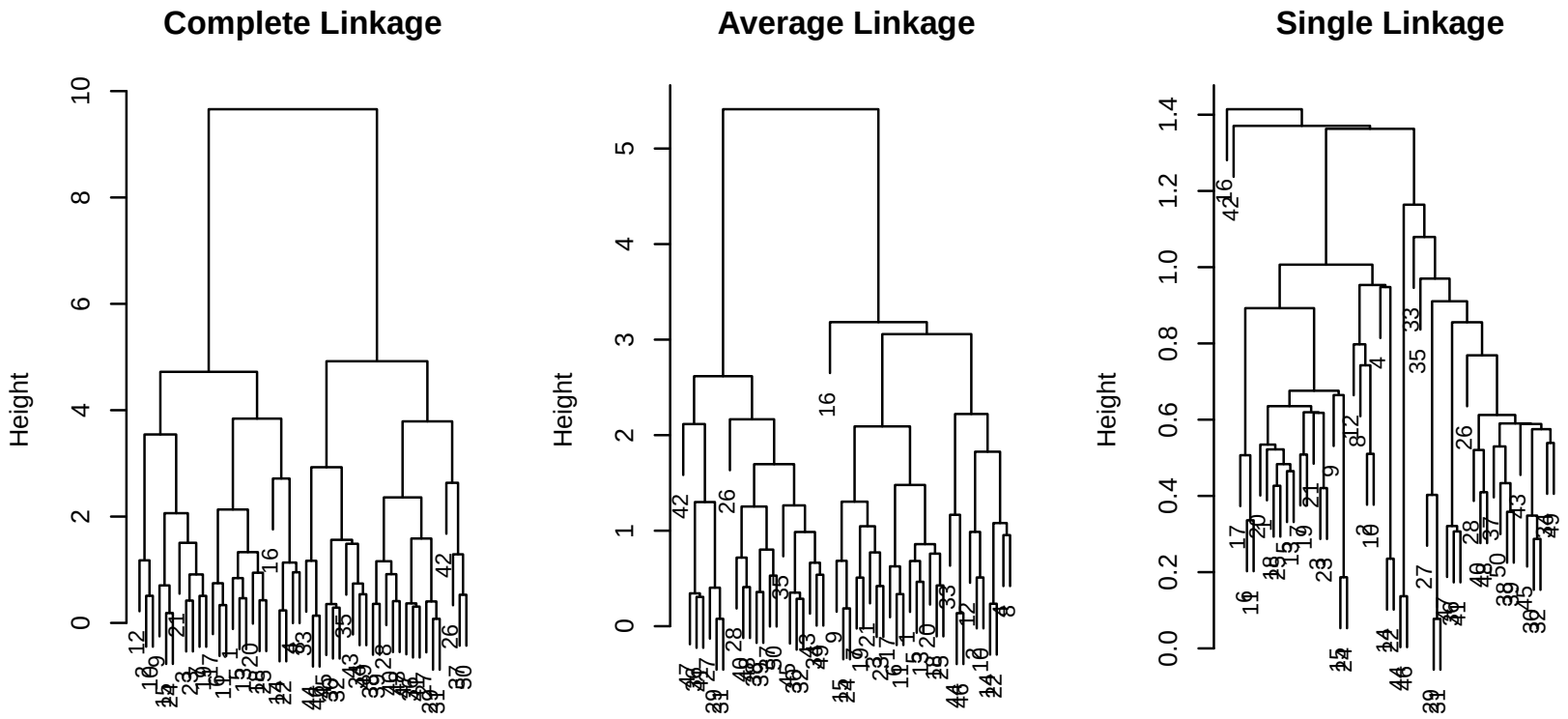
계층군집화를 수행하는 함수는 hcluster이다. dist(x)는 자료들 간의 거리를 구해주는 함수이다. method는 연결방법을 지정한다.

```
dist(x[1:4,])
```

```
##          1        2        3
## 2 3.099491
## 3 2.500046 2.979541
## 4 2.126855 1.534550 3.281285
```

```
par(mfrow=c(1,3))
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
plot(hc.average, main="Average Linkage", xlab="", sub="", cex=.9)
plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
```

계층 군집화 R 코드 II



그림을 그릴 때는 plot 함수를 사용한다.

계층 군집화 R 코드 III

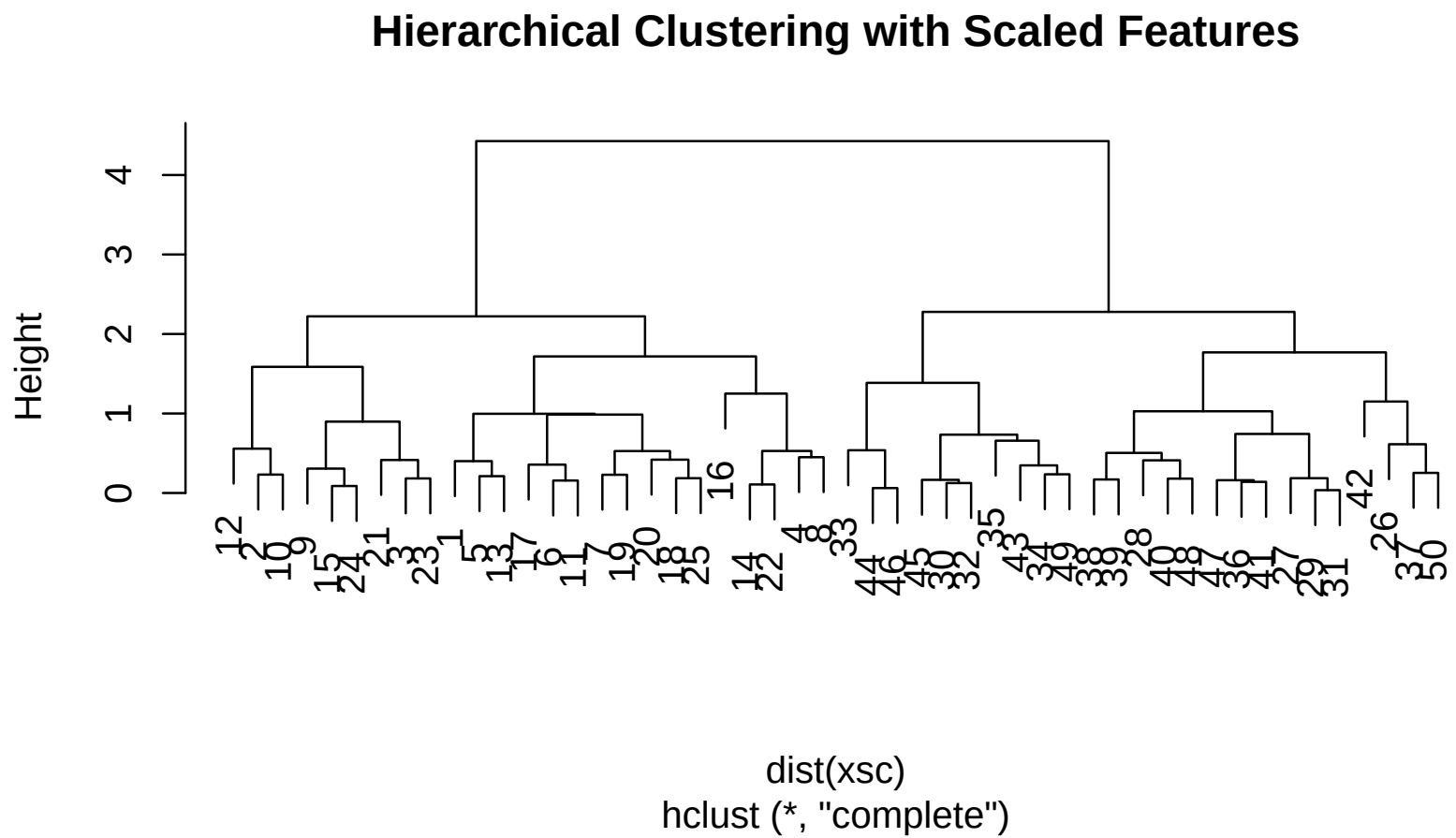
cutree(tree, k = NULL, h = NULL)

```
cutree(hc.complete, 2)

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
## [36] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

cutree(hc.average, 2)

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 2 2
## [36] 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2

cutree(hc.single, 2)

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

cutree(hc.single, 4)

##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3
## [36] 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3
```

군집의 인덱스를 구하는 함수는 cutree이다. k=2는 군집의 개수를 지정한다.

계층 군집화 R 코드 IV

```r
xsc=scale(x)
plot(hclust(dist(xsc), method="complete"),
 main="Hierarchical Clustering with Scaled Features")
```



변수를 표준화해서 계층군집화를 수행하였다.

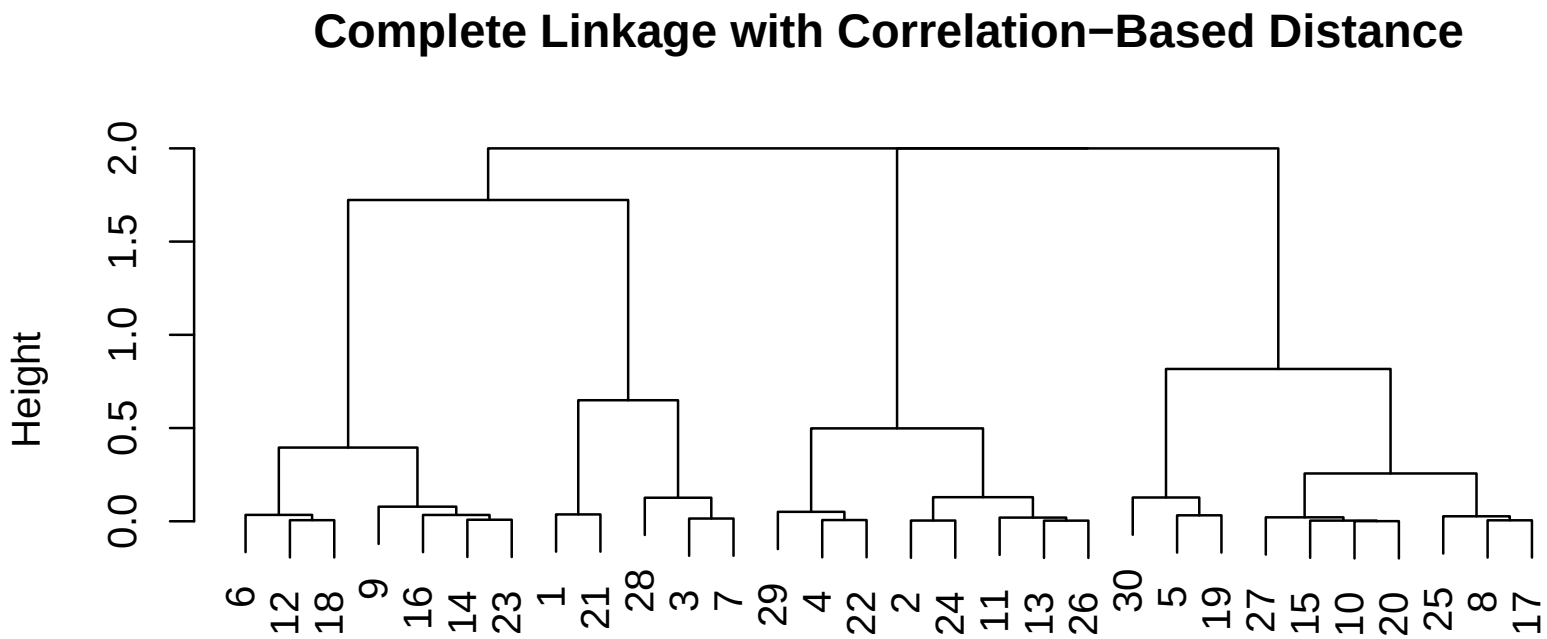
계층 군집화 R 코드 V

```
x=matrix(rnorm(30*3), ncol=3)
dd=as.dist(1-cor(t(x)))
```

as.dist는 주어진 행렬을 거리행렬로 바꾸어준다.

```
plot(hclust(dd, method="complete"),
 main="Complete Linkage with Correlation-Based Distance", xlab="", sub="")
```

계층 군집화 R 코드 VI


Complete Linkage with Correlation−Based Distance
Height
0.0
0.5
1.0
1.5
2.0
6
12
18
9
16
14
23
1
21
28
3
7
29
4
22
2
24
11
13
26
30
5
19
27
15
10
20
25
8
17

거리 – Distance

dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

x
a numeric matrix, data frame or "dist" object.

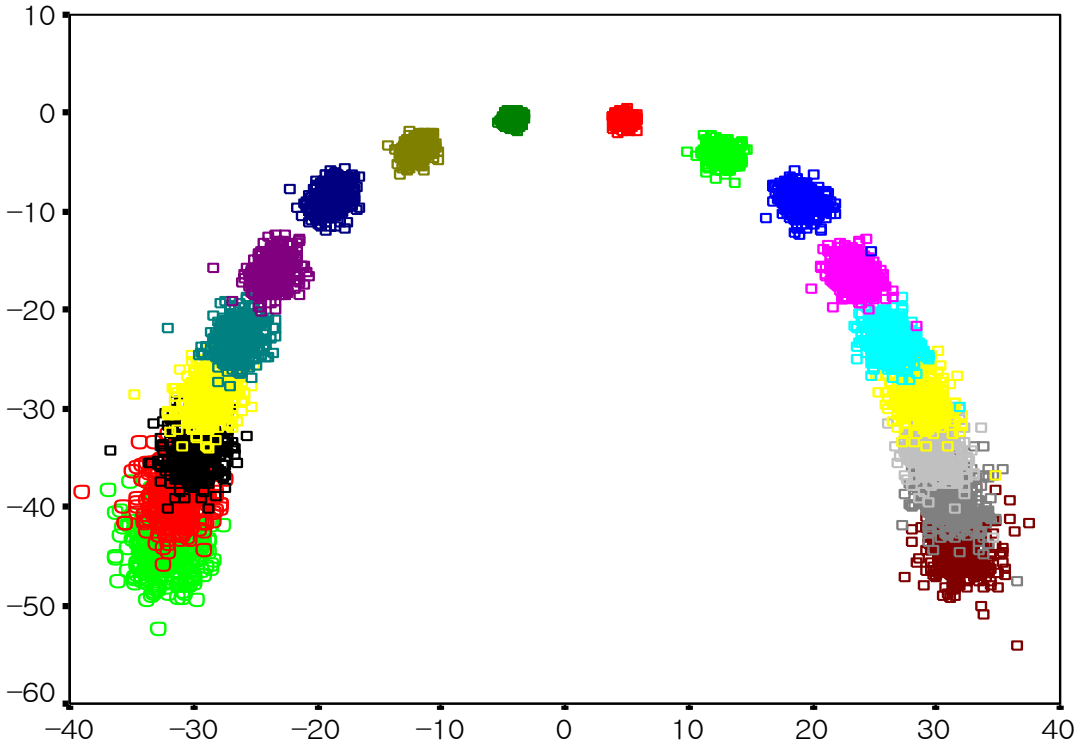
method
the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.

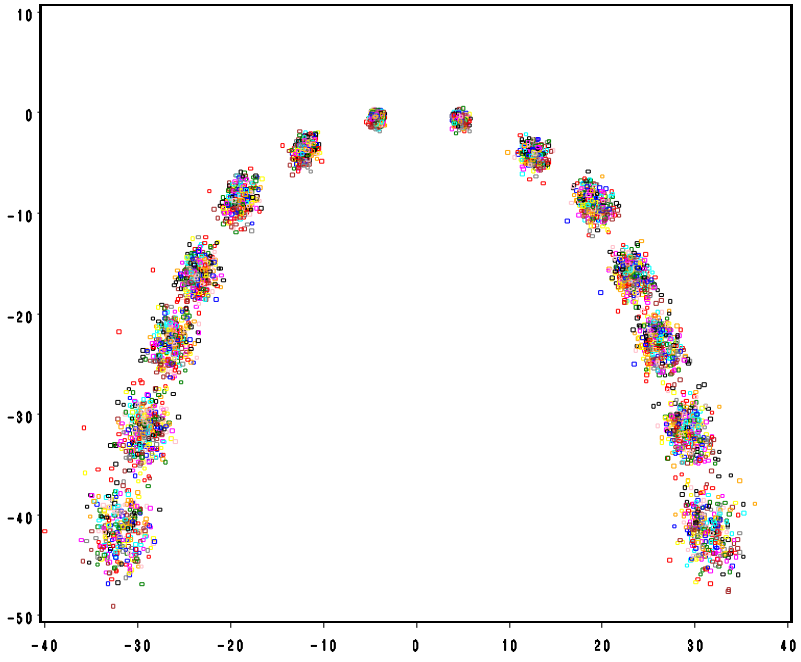
거리(distance):

1. 개체들 사이의 거리
2. 변수들 사이의 거리
3. 위의 distance말고 문제에 따라 적합한 거리를 정의한다. 사례연구 참조.

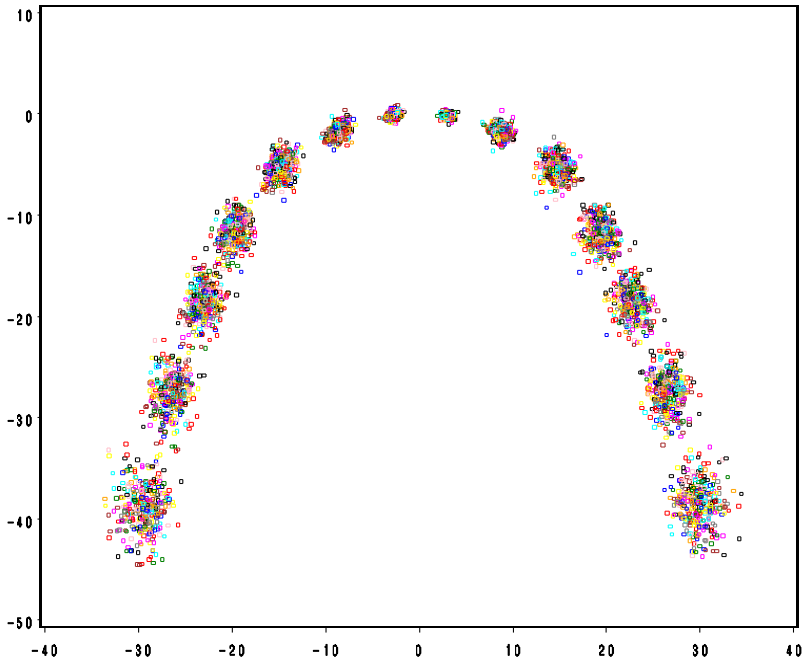
Basics

- **Data:**
  - **306 normal occulusions**
  - **“maxilla” (upper jaw) and “mandible” (lower jaw)**
  - **Example: “mandible”**

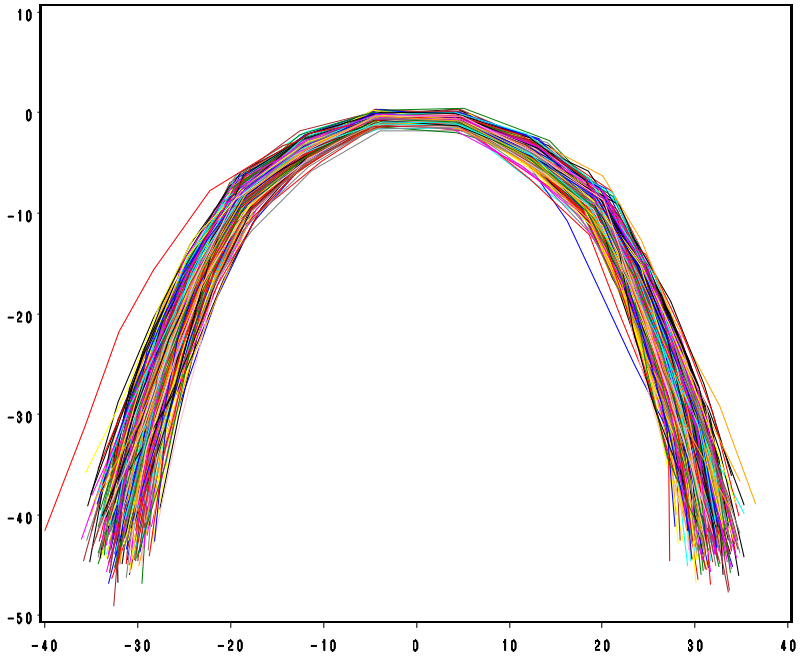



10
0
-10
-20
-30
-40
-50
-40
-30
-20
-10
0
10
20
30
40

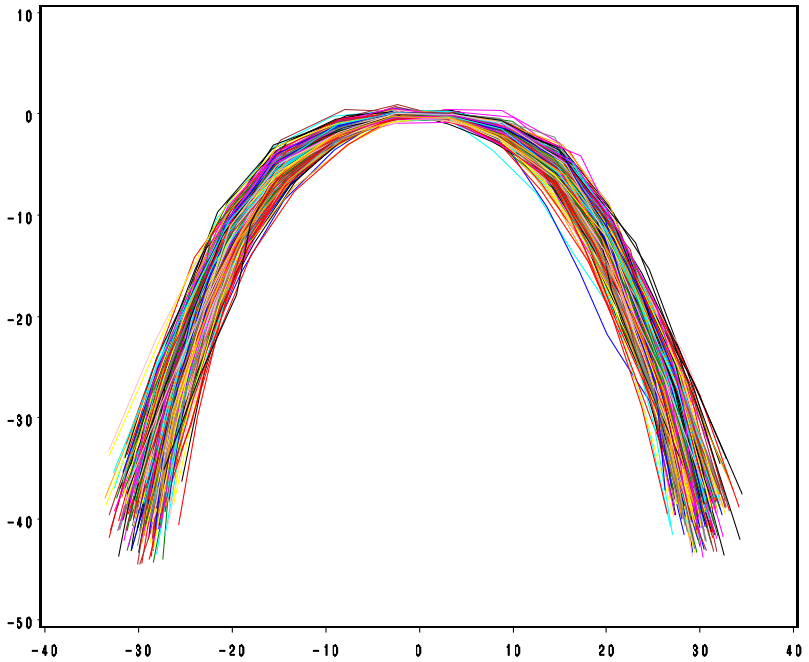
상악


10
0
-10
-20
-30
-40
-50
-40
-30
-20
-10
0
10
20
30
40

하악


10
0
-10
-20
-30
-40
-50
-40
-30
-20
-10
0
10
20
30
40

상악


10
0
-10
-20
-30
-40
-50
-40
-30
-20
-10
0
10
20
30
40

하악

Basics

- We notate as:

$$w_i = (x_{ij}, y_{ij}), \quad i = 1,\mathrm{L}\ 306, \quad j = 1,\mathrm{L}\ ,14$$

- Goal:
  - “Clustering 306 arch forms” and
  - “Find a standard arch form for each cluster”

w_i

Classical Approach

- Treat w_i as a 28 dimension vector with Euclidian distance

$$d(w_i, w_k) = \sqrt{\sum_{j=1}^{14} \left[(x_{ij} - x_{kj})^2 + (y_{ij} - y_{kj})^2 \right]}$$

- Then use

  hierarchical clustering
  VQ (k-means clustering): medical imaging people
  PAM etc.

- Weak point:

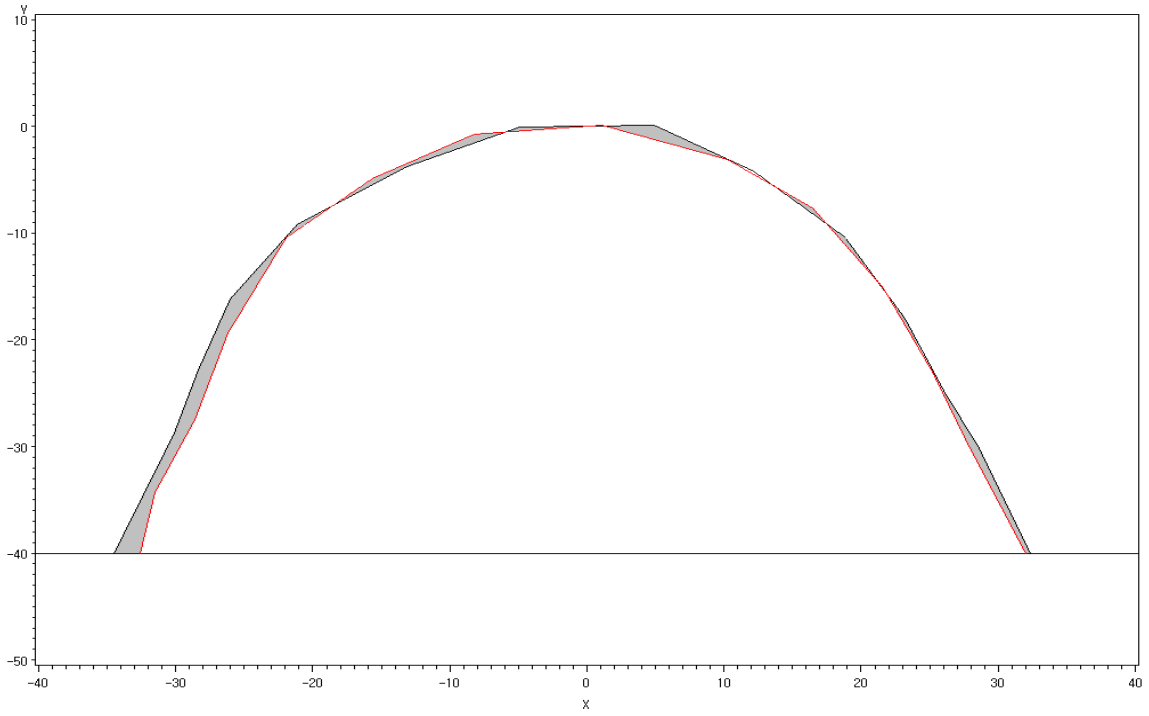
  Practical meaning of Euclidian distance
  Measurement errors and registration problem

Outline of Suggested Method

- **Rotation-Shift Invariant Distance Measure**
- **Clustering with PAM (Partition Around Medoids)**
- **Find the a standard arch form of each cluster using**
  symmetric cubic smoothing spline

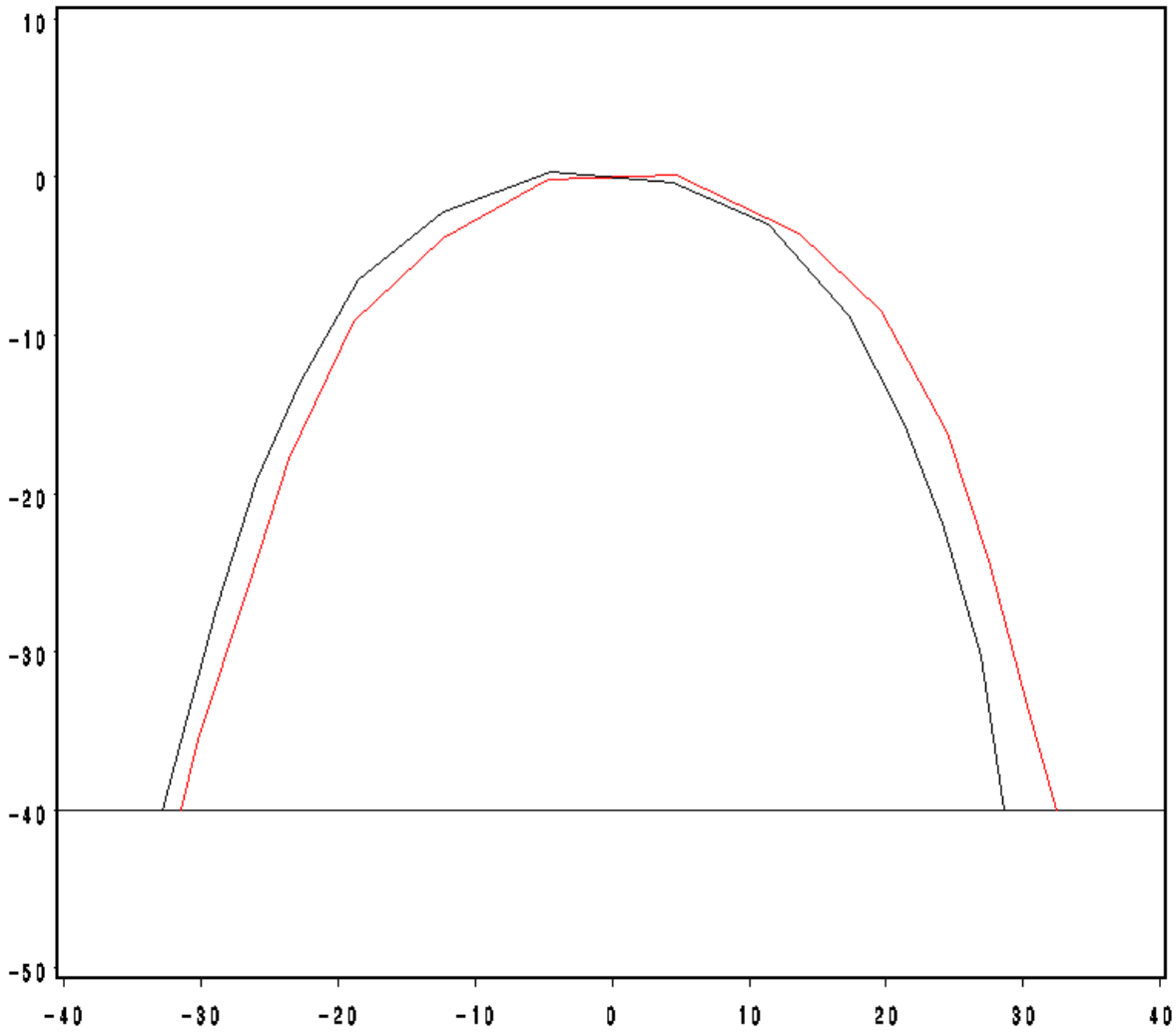
Rotation-Shift Invariant Distance Measure

- Measurements in each observed arch form.
  - shift variation
  - rotation variation
- Fix one (piecewisely connected) arch form, and then apply **shift and rotation transformation** to the second curve
- **Minimize the area between two arch forms**



Distance example : subject 95 vs. subject 158

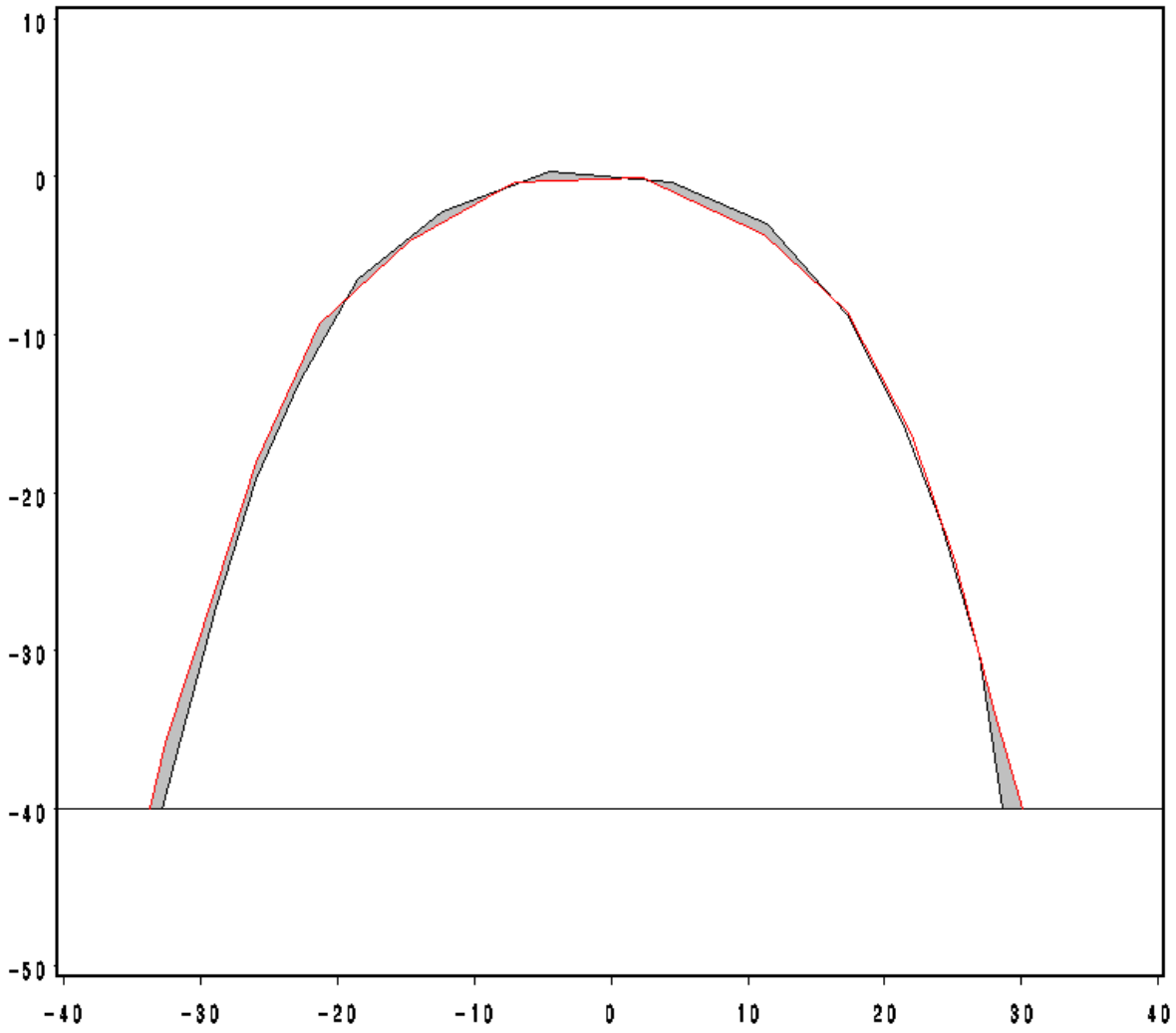
before transformstion



distance=177.70

Example distance : subject 95 vs. subject 158

after transformstion



distance=46.86

Clustering with PAM (Partition around Medoid = median을 이용한 K-means 방법)

pam {cluster} R Documentation
Partitioning Around Medoids
Partitioning (clustering) of the data into k clusters "around medoids", a more robust version of K-means.

Usage
```
pam(x, k, diss = inherits(x, "dist"), metric = "euclidean",
    medoids = NULL, stand = FALSE, cluster.only = FALSE,
    do.swap = TRUE,
    keep.diss = !diss && !cluster.only && n < 100,
    keep.data = !diss && !cluster.only,
    pamonce = FALSE, trace.lev = 0)
```

- **Suggested by Rousseeuw (1987)**
- **Similar to K-means**
- **PAM finds the representative subject of each cluster which minimizes the distance to all other subjects in the same cluster**

  "medoid" = the representative subject

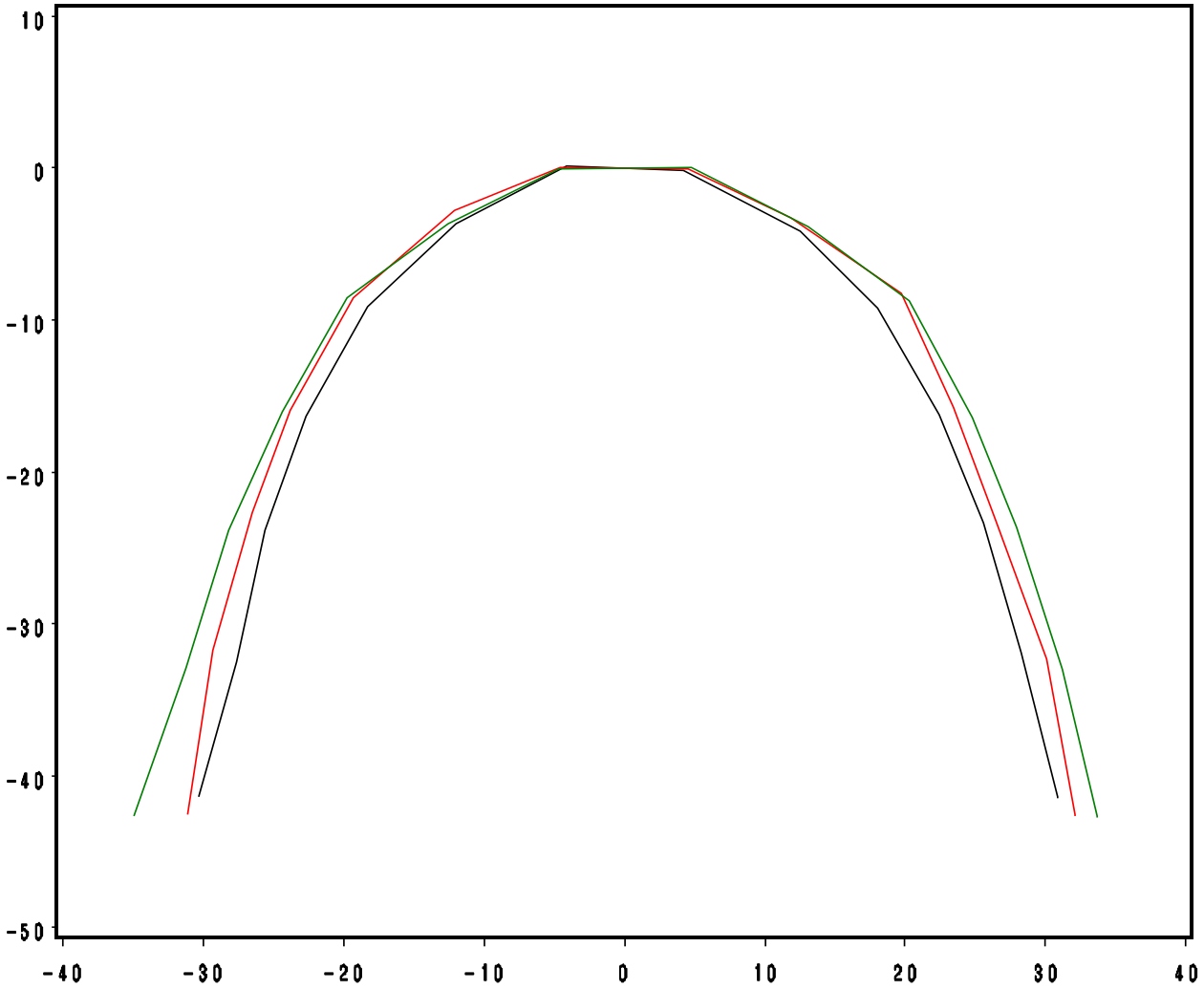
- **Robust to outlier but computationally expensive**
- **Cluster number by average silhouette width**

Cluster number

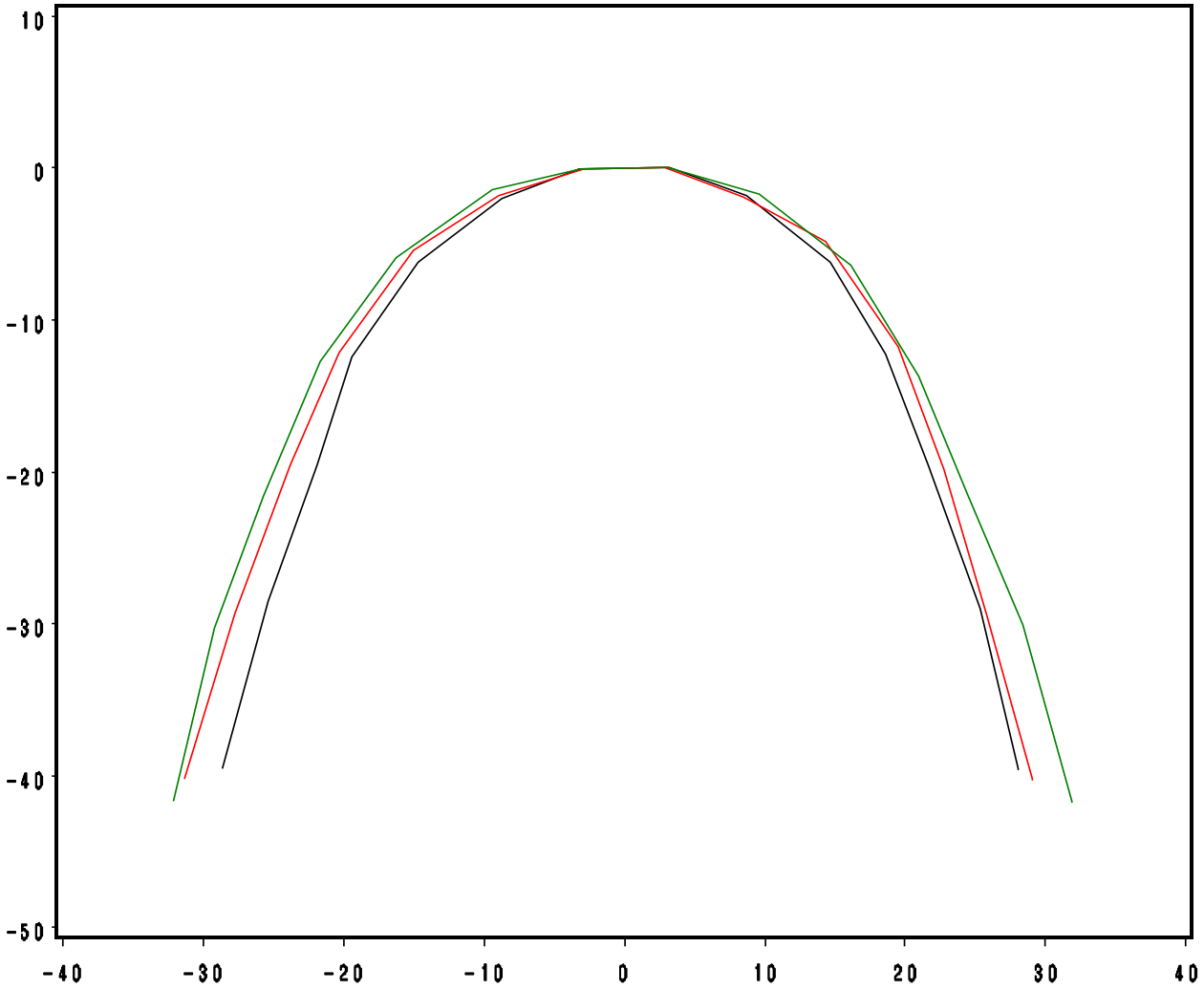
- choose number of clusters=3
  - reasonable size of patients
  - average silhouette width < 0.25 for number of clusters > 3
- Use average silhouette width
  - “high average silhouette width” implies “good cluster”
  - Kaufman과 Rousseeuw (1990) :
    average silhouette width should be larger than 0.25

Number of Clusters	Maxilla	Mandible
2	0.412	0.411
3	0.322*	0.316*
4	0.250	0.252
5	0.237	0.211
6	0.210	0.221

Representatives of maxillary teeth



Representatives of mandibular teeth



Maxilla vs Mandible

- High association between maxilla and mandible clusters
- Remove off-diagonal subjects for standard arch forms.

		mandible			
		narrow	middle	wide	합
maxilar	narrow	60	33	3	96
	middle	9	90	22	121
	wide	0	14	75	89
	합	69	137	100	306

Smoothing

- **Symmetric cubic spline with symmetric restiction at x=0.**

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \gamma_3 x_+^3$$

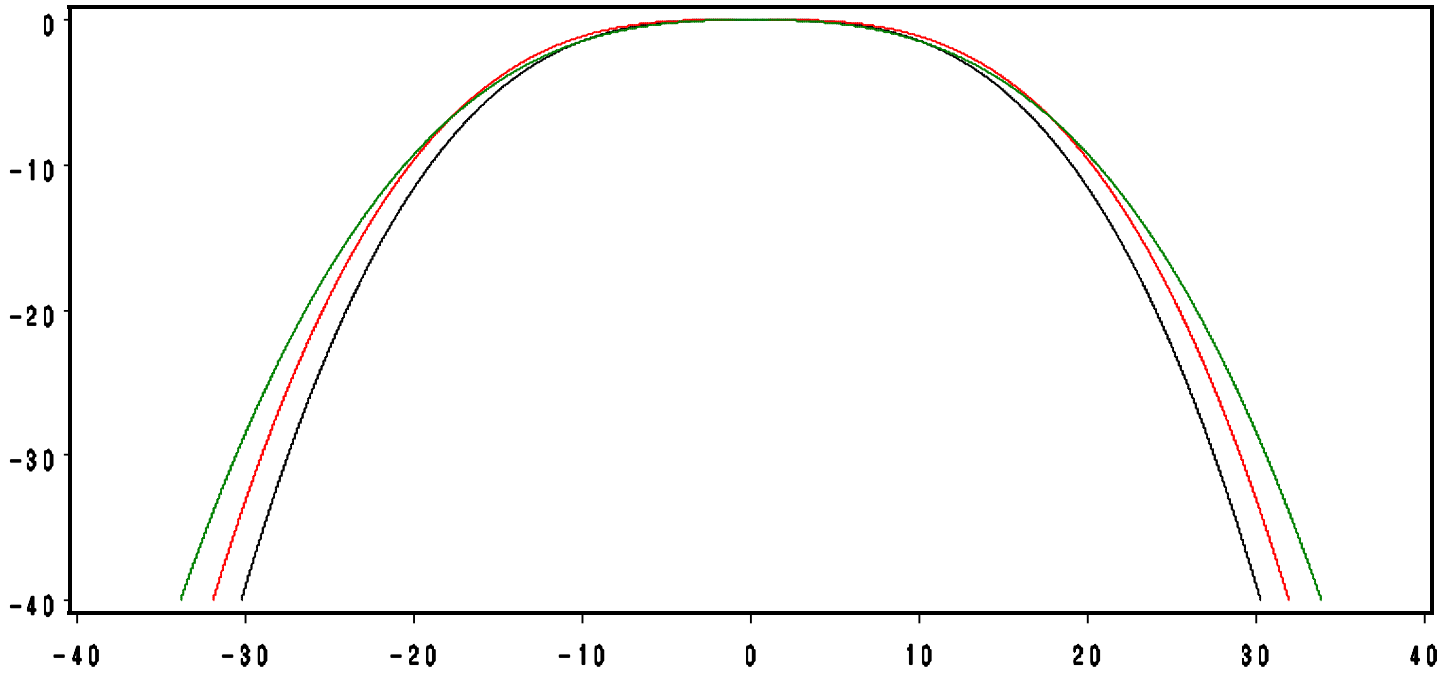
$$x_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$\beta_1 = 0,\ 2\beta_3 + \gamma_3 = 0.$$

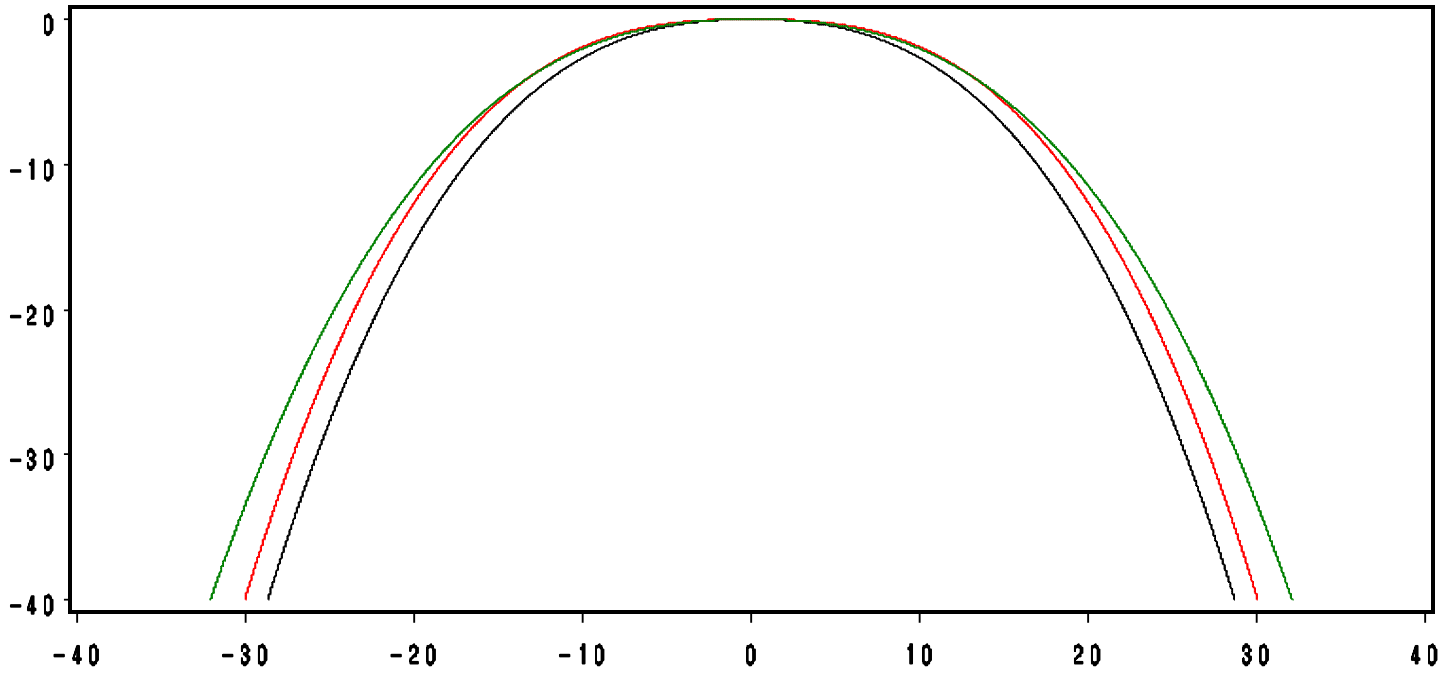
- **For each cluster, estimate parameter** $(\beta_0, \beta_1, \beta_2, \beta_3, \gamma_3)$ **to minimize**

$$\sum \int | g_i(x) - f(x; \beta_0, \beta_1, \beta_2, \beta_3, \gamma_3) | \, dx$$

Standard arch forms of maxilla



Standard arch forms of Mandible



Assignment rule for new individuals

- Compute shift-rotation invariant distance between
  newly observed piecewisely connected arch form
  and each representative arch form

- New individual is assigned to the cluster with minimum distance

그 밖의 잘 알려진 군집화 방법

Model based clustering (mclust):

A quick tour of m-clustering:
https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

R package: "mclust"
https://cran.r-project.org/web/packages/mclust/mclust.pdf

참고문헌

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (p. 6). New York: springer.
이 책에서 제공하는 그림을 이용하여 슬라이드를 작성하였다.

참고문헌

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (p. 6). New York: springer.
이 책에서 제공하는 그림을 이용하여 슬라이드를 작성하였다.