1장 다변량 데이터 이해

덕성여자대학교 정보통계학과 김 재희

덕성여자대학교
DUKSUNG WOMEN'S UNIVERSITY

1. 다변량 데이터(multivariate data)

1.1 다변량 데이터 예

[표 1.1] 다변량 데이터 구성 예

표본단위	다변량 데이터를 구성하는 변수들
학생	한 과목 강좌에서 3 번 본 시험의 점수
학생	수강한 여러 과목들 각각의 점수
회사	광고비, 노동자 수, 자산, 부채액, 매출액, 순이익률
은행대출자	연수입, 교육정도, 거주 년 수, 예금 액수, 부채
새	주요 뼈들의 길이, 둘레
사람	키, 몸무게, 지방의 비중, 분당 심장 박동 수
두개골	길이, 둘레
유치원생	나이, 놀이 시간, 집중 시간
신용카드 사용자	나이, 학력 정도, 연봉, 연체액, 사용액, 사용 회수
한우	육량등급, 나이, 등심단면적, 도체중

1.2 다변량 데이터 구조와 기술통계량

n개 개체에 대해 p개 변수 측정값인 전체 데이터

$$\begin{array}{c|cccccc}
 & \text{변수}1 & \text{변수}2 & \cdots & \text{변수}j & \cdots & \text{변수}p \\
\hline
\text{개체}1 & X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\
\text{개체}2 & X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2p} \\
\vdots & \vdots & & \vdots & & \vdots & \\
\text{개체}i & X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\
 & \vdots & \vdots & & \vdots & & \vdots \\
\text{개체}n & X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np}
\end{array}$$

n : 표본단위 수, 케이스 수 또는 개체

X_{ij} : i 번째 개체에 대한 j 번째 변수의 관측값

열벡터(column vector) $\boldsymbol{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad i = 1, 2, \ldots, n$

표본평균벡터 $\overline{\boldsymbol{X}} = \begin{pmatrix} \overline{X_1} \\ \overline{X_2} \\ \vdots \\ \overline{X_p} \end{pmatrix}$

표본공분산행렬 $\boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{pmatrix}$ 표본상관행렬 $R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$

•j 번째 변수의 표본평균(sample mean): $\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad j = 1,2,...,p$.

•j번째 변수의 표본분산(sample variance): $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2 = s_{jj}$

•j번째 변수의 표본표준편차(sample standard deviation): $\sqrt{s_{jj}}, \quad j = 1,2,...,p$.

•j번째 변수와 k 번째 변수의 표본공분산(sample covariance):

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k}), \quad j = 1,2,\cdots,p, \quad k = 1,2,\cdots,p$$

•j번째 변수와 k번째 변수의 표본상관계수(sample correlation coefficient):

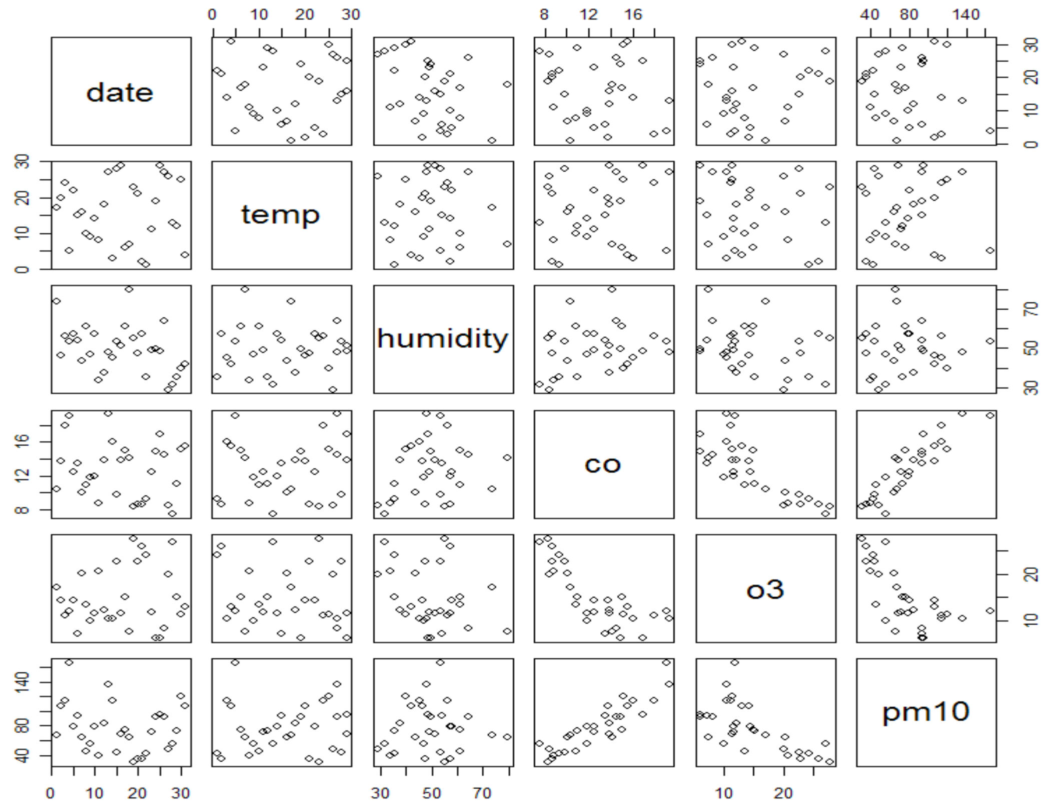
$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})}{\sqrt{\sum_{i=1}^{n}(X_{ij} - \overline{X_k})^2}\sqrt{\sum_{i=1}^{n}(X_{ik} - \overline{X_k})^2}}$$

1.3 그림을 통한 다변량 데이터 표현

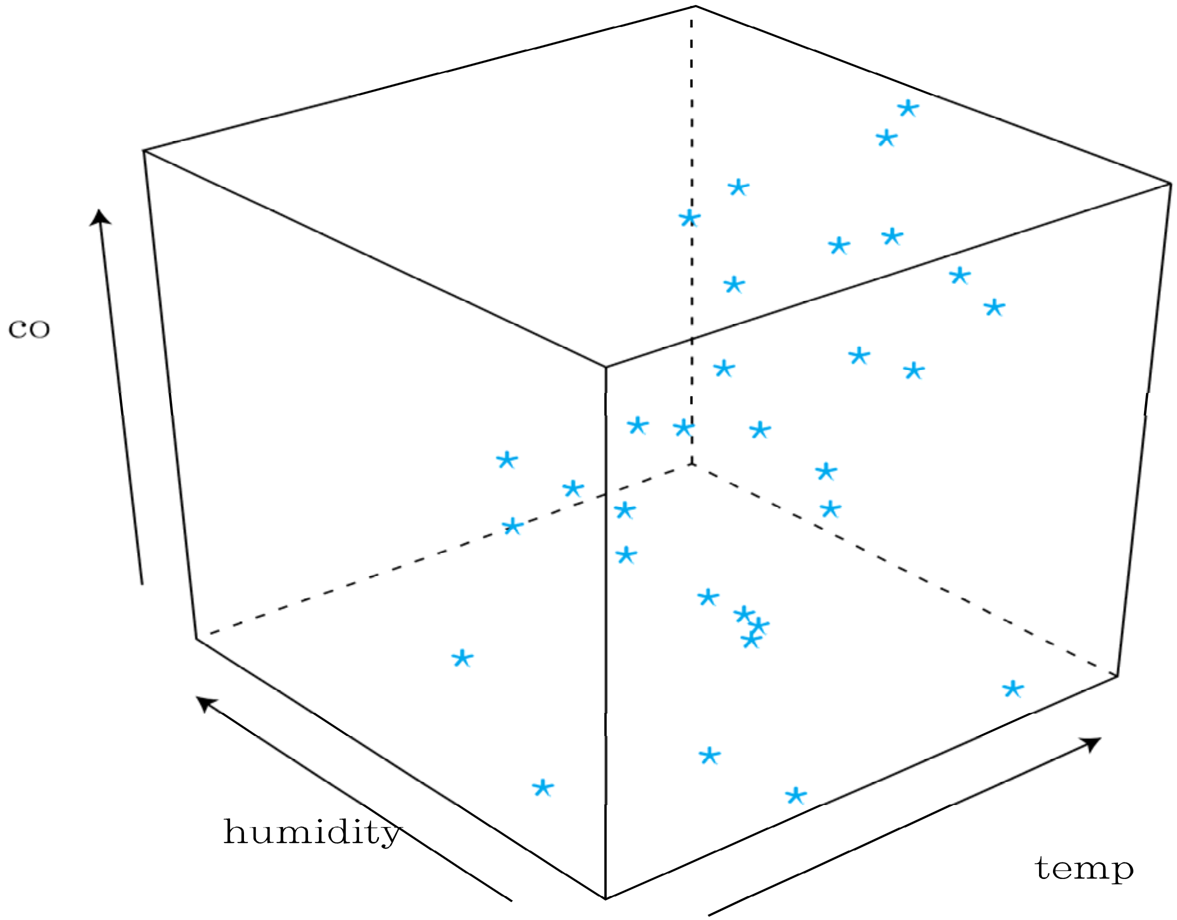
[표 1.2] 1999년 3월 서울의 기상과 대기오염 자료

측정량	날짜	온도	습도	CO	O3	PM10
측정단위		℃	%	ppm	ppb	$\mu g/m^3$
변수명	date	temp	humidity	co	o3	pm10
	1	5.6	73.8	10.40	17.10	67.59
	2	6.2	46.5	13.81	14.37	106.80
	3	7.7	56.4	17.97	11.24	114.40
	4	11.1	53.3	19.10	12.06	165.98
	5	6.7	57.6	12.44	14.46	79.86
	6	4.9	54.0	13.56	7.26	93.34
	7	5.2	43.4	10.04	20.30	63.92
	8	3.3	61.0	10.91	13.55	45.74
	9	3.2	47.0	11.87	10.02	55.61
	10	4.1	57.3	11.91	11.62	79.56
		(중략)				
	31	10.8	42.0	15.54	13.07	106.59

1.3.1 산점도행렬과 3차원 산점도

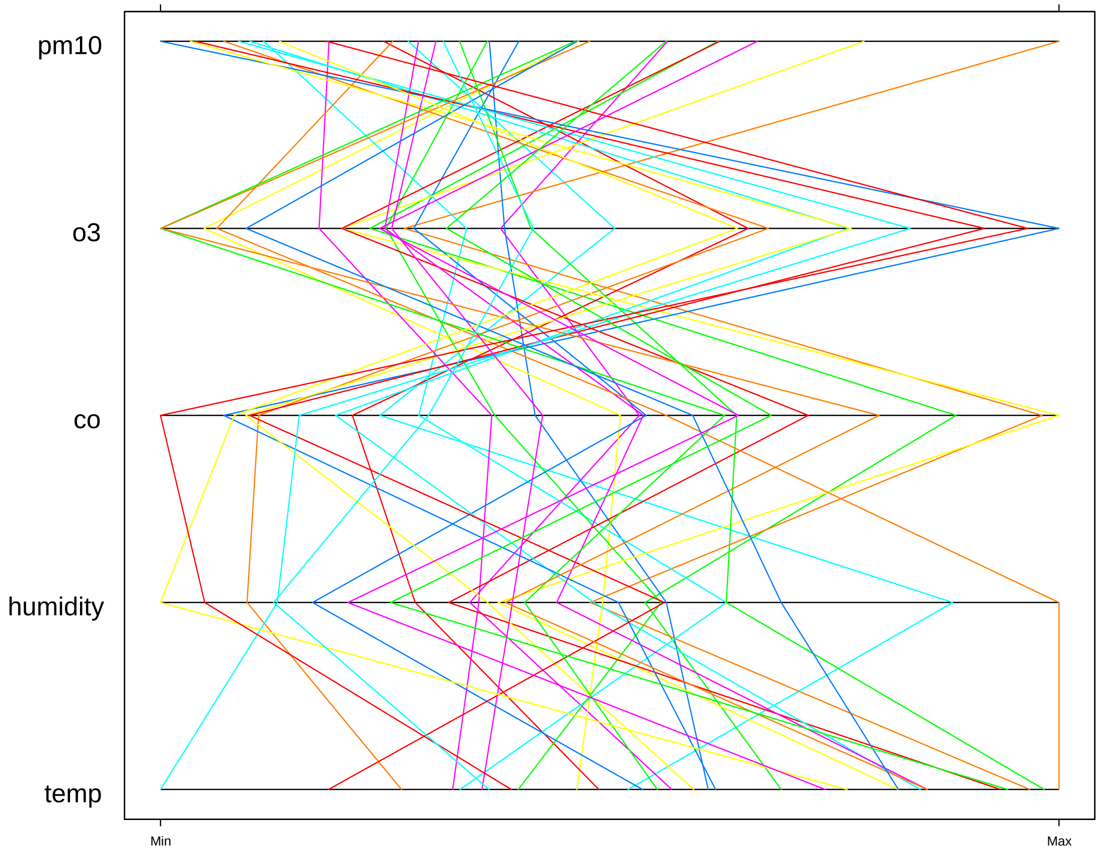


[그림 1.1] 기상과 오염자료에 대한 산점도행렬

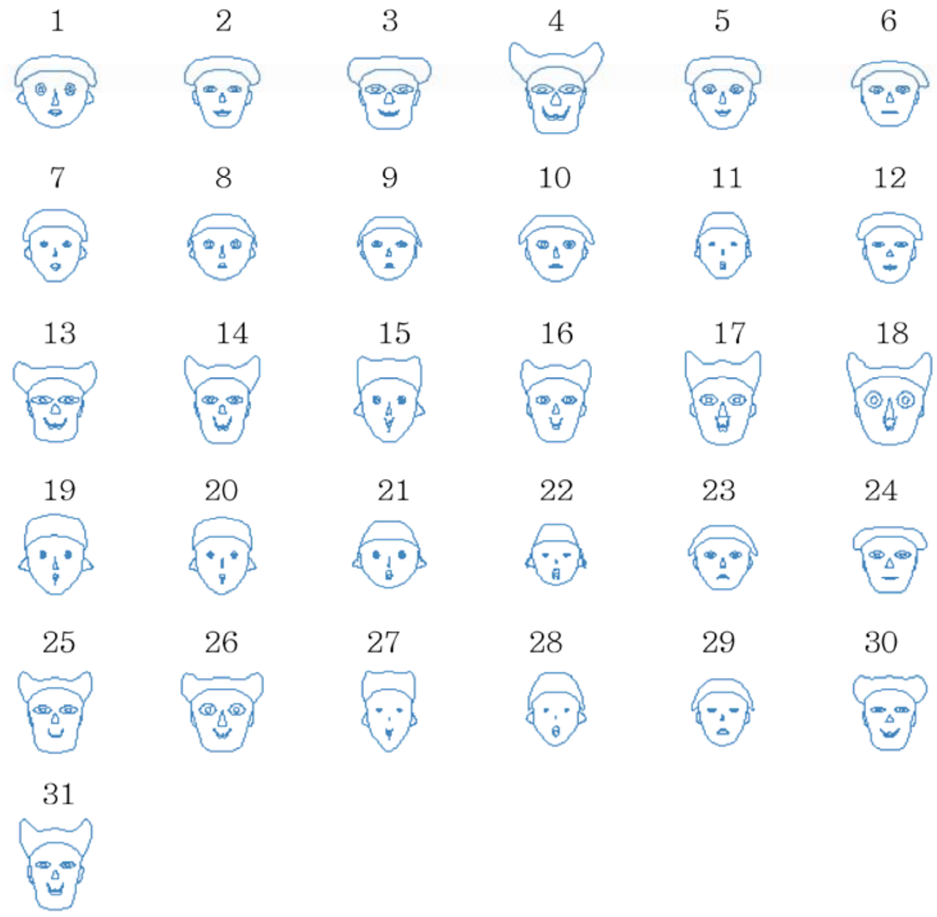


[그림 1.2] 3차원 산점도

1.3.2 Trellis 그래프(Parallel 그래프)



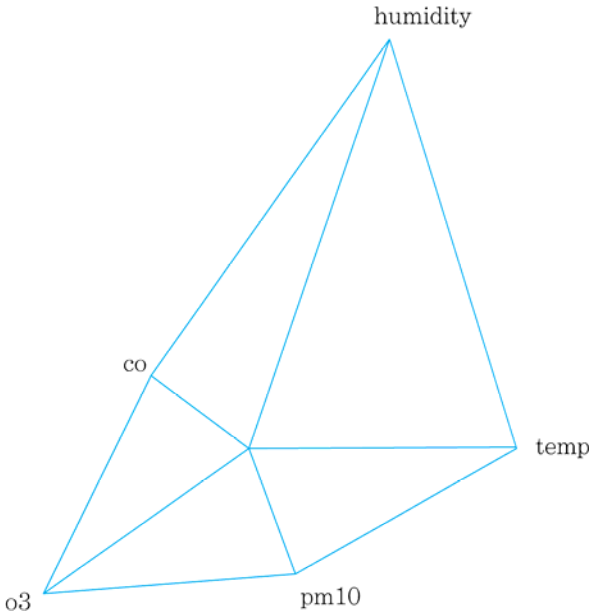
1.3.3 Chernoff의 얼굴그림



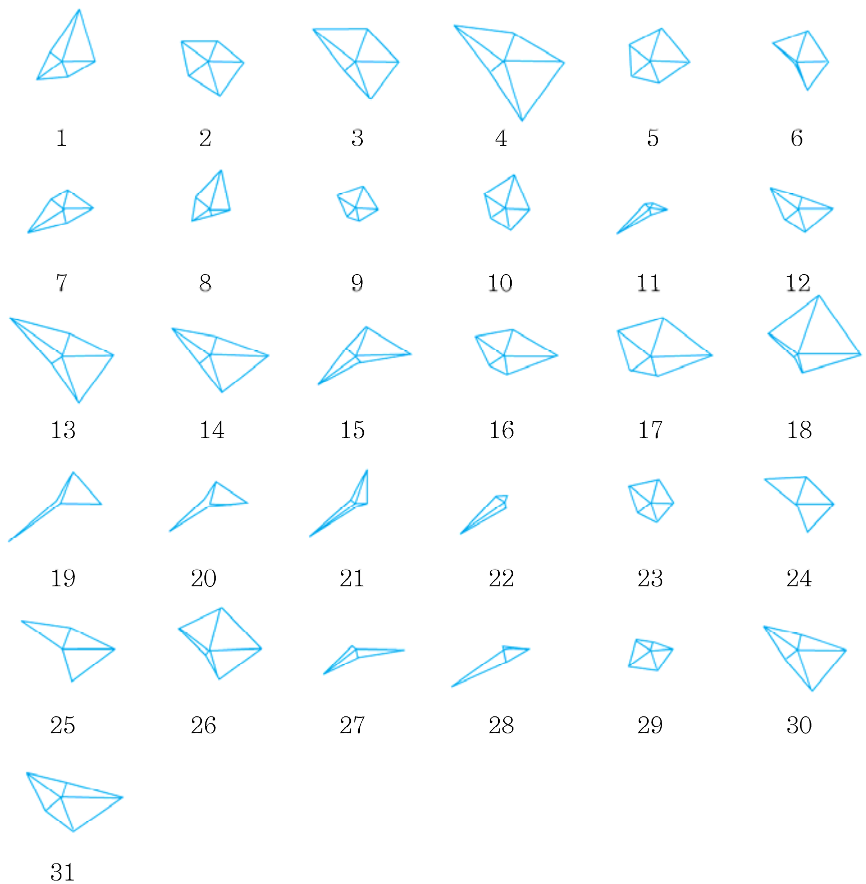
1. 얼굴 길이(height of face)
2. 얼굴 너비(width of face)
3. 얼굴 모양(shape of face)
4. 입의 높이(height of mouth)
5. 입의 너비(width of mouth)
6. 미소 곡선(curve of smile)
7. 눈의 위치(height of eyes)
8. 눈간 거리(width of eyes)
9. 머리 높이(height of hair)
10. 머리 너비(width of hair)
11. 머리 스타일(styling of hair)
12. 코의 높이(height of nose)
13. 코의 너비(width of nose)
14. 귀의 너비(width of ears)
15. 귀의 높이(height of ears)

[그림 1.4] Chernoff의 얼굴그림

1.3.4 별그림



[그림 1.5] 별의 축 표시



[그림 1.6] 별그림

1.4 거리 측도

i번째 관측벡터 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, ..., X_{ip})'$ 와 k번째 관측벡터 $\boldsymbol{X}_k = (X_{k1}, X_{k2}, ..., X_{kp})'$

1.4.1 유클리드 거리

$$d_{ik} = \left[\sum_{j=1}^{p} (X_{ij} - X_{kj})^2 \right]^{1/2} = \left[(\boldsymbol{X}_i - \boldsymbol{X}_k)'(\boldsymbol{X}_i - \boldsymbol{X}_k) \right]^{1/2}$$

1.4.2 표준화 거리(통계적 거리): 분산 고려

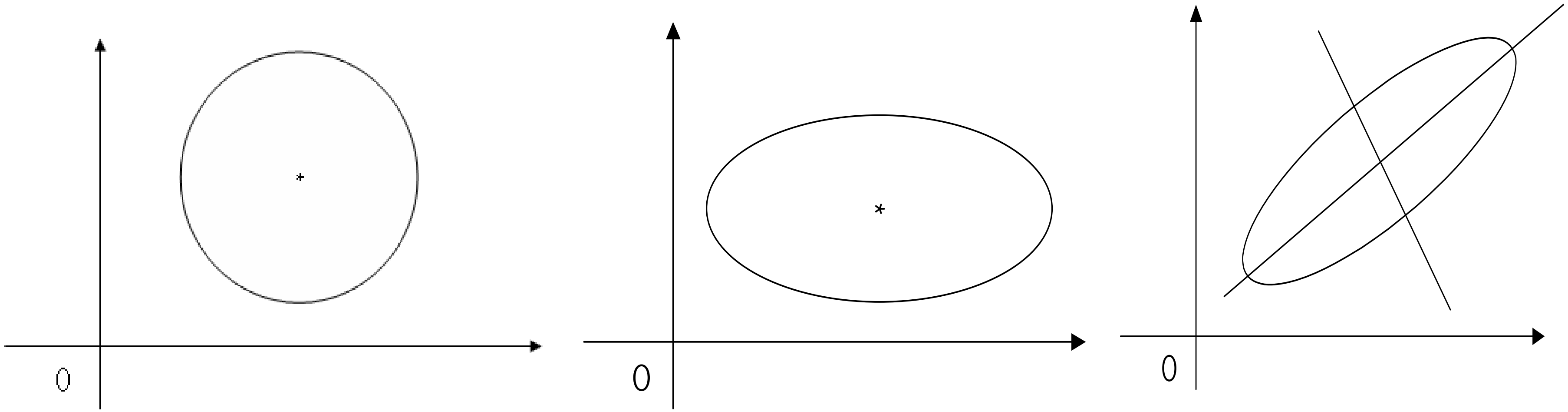
$$d_{ik} = \left[\sum_{j=1}^{p} \frac{(X_{ij} - X_{kj})^2}{s_{jj}} \right]^{1/2} = \left[(\boldsymbol{X}_i - \boldsymbol{X}_k)' \boldsymbol{D}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_k) \right]^{1/2}$$

여기서 $\boldsymbol{D} = diag\{s_{11}, s_{22}, \cdots, s_{pp}\}$, s_{jj}는 j번째 변수의 분산

1.4.3 마할라노비스 거리 : 공분산 고려

$$d_{ik} = \left[(\boldsymbol{X}_i - \boldsymbol{X}_k)' \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_k) \right]^{1/2}$$

<그림> 평균벡터로부터 거리가 같은 벡터들의 집합



(1) 유클리드 거리　　(2) 표준화 거리　　(3) 마할라노비스 거리

1.5 R을 이용한 기초 통계량

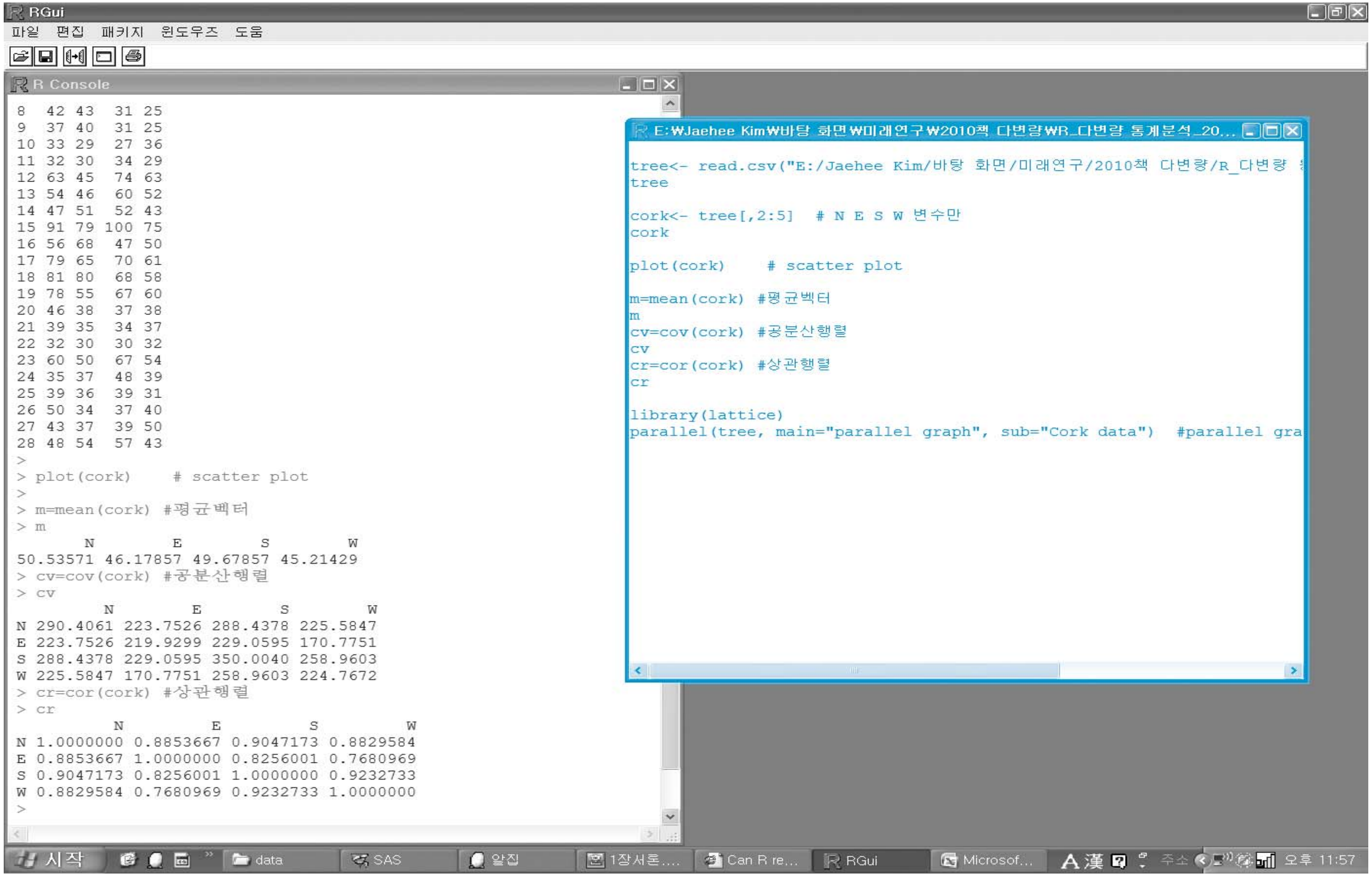
≪예제 1.3≫ 28 그루의 나무를 대상으로 북쪽(N), 동쪽(E), 남쪽(S), 서쪽(W) 방향으로 형성된 코르크 보오링의 깊이를 측정하여 [표 1.4]의 자료를 얻었다. 다변량 데이터에 대한 기초적인 통계량과 그래프 표현을 해보고자 한다.

▶표 1.4 코르크 방향 자료

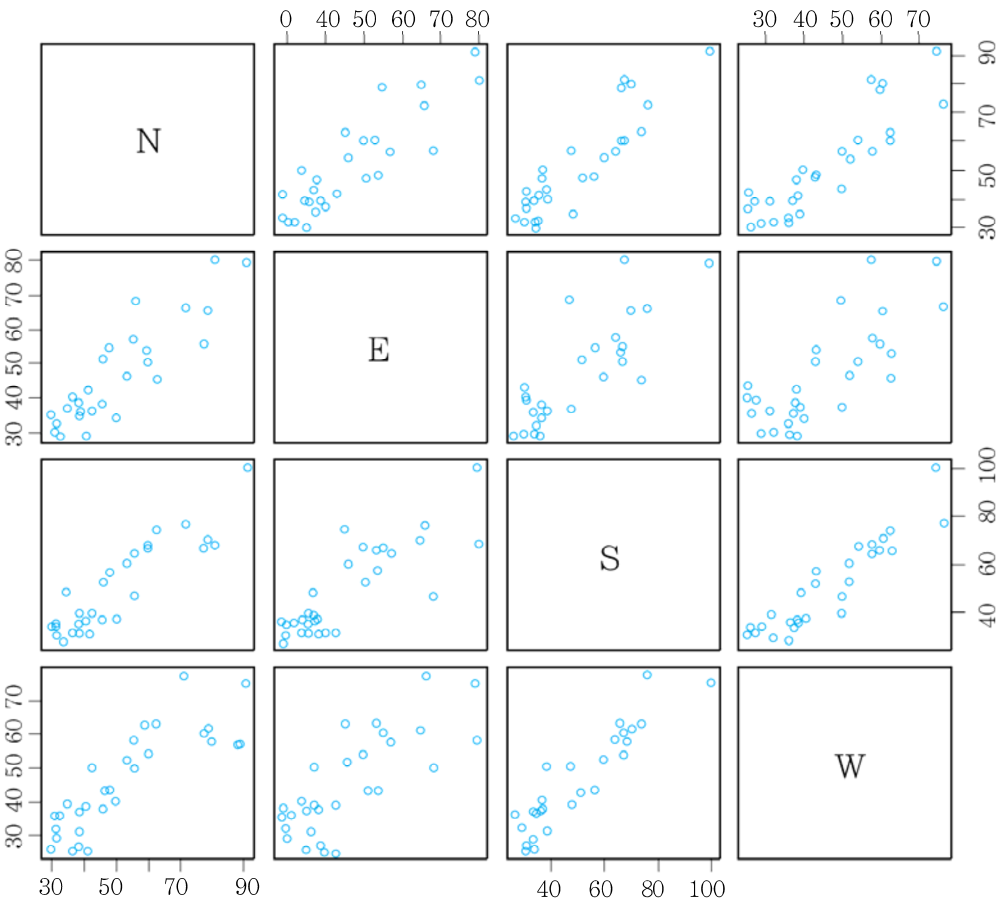
Tree	N	E	S	W	Tree	N	E	S	W
1	72	66	76	77	15	91	79	100	75
2	60	53	66	63	16	56	68	47	50
3	56	57	64	58	17	79	65	70	61
4	41	29	36	38	18	81	80	68	58
5	32	32	35	36	19	78	55	67	60
6	30	35	34	26	20	46	38	37	38
7	39	39	31	27	21	39	35	34	37
8	42	43	31	25	22	32	30	30	32
9	37	40	31	25	23	60	50	67	54
10	33	29	27	36	24	35	37	48	39
11	32	30	34	29	25	39	36	39	31
12	63	45	74	63	26	50	34	37	40
13	54	46	60	52	27	43	37	39	50
14	47	51	52	43	28	48	54	57	43

[프로그램 1.2] tree.R

```
tree<- read.csv("C:/data/tree.csv", header=T)
tree
cork<- tree[,2:5]  # N E S W 변수만
cork
plot(cork)         # scatter plot 그림 1.11
m=mean(cork)  # 평균
m
cv=cov(cork)    # 공분산
cv
cr=cor(cork)    # 상관관계
cr
library(lattice)
parallel(tree, main="parallel graph")                    # parallel  그림 1.12
stars(cork, labels = tree[,1], main="star graph")     # 그림 1.13
library(aplpack)
faces(cork, main="face plot for cork")        # face plot 그림 1.14
library(lattice)
cloud(N~ E* W , data=cork)                    # 3차원 산점도 그림 1.15
```



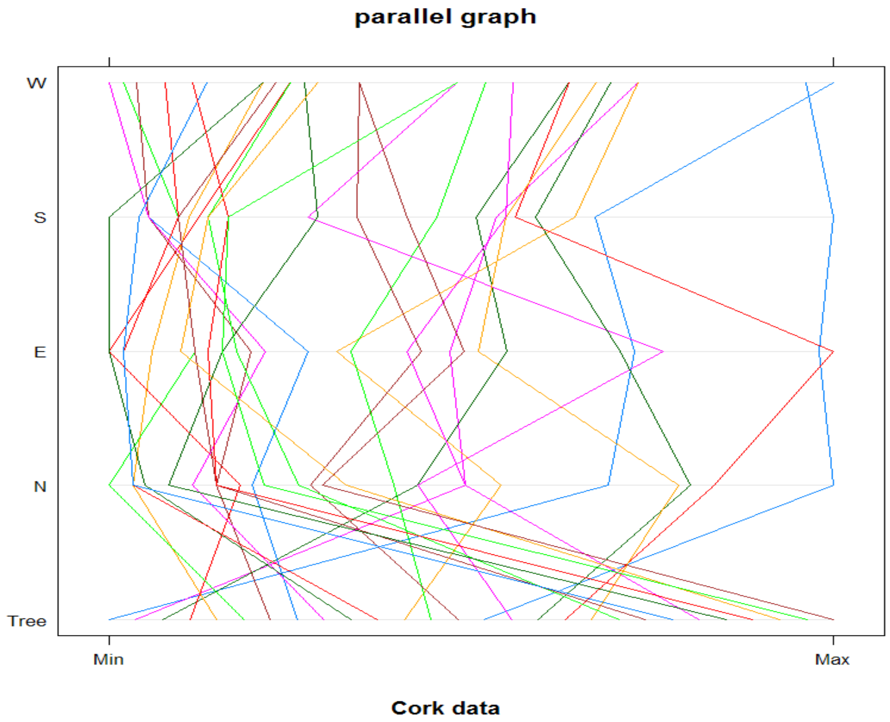
[화면 1.1] R 콘솔과 tree.R 프로그램



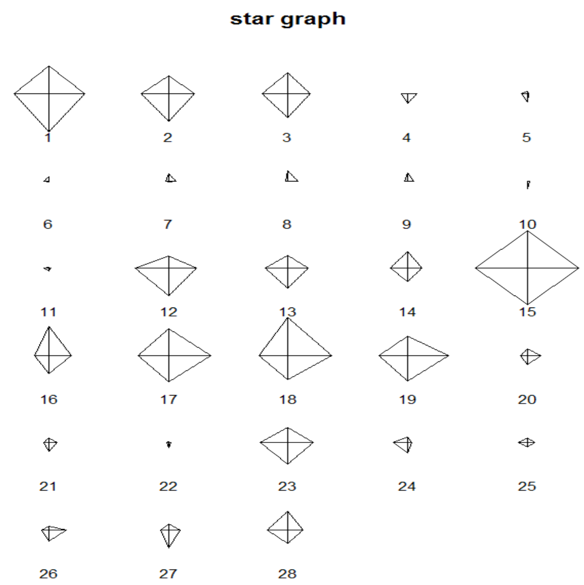
[그림 1.11] 코르크 자료에 대한 산점도행렬

▶표 1.5 코르크 자료에 대한 기술통계량

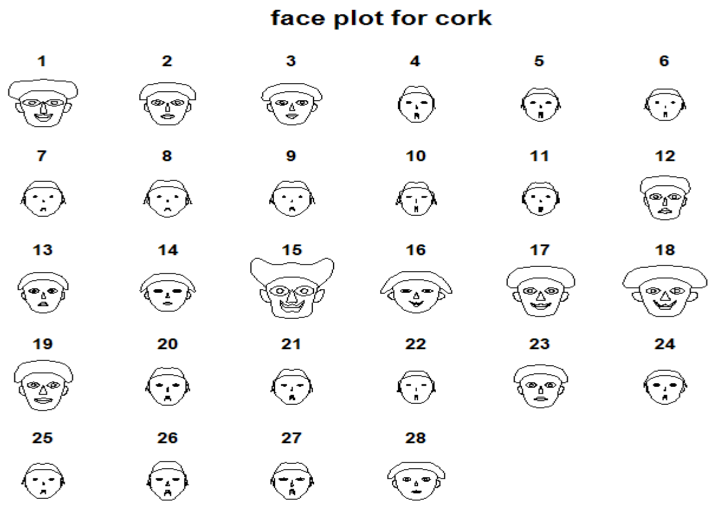
```
> m=mean(cork)  #평균벡터
> m
       N        E        S        W
50.53571 46.17857 49.67857 45.21429
> cv=cov(cork)  #공분산행렬
> cv
         N        E        S        W
N 290.4061 223.7526 288.4378 225.5847
E 223.7526 219.9299 229.0595 170.7751
S 288.4378 229.0595 350.0040 258.9603
W 225.5847 170.7751 258.9603 224.7672
> cr=cor(cork)  #상관행렬
> cr
          N         E         S         W
N 1.0000000 0.8853667 0.9047173 0.8829584
E 0.8853667 1.0000000 0.8256001 0.7680969
S 0.9047173 0.8256001 1.0000000 0.9232733
W 0.8829584 0.7680969 0.9232733 1.0000000
```



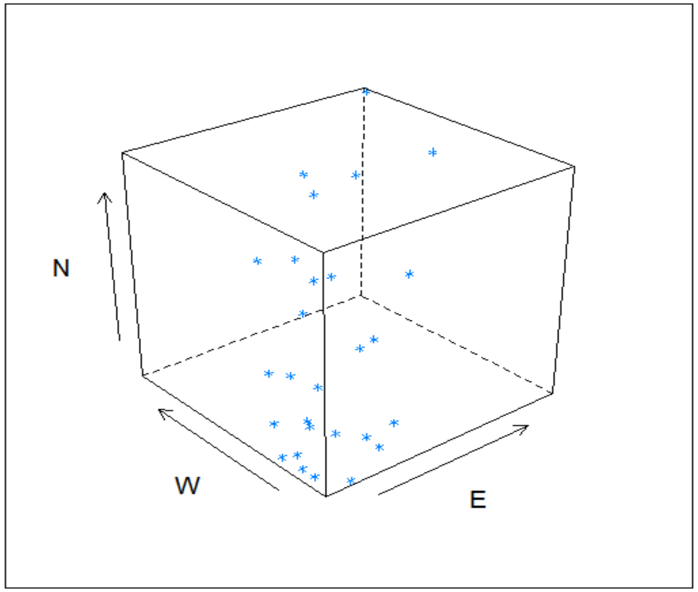
[그림 1.12] 코르크 자료에 대한 평행그림



[그림 1.13] 코르크 자료에 대한 별그림



[그림 1.14] 코르크 자료에 대한 체르노프 얼굴그림



[그림 1.15] 코르크 자료에 대한 3차원 산점도 그림