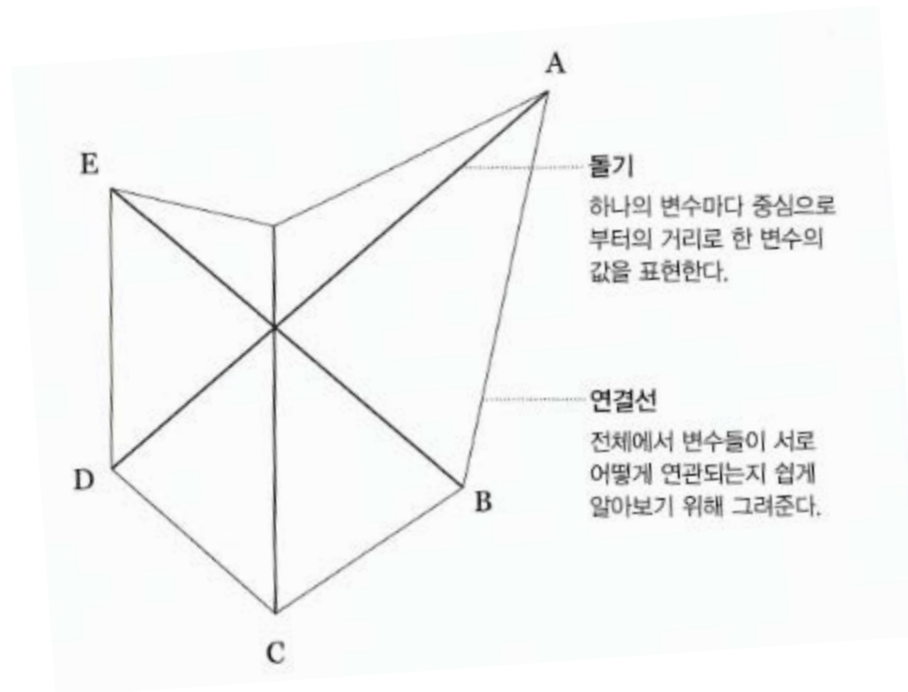


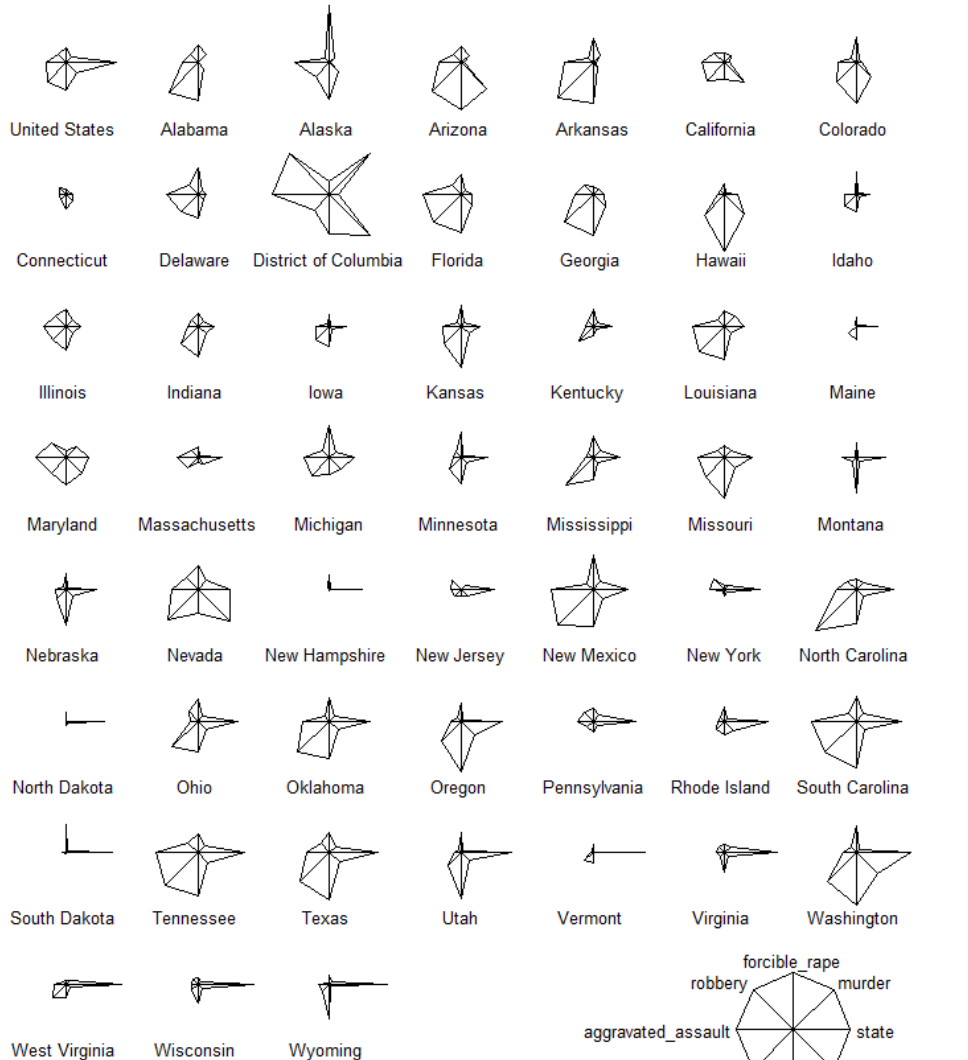
다변량 자료의 탐색

별그림 (Star Chart, Spider Chart)

- 중점: 축이 나타내는 최소값
- 가장 먼 끝은 최대값

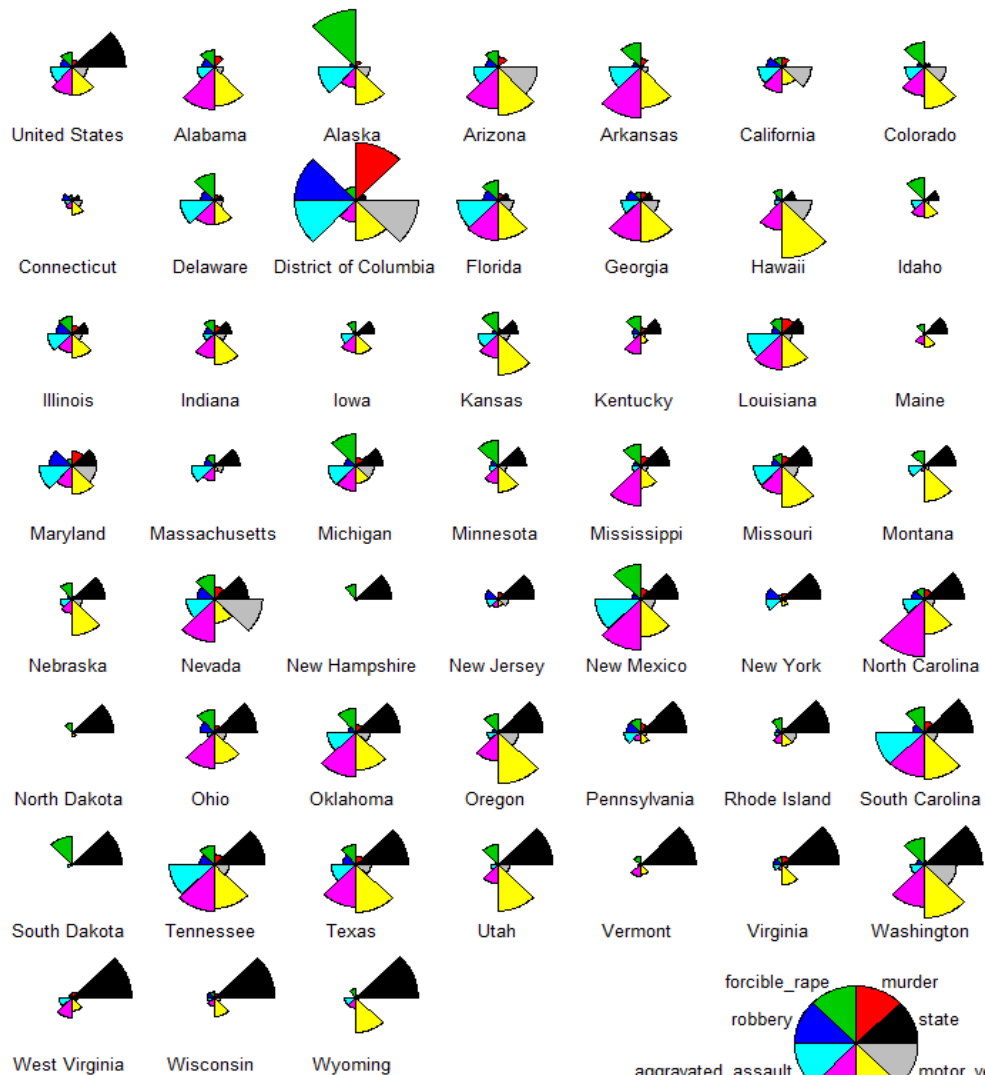


별그림 (Star Chart, Spider Chart)

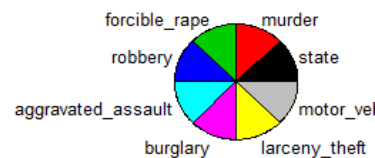


```
crime=read.csv("http://datasets.flowingdata.com
/crimeRatesByState-formatted.csv")
rownames(crime)=crime$state
stars(crime,flip.labels=FALSE)
```

나이팅게일 차트



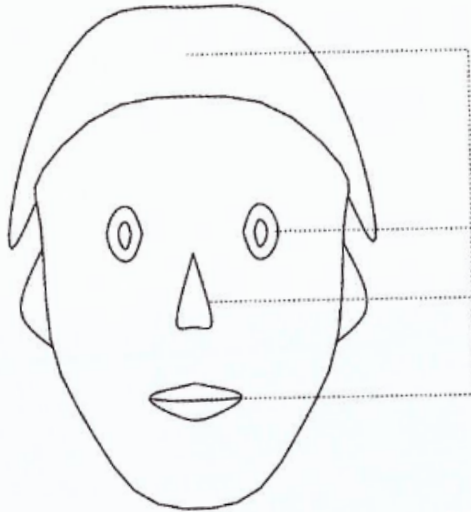
```
stars(crime, flip.labels=FALSE, key.loc=c(15,1.5), draw.segments=TRUE)
```



체르노프 페이스

얼굴

얼굴 전체로서
하나의 대상, 또는
데이터의 한 줄을
나타낸다.



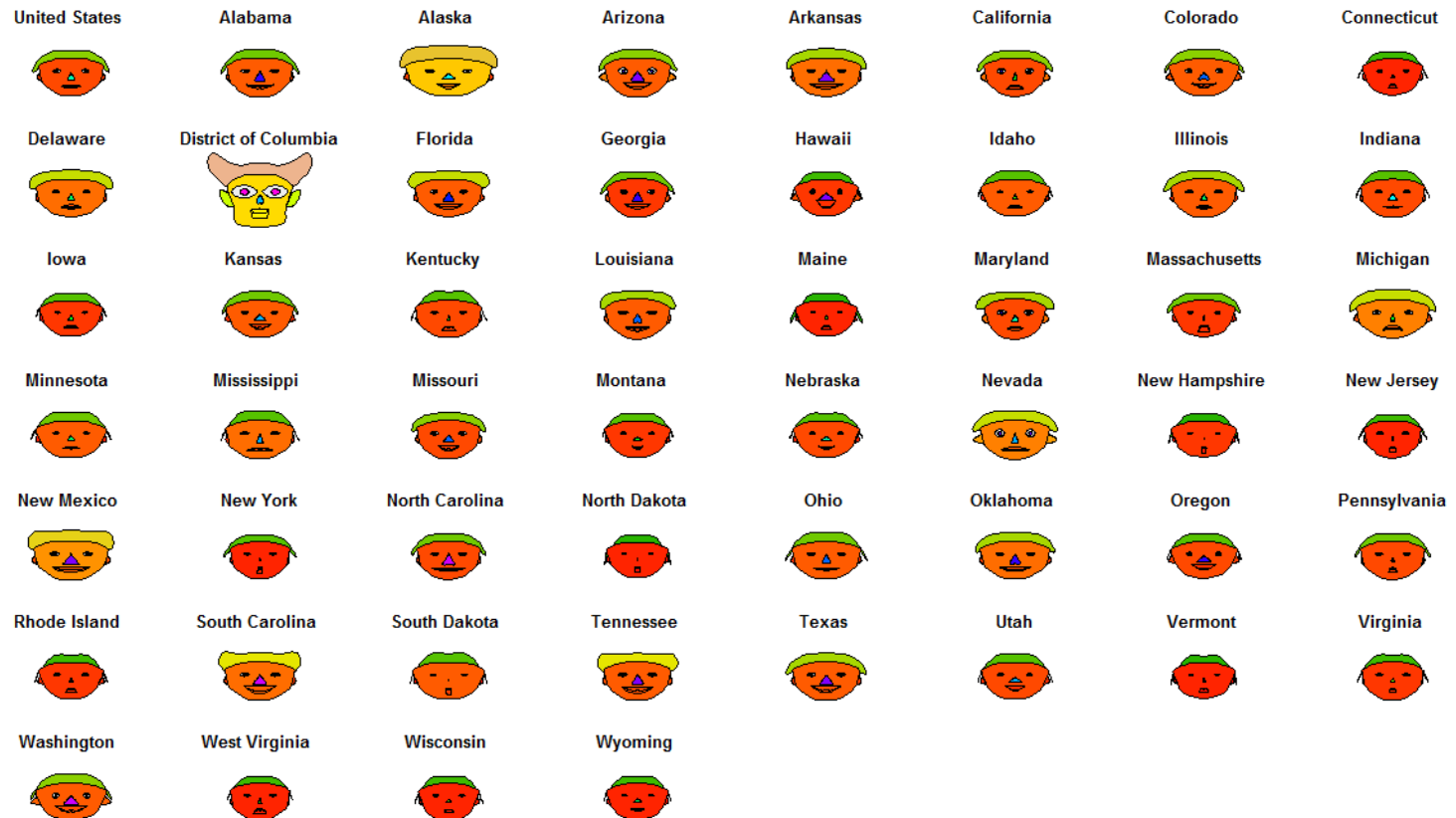
특징

머리 모양의 높이,
눈과 코의 크기, 입술
모양 등이 데이터에
따라 다르게 나타난다.

열 번호에 따라 아래
순서대로 대응되는 특징

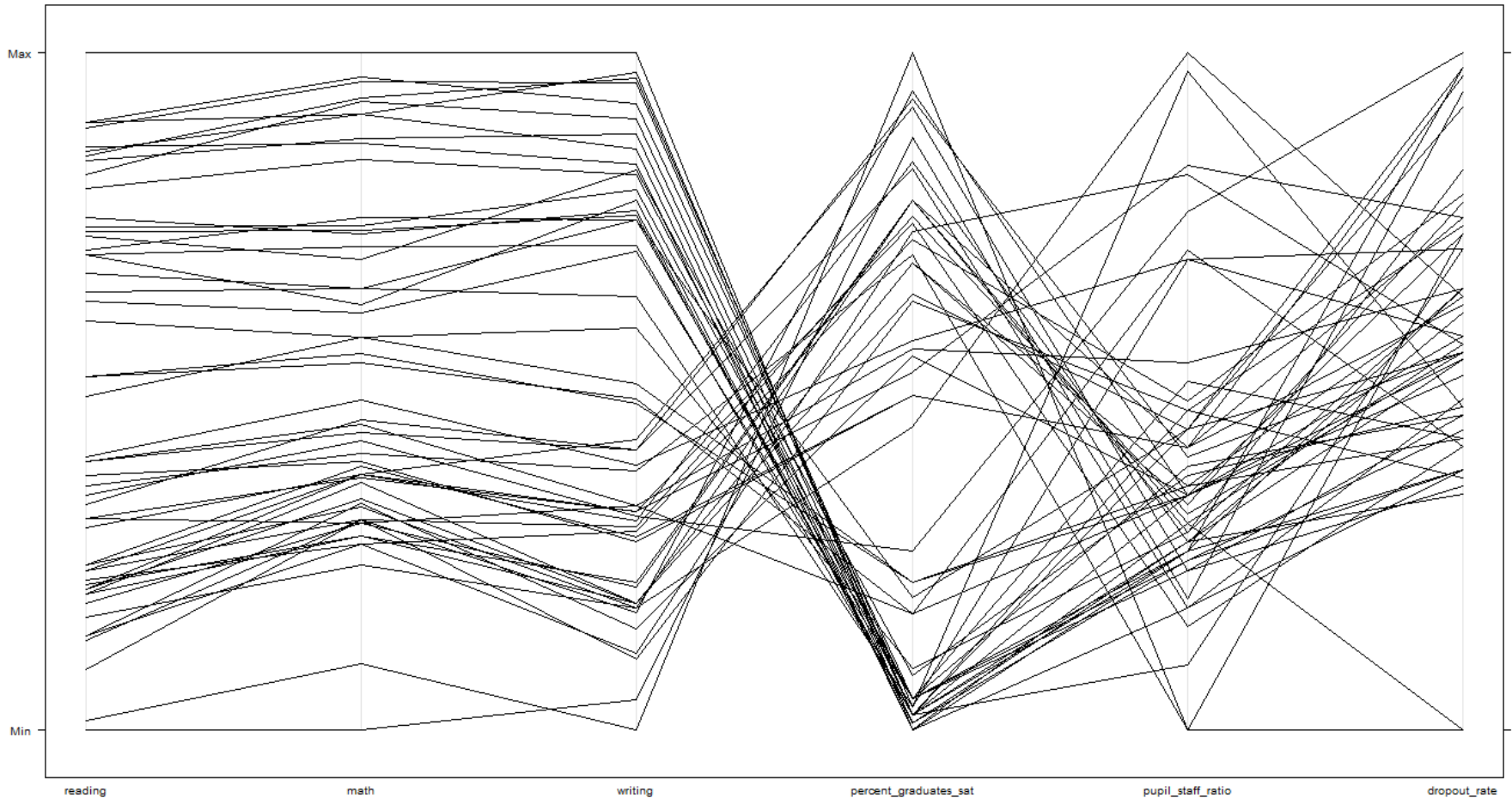
- 얼굴 길이
- 얼굴 너비
- 얼굴 윤곽
- 입의 높이
- 입의 너비
- 입모양
- 눈의 높이
- 눈의 너비
- 머리카락 높이
- 머리카락 너비
- 머리카락 모양
- 코의 높이
- 코의 너비
- 귀의 너비
- 귀의 높이

체르노프 페이스



```
library(aplpack)  
faces(crime[,2:8])
```

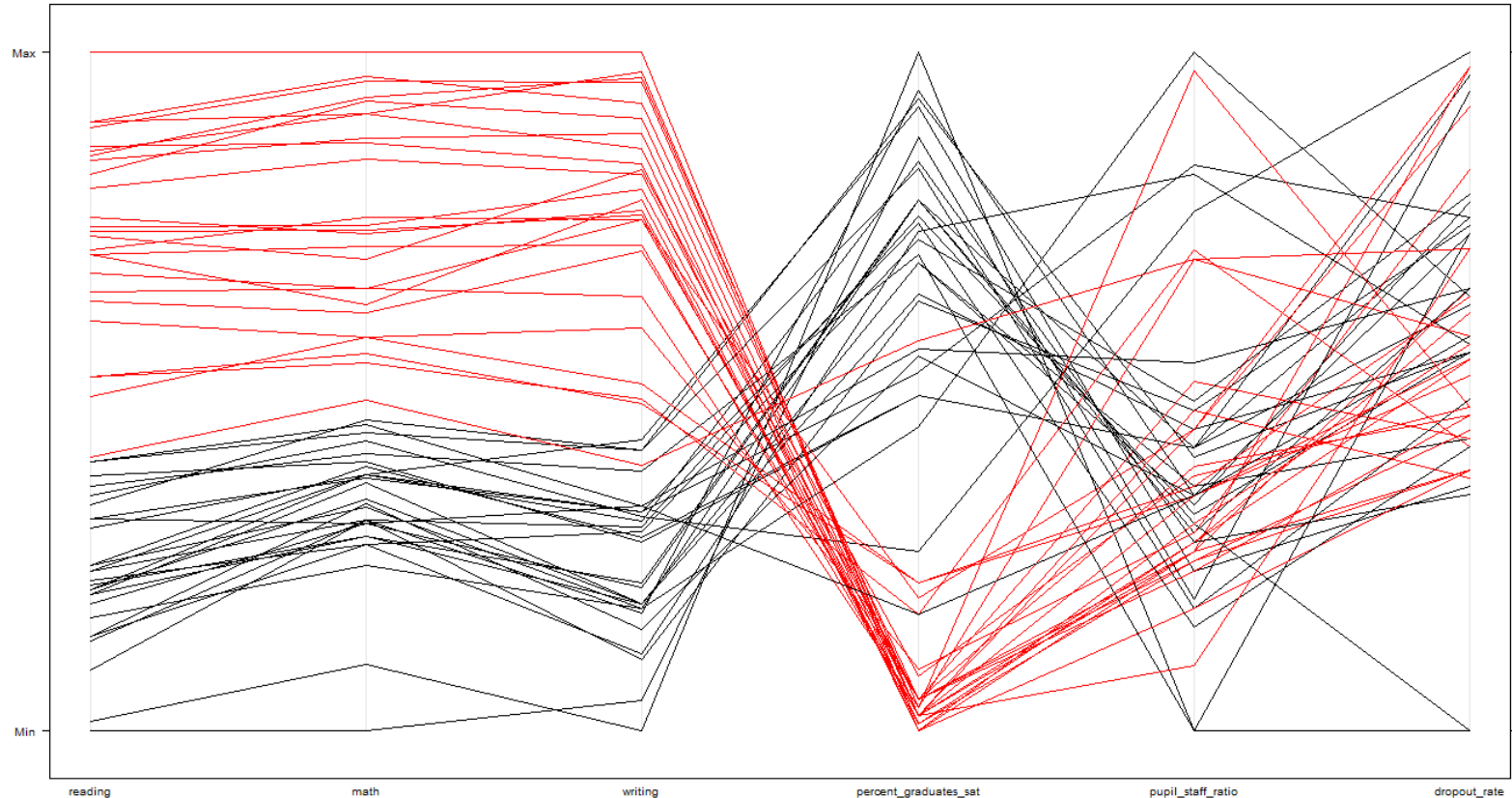
평행좌표플롯 (parallel coordinate plot)



```
education=read.csv("http://datasets.flowingdata.com/education.csv")  
library(lattice)  
parallel(education[,2:7],horizontal.axis=FALSE,col=1)
```

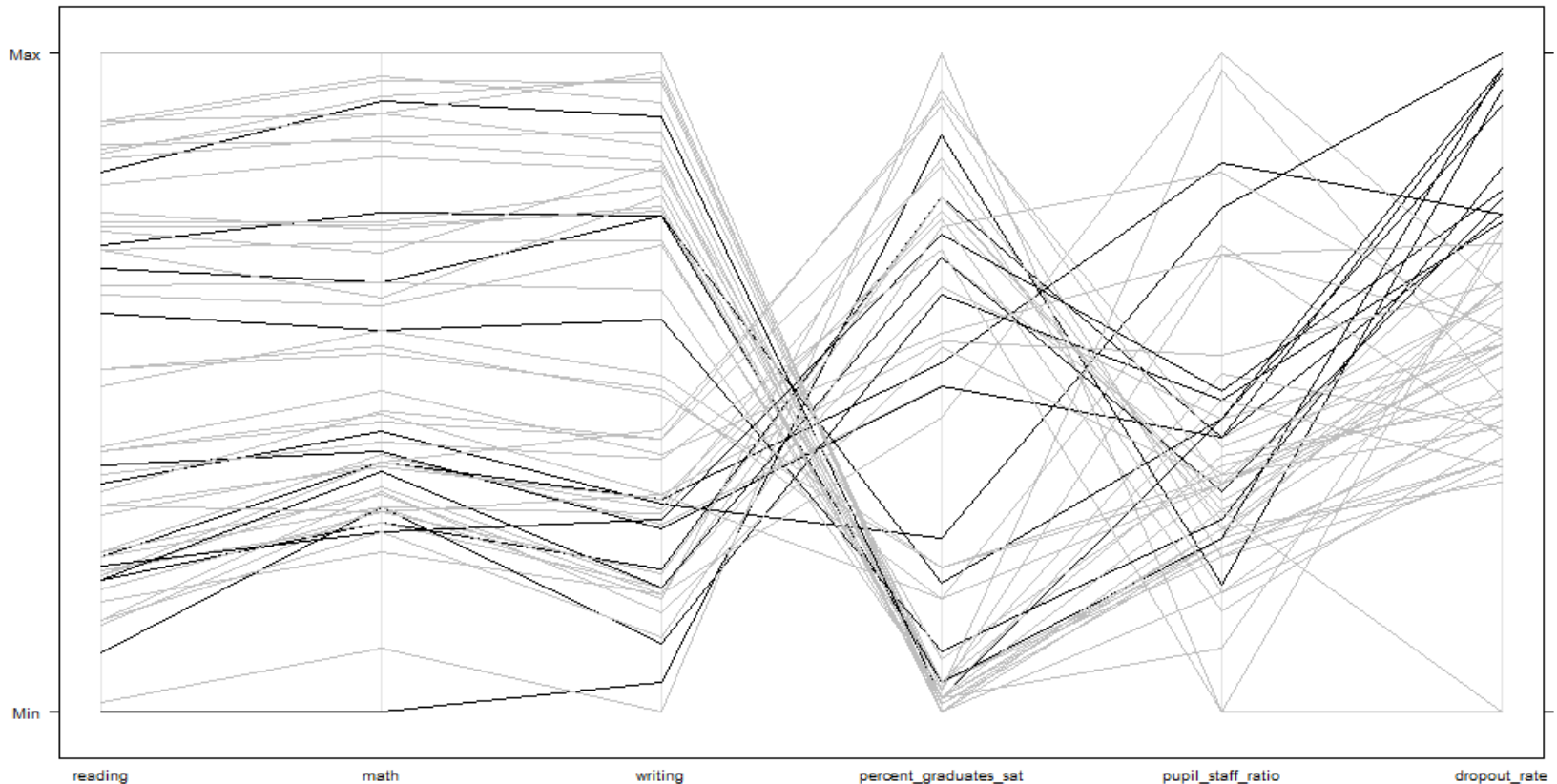
평행좌표플롯 (parallel coordinate plot)

- reading을 기준으로 상위 50% 선으로 나누어 상위 50%와 하위 50%를 다른 색으로 구분



평행좌표플롯 (parallel coordinate plot)

- dropout_rate(중퇴율)을 기준으로 상위 25%를 다른 색으로 표시



주성분 분석

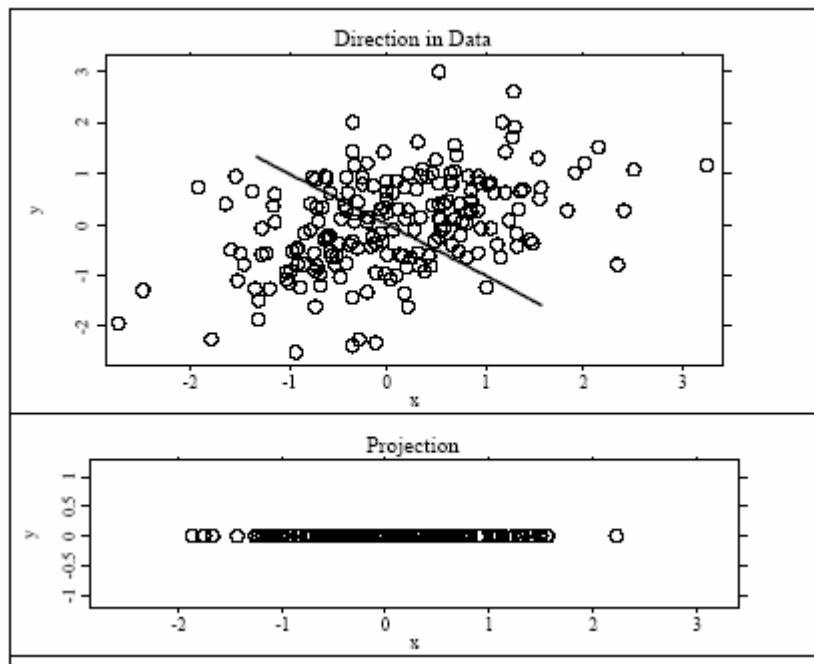
Principal Component Analysis

주성분 분석

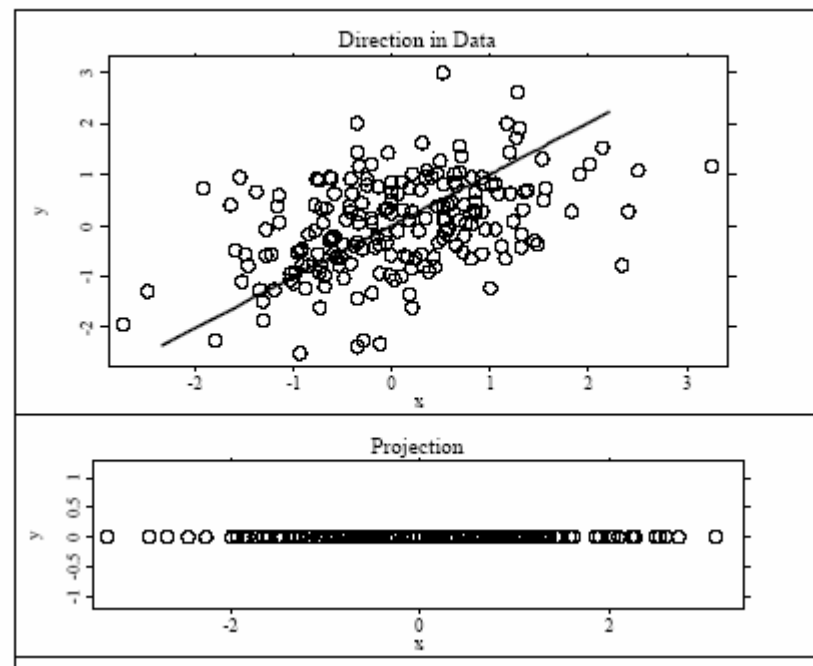
- 본래의 변수들의 변이를 적은 수의 변환된 변수로 설명
- 무수히 많은 선형결합 가운데에서 그 중 가장 높은 설명력을 가지는 선형결합 형태
- 측정변수는 줄지않고 설명요인으로 묶임
- 목적
 - 차원 축약 (Dimension reduction)
 - 회귀분석, 군집분석 등에서 변수를 제거하기 위한 분석으로 사용
 - 자료탐색
 - 이상치 판별
 - 자료의 그룹화

주성분 분석의 원리

Good

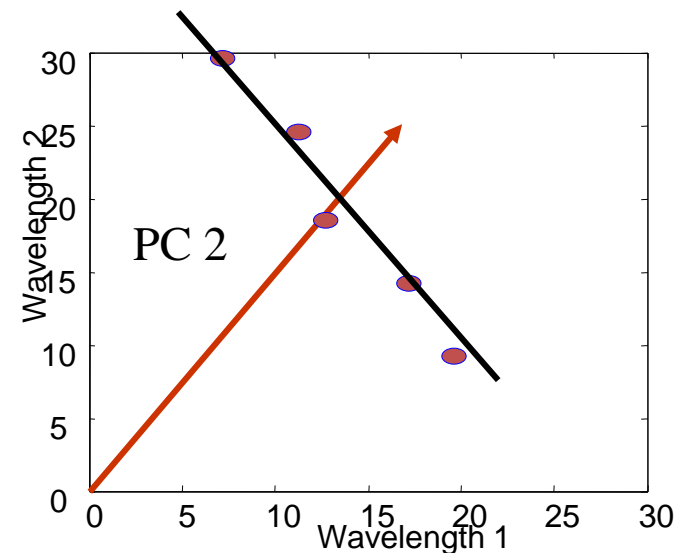
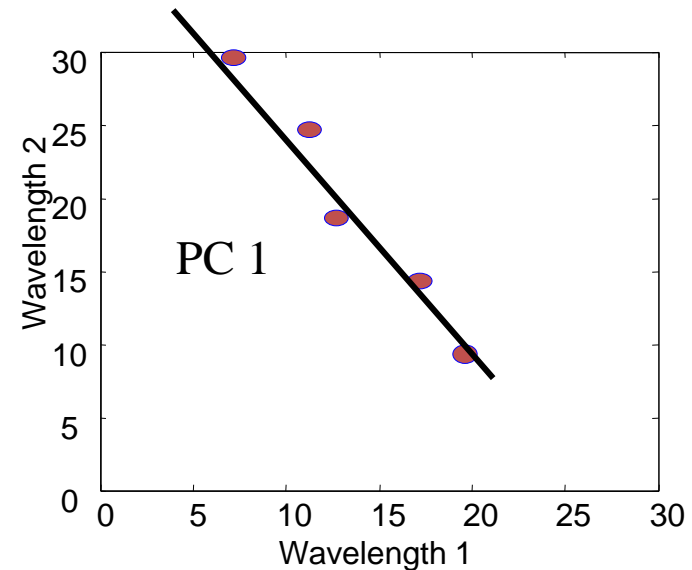


Better



주성분 분석의 원리

- First PC is direction of maximum variance from origin
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance



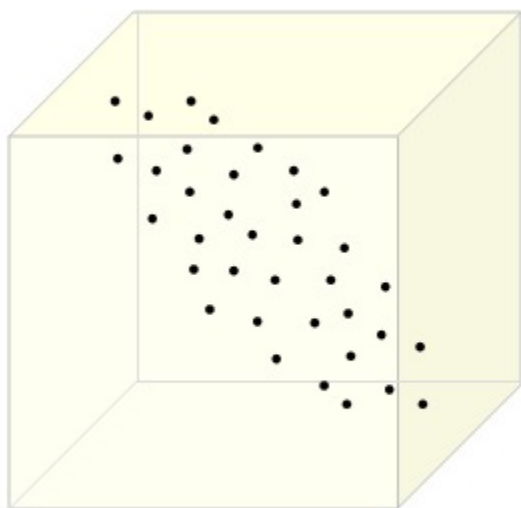
주성분 분석의 이해

- 확률벡터 $\mathbf{x} = (x_1, \dots, x_p)^T$
- 평균 $\boldsymbol{\mu}$, 공분산 행렬 Σ
- 공분산 행렬을 spectral decomposition

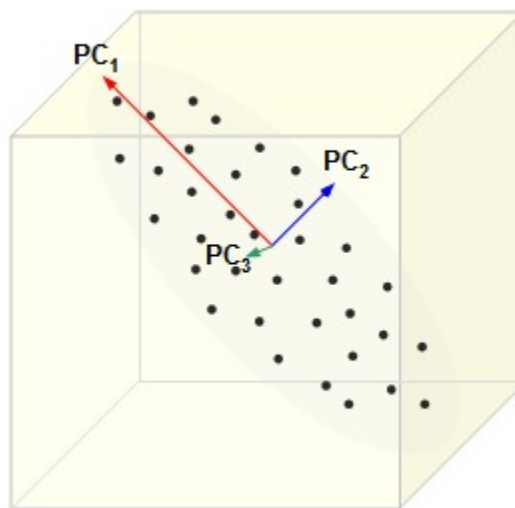
$$\Sigma = \Gamma \Lambda \Gamma^T = \sum_i \lambda_i \gamma_i \gamma_i^T$$

- Γ : eigenvector(γ_i)로 이루어진 직교행렬
- Λ : eigenvalue로 이루어진 대각행렬

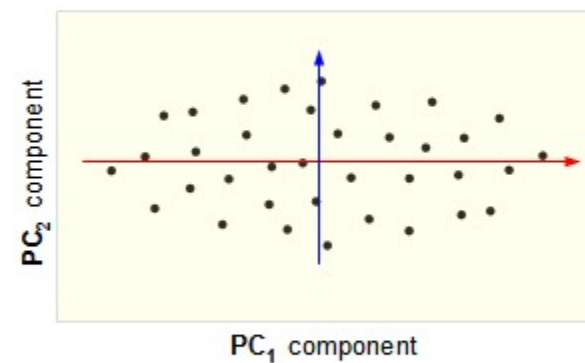
주성분 분석의 이해



a



b



c

PCA에서 의 고려사항

■PCA에서 의 고려사항

- 상관 행렬과 공분산 행렬 중 어느것을 선택할 것인가?
- 주성분의 개수를 몇 개로 할 것인가?
- 주성분에 영향을 미치는 변수는 어떤 변수를 택할 것인가?

■주성분 개수 결정(주성분을 결정하는 기준)

- 고유치가 1보다 클 것 (상관계수 이용의 경우)
- 주성분의 설명력이 원자료 변량의 80-90%이상 설명할 수 있을 것
- 고유치를 그래프로 그려 급격히 감소하기 전까지 선택(Scree Analysis)
- 주성분에 대한 해석이 가능할 것

예: 평균건조무게

- 54종의 식물에 대해 89개 구역에서 측정한 평균건조무게
- plot, time(실험조건), species (종수), hay (건초량), pH(토양 산성도)는 분석에서 제외

```
data.r=read.table("pgfull.txt",header=T)
data=data.r[,1:54]
model=prcomp(data,scale=TRUE)
summary(model)
```

```
> summary(model)
```

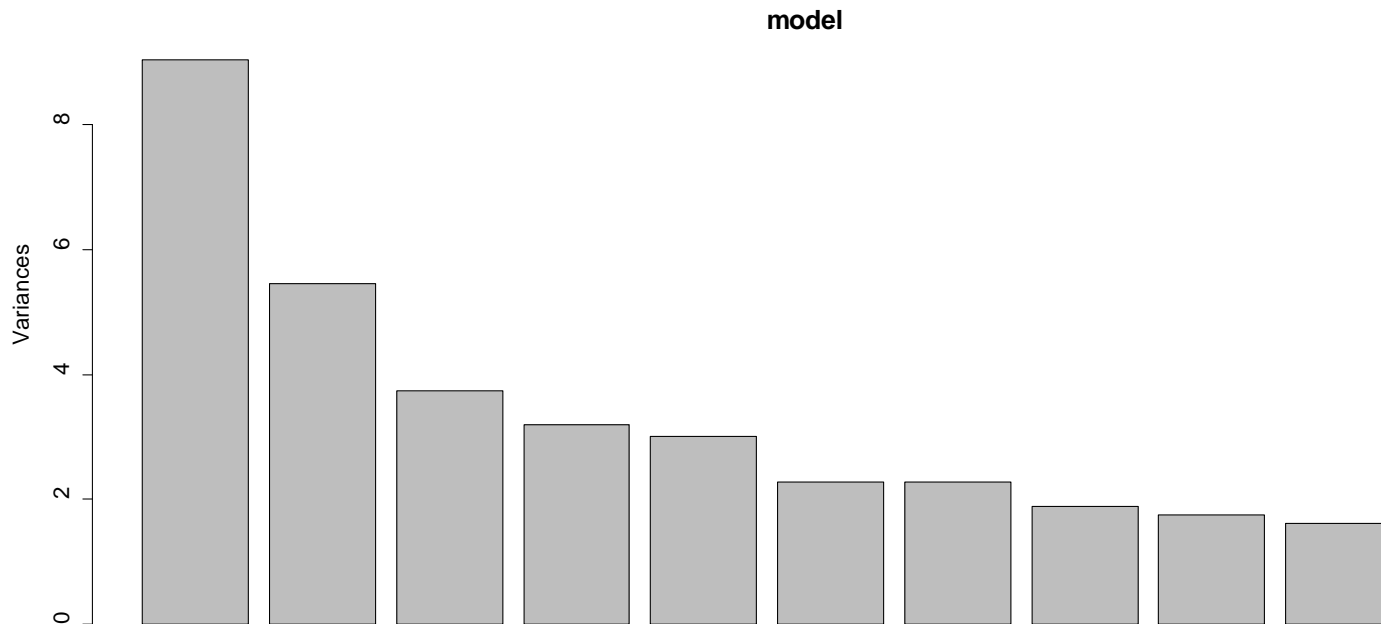
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.0048	2.3358	1.9317	1.78562	1.73303	1.51187	1.50878	1.37586	1.32441	1.27318
Proportion of Variance	0.1672	0.1010	0.0691	0.05904	0.05562	0.04233	0.04216	0.03506	0.03248	0.03002
Cumulative Proportion	0.1672	0.2682	0.3373	0.39639	0.45201	0.49434	0.53649	0.57155	0.60403	0.63405

Scree Plot

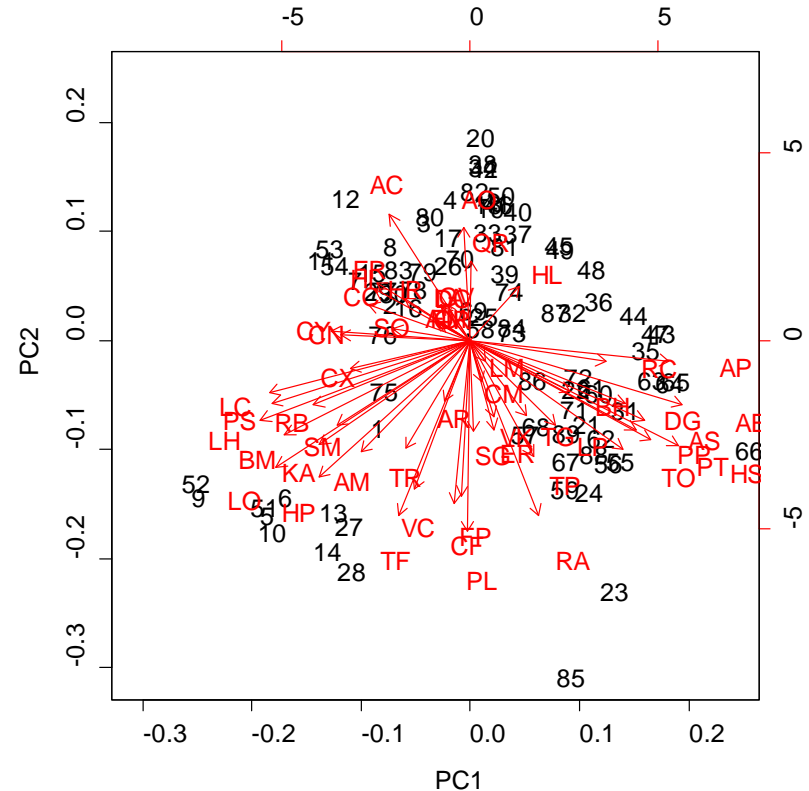
- 각 주성분의 분산을 그래프로 표현
- 주성분 수를 늘렸을 때 설명되는 분산의 증가량 시각화

```
plot(model)
```



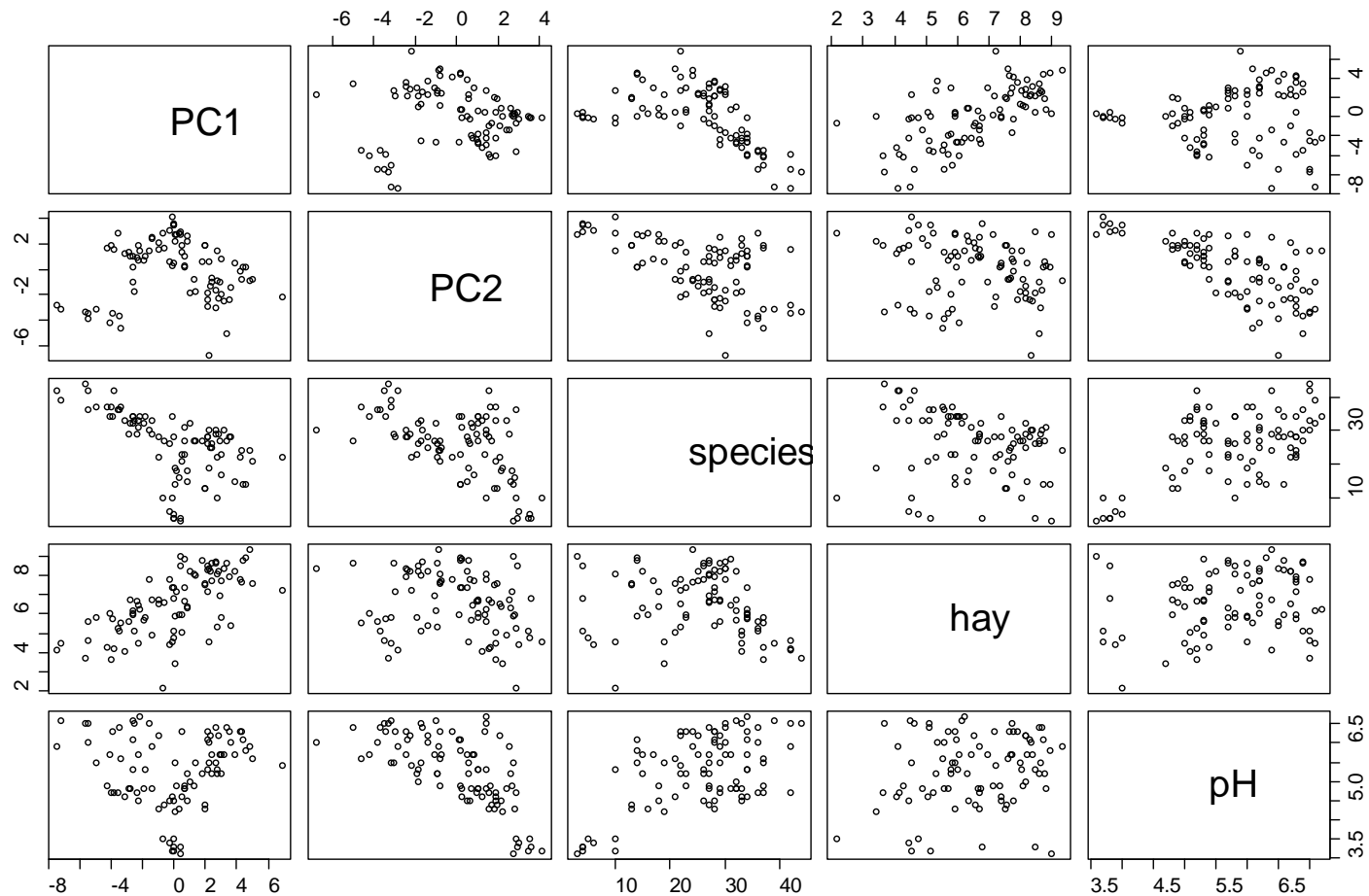
행렬도 (Biplot)

- x축: 각 관찰치의 PC1의 값
- y축: 각 관찰치의 PC2의 값
- 벡터와 축의 각의 cosine: 변수가 해당 PC에 기여한 정도
- 두 벡터가 이루는 각의 cosine: 두 변수의 상관
- 벡터 길이의 제곱: 변수의 분산

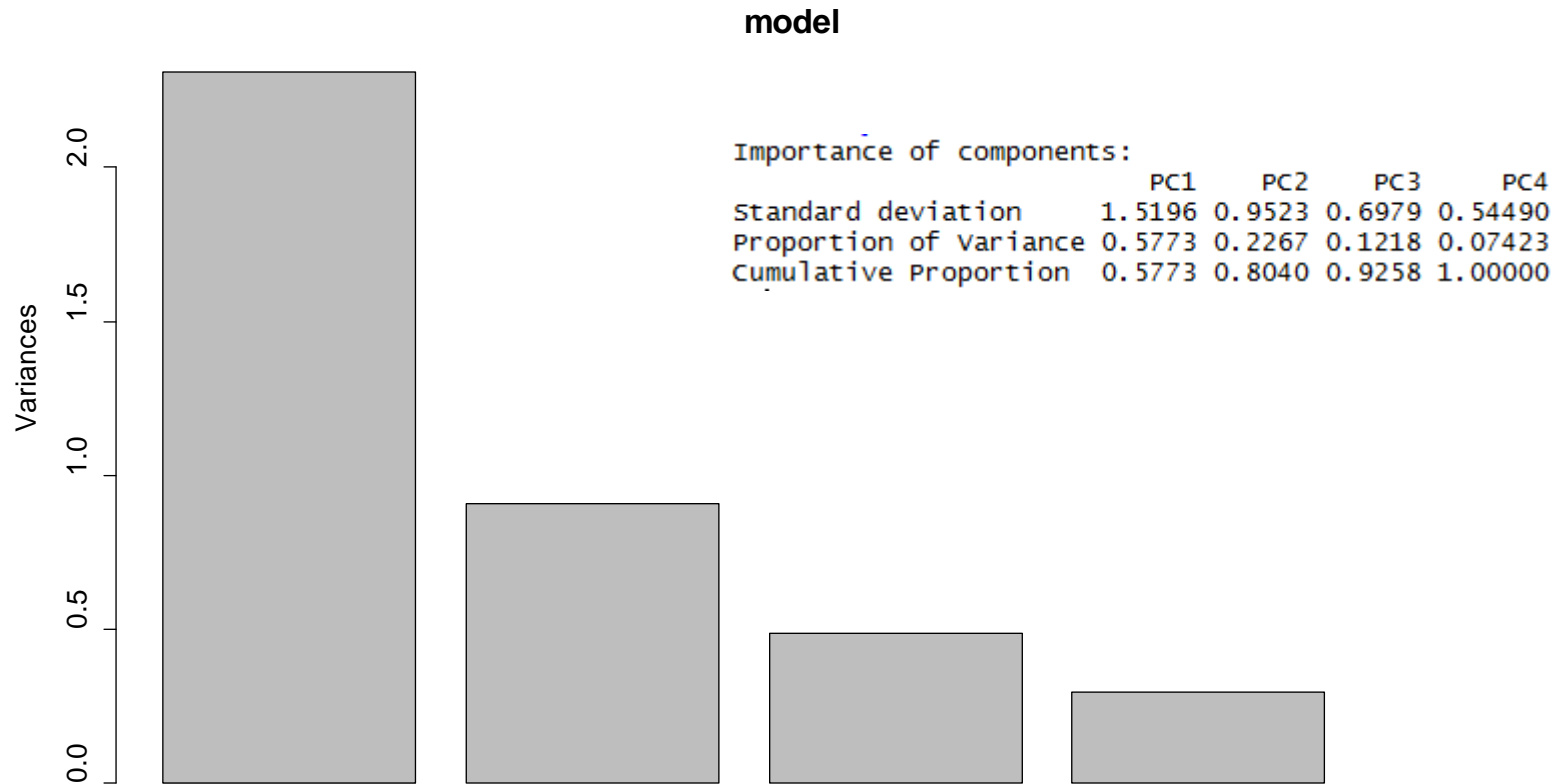


주성분과 공변량 관계 탐색

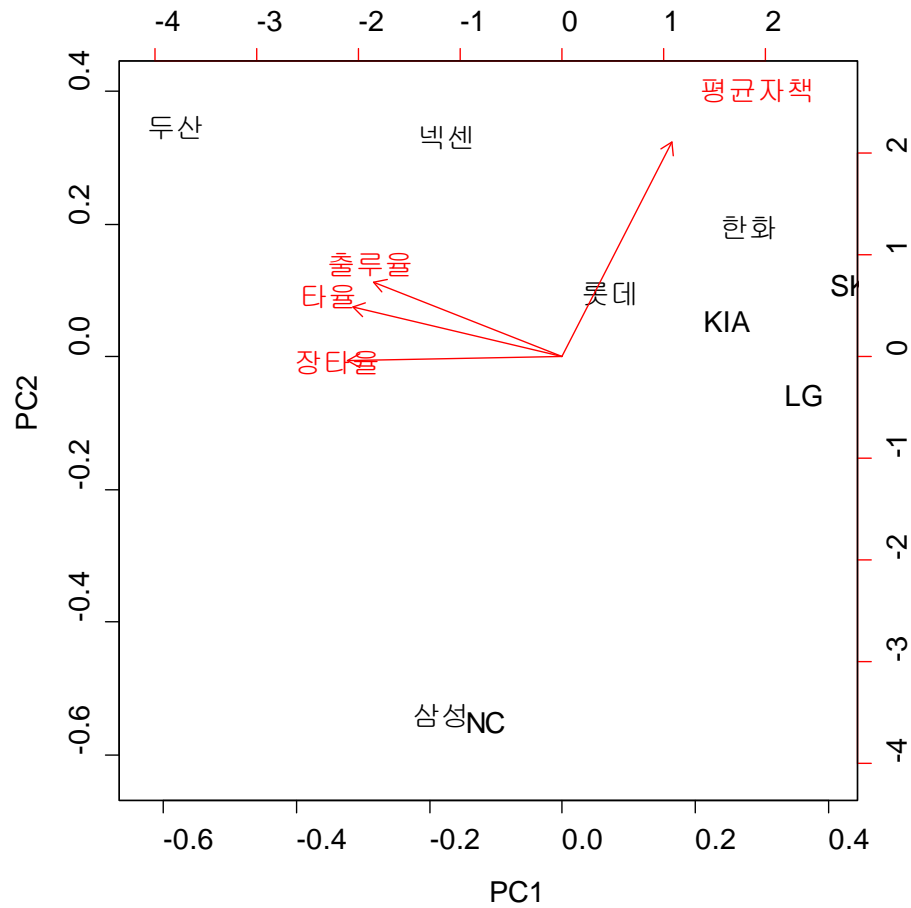
```
pairs(cbind(predict(model)[,1:2],data.r[,57:59]))
```



예: 2014년 한국 프로야구



예: 2014년 한국 프로야구



예: 2014년 한국 프로야구

