주성분분석

Jinseog Kim

October 8, 2007

1 주성분분석 (Principal Component Analysis)

Page 151
주성분분석 (Principal Component Analysis)은 다차원자료를 설명력이 높은 몇개의 차원으로 축소하기 위한 분석방법이다.

p 차원 확률벡터 $\mathbf{x} = (x_1, \ldots, x_p)'$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산행렬이 Σ 일 때,

$$\mathbf{y} = P'\mathbf{x}$$

인 선형변환을 고려하자. 이 때, 선형변환의 결과 $\mathbf{y}$의 공분산은

$$Cov(\mathbf{y}) = P'Cov(\mathbf{x})P = P'\Sigma P.$$

이 된다. 만일 $\mathbf{x}$의 선형변환 중에서 변환된 결과의 공분산행렬이 대각선원

소를 제외하고 모두 0이면, 즉

$$Cov(\mathbf{y}) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}$$

변환된 결과 $\mathbf{y} = (y_1, \dots, y_p)'$ 는 모두 서로 독립이다. 또한 $Cov(\mathbf{y})$의 대각선 원소는 일변량 변수 $y_1, \dots, y_p$ 각각에 대한 분산이 된다. 여기서 분산의 크기가 $Var(y_1) \geq \dots Var(y_p)$라고 가정하자.

그러면 이러한 조건을 만족하는 선형변환 (P)은 어떻게 구할까?

Σ의 스펙트럴분해 (spectral decomposition)은

$$\Sigma = \Gamma \Lambda \Gamma'.$$

여기서 Γ 는 직교행렬(orthogonal)이고, Σ의 고유벡터로 구성되며, Λ는 대각행렬로써 대각선원소는 Σ의 고유치로 구성된다. 즉,

$$\Sigma = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \lambda_p \end{pmatrix} \begin{pmatrix} \boldsymbol{e}_1' \\ \vdots \\ \boldsymbol{e}_p' \end{pmatrix}.$$

이런 사실을 이용하면

$$Cov(\mathbf{y}) = P'\Sigma P = P'\Gamma\Lambda\Gamma' P = \Lambda.$$

이기 위해서는

$$P = \Gamma$$

이면 된다. 이 때, $\mathbf{y}$를 $\mathbf{x}$의 주성분(Principal component)라고 부르며, 이 경우 공분산행렬의 대각원소의 합은 동일하다.

$$\sum_{j=1}^{p} Var(x_j) = \mathrm{tr}(\Sigma) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(y_j) \tag{1}$$

```
X<-cov(Prestige[,1:4]) #or var(...)
ei<-eigen(X)
LAMBDA<-diag(ei$values)
GAMMA<-ei$vectors
Z<-t(GAMMA)%*%X%*%GAMMA
diag(Z)
    [1] 1.802821e+07 8.287121e+02 1.298184e+02 1.816137e+00
sum(diag(Z))
    [1] 18029165
sum(diag(X))
    [1] 18029165
```

1.1 PCA 의 선택

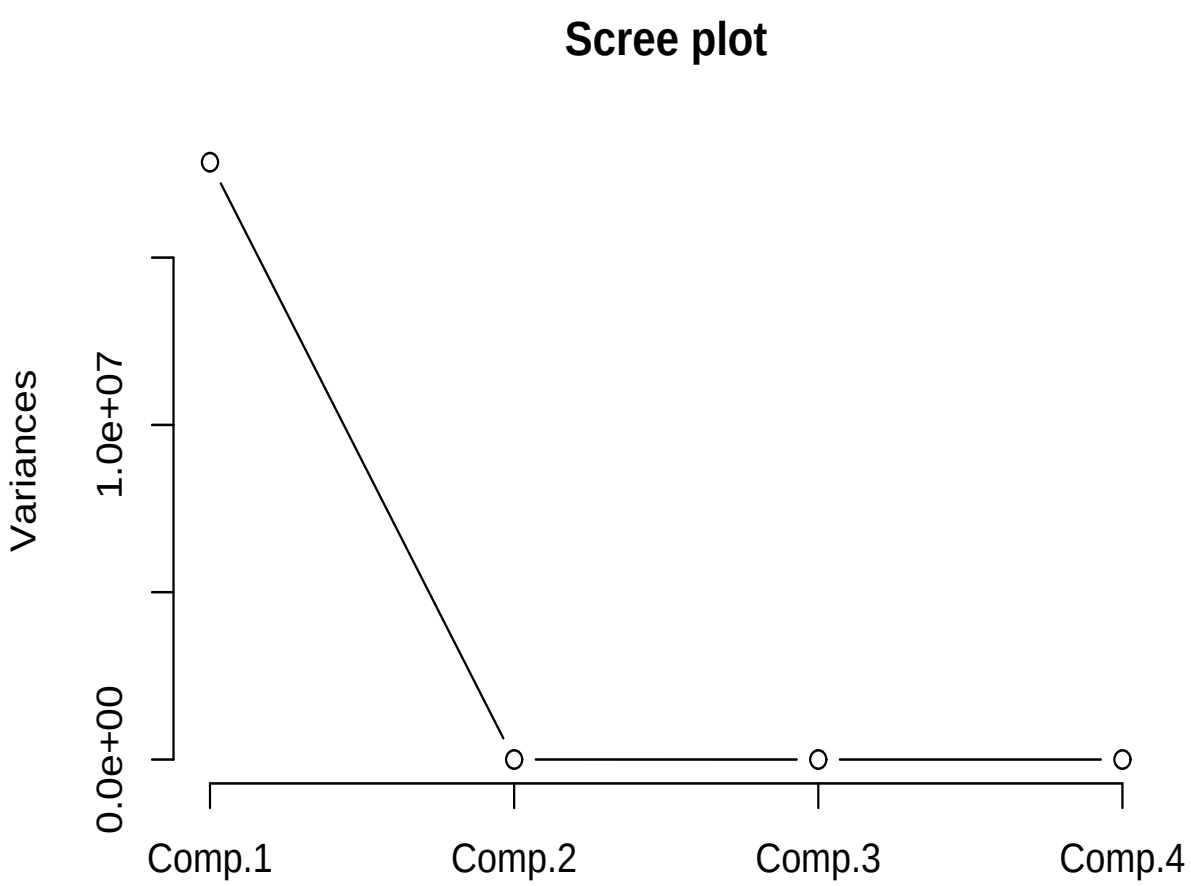
Scree Plot의 이용: 전체분산(변동) 중 주성분이 설명하는 변동의 양을 이용
전체 p개의 변수가 있을 때 다음을 계산하여 그림으로 표현해 준다.

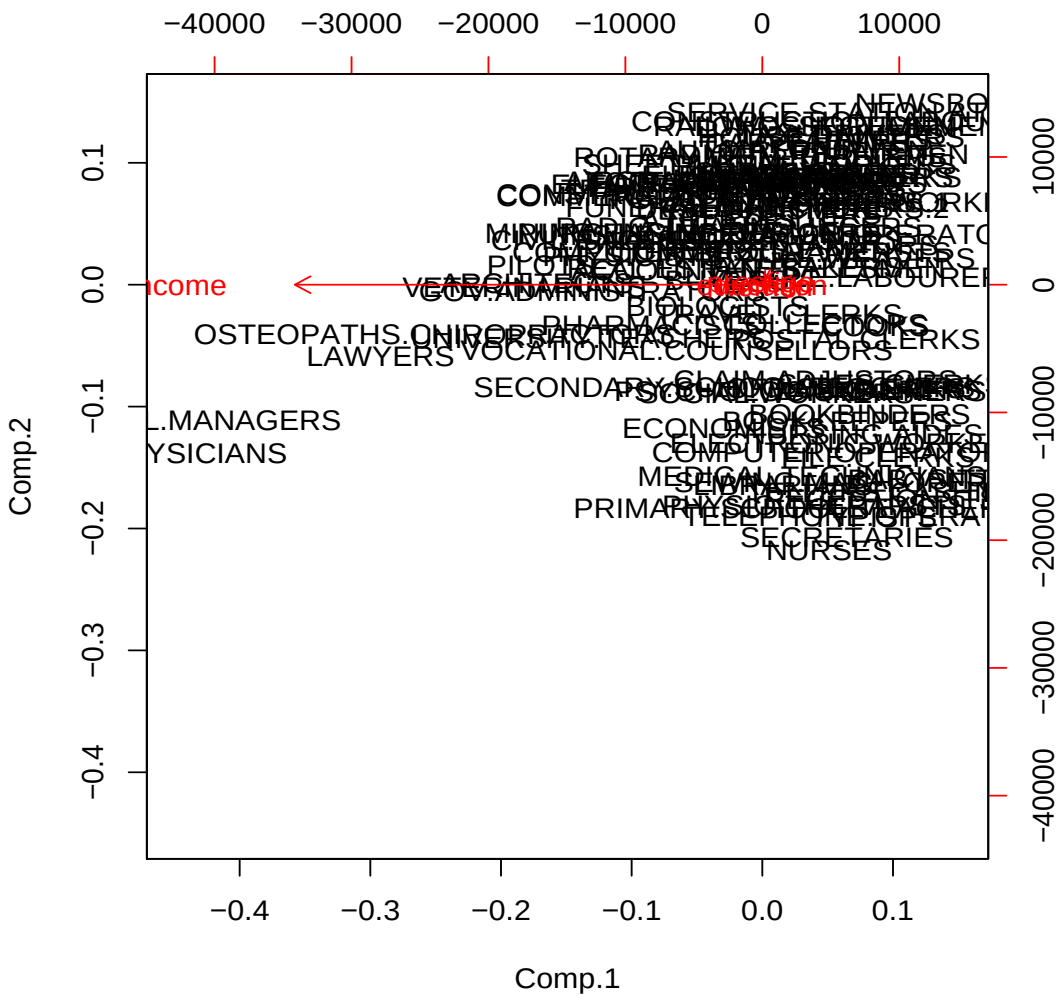
$$Var(y_1 + \ldots + y_q) = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, q = 1, \ldots, p.$$

1.2 R을 이용한 PCA

```
>pr1<-princomp(Prestige[,1:4])
>summary(pr1)
#--
Importance of components:
                          Comp.1  Comp.2     Comp.3    Comp.4
Standard deviation     4225.098 2.864e+01 1.133e+01 1.341e+00
Proportion of Variance    0.999 4.596e-05 7.200e-06 1.007e-07
```

```
Cumulative Proportion     0.999 9.999e-01 9.999e-01 1.000e+00
#--

>screeplot(pr1, type="lines", main="Scree plot")

>loadings(pr1) # Gamma: matrix of eigen vectors

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4
education                 0.126  0.991
income    -1.000
women             -0.987 -0.160
prestige          -0.156  0.979 -0.130


>biplot(pr1)
```


Scree plot
Variances
0.0e+00
1.0e+07
Comp.1
Comp.2
Comp.3
Comp.4


−40000
−30000
−20000
−10000
0
10000
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Comp.2
Comp.1
10000
0
−10000
−20000
−30000
−40000
OSTEOPATHS
LAWYERS
VOCATIONAL.COUNSELLORS
SECONDARY
.MANAGERS
YSICIANS
PRIMA
SECRETARIES
NURSES

1.3 PCA의 활용

- 차원축소: 저차원을 이용한 Data Visualization
- 중회귀분석: 입력변수간의 다중공선성이 있을 때
- 요인분석, 판별분석, 집락분석, 이상치의 탐색 등.

1.4 실습 및 과제

- (과제) 교과서 p. 164 표8.3에 있는 자료에 대하여 주성분 분석을 하시오. (중간고사 전까지)
- (실습) MASS package 에 있는 UScrime data set을 이용하여 중회귀분석을 하시오.