**주성분 분석(Principal Component Analysis)**

http://contents.kocw.or.kr/document/lec/2012/DukSung/KimJaehee/09.pdf
http://contents.kocw.or.kr/KOCW/document/2015/chungbuk/najonghwa1/5.pdf

많은 변수로 구성된 데이터에 대해 주성분이라는 새로운 변수를 만들어 기존 변수보다 차원을 축소하여 분석을 수행

주성분 P1은 데이터 분산을 가장 많이 설명할 수 있는 것을 선택하고 P2는 P1과 수직인 주성분을 만들어 다중 공선성 문제를 해결

다중 공선성(MultiCollinearity): 독립 변수사이에 강한 상관관계가 나타나서 종속변수에 영향을 미치는 경우
완전 공선성: 독립 변수들 사이에 정확한 선형 관계가 존재하는 경우
다중 공선성 문제는 분석과 예측의 정확성을 위해서 피하거나 해결해야 한다.

실습.

```
================
### PCA

data("iris")
head(iris)

# 변수간 상관관계 확인
cor(iris[1:4])

# 변수간 S.L와P.L, S.L와 P.W간의 상관관계 높음
# 다중공선성 문제 발생 예상
# 독립변수 새롭게 설계 필요

# 전처리 과정
iris2 <- iris[, 1:4]
ir.species <- iris[,5]

# 중앙을 0, 분산은 1로 설정
```

```
prcomp.result2 <- prcomp(iris2, center=T, scale=T)
prcomp.result2

# 결과
summary(prcomp.result2)

# 주성분 수 설정
plot(prcomp.result2, type="l")

# result2의 데이터 확인
prcomp.result2$rotation
iris2

# iris2 데이터와 prcomp.result2데이터를 행렬곱하여 변환
Result3 <- as.matrix(iris2) %*% prcomp.result2$rotation

# 변환결과 확인
head(Result3)

# 종 데이터와 Result3의 데이터프레임을 열병합
final2 <- cbind(ir.species, as.data.frame(Result3))
final2

# factor형으로 변환
final2[,1] <- as.factor(final2[,1])

# 컬럼명을 label1로 명명
colnames(final2)[1] <- "label1"

# final2 확인
final2

# 새로 구성된 데이터로 회귀 분석 실시
fit3 <- lm(label1 ~ PC1 + PC2, data=final2)

fit3_pred <-predict(fit3, newdata=final2)
b2 <- round(fit3_pred)
a2 <- ir.species
table(b2,a2)
```

==========

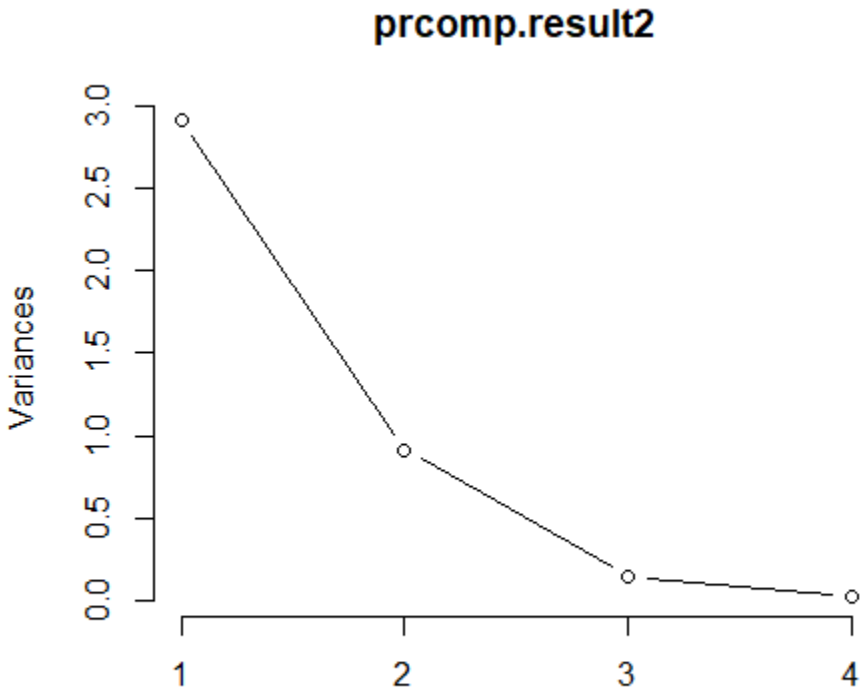
```
> data("iris")
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
> # 변수간 상관관계 확인
> cor(iris[1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
>
```

```
> prcomp.result2 <- prcomp(iris2, center=T, scale=T)
> prcomp.result2
Standard deviations (1, .., p=4):
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation (n x k) = (4 x 4):
                    PC1         PC2        PC3         PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
>
> # 결과
> summary(prcomp.result2)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
>
> # 주성분 수 설정
> plot(prcomp.result2, type="l")
>
> # result2의 데이터 확인
> prcomp.result2$rotation
                    PC1         PC2        PC3         PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

```
> # 변환결과 확인
> head(Result3)
          PC1        PC2       PC3         PC4
[1,] 2.640270 -5.204041 2.488621 -0.1170332
[2,] 2.670730 -4.666910 2.466898 -0.1075356
[3,] 2.454606 -4.773636 2.288321 -0.1043499
[4,] 2.545517 -4.648463 2.212378 -0.2784174
[5,] 2.561228 -5.258629 2.392226 -0.1555127
[6,] 2.975946 -5.707321 2.437245 -0.2237665

> table(b2,a2)
   a2
b2  setosa versicolor virginica
  1     50          0         0
  2      0         44         5
  3      0          6        45
>
```




실습2

http://contents.kocw.or.kr/KOCW/document/2015/chungbuk/najonghwa1/6.pdf