# 13장 Principal Component Regression

# CONTENTS

13.1 서론

**13.2 주성분분석 소개**

**13.3 주성분회귀**

# 13.1 서론

- 주성분회귀는 주성분분석에 기초한 회귀분석 기법이다. 주성분회귀에서는, 반응변수를 직접 설명변수에 회귀하는 대신, 설명변수들의 주성분들을 회귀변수로 사용한다. 통상적으로 전체 주성분들의 부분집합(일부분)을 회귀변수로 사용하며, 이 과정은 주성분회귀를 일종의 정칙화 (regularized) 절차로 이해되도록 한다.
- 주성분회귀는 다중공선성에 대한 해결책으로 주로 이용된다. 주성분회귀는 회귀 단계에서 작은 변동을 가지는 주성분들을 제외시킴으로써 이러한 문제를 적절히 다룰 수 있다. 또한, 주성분의 일부만을 회귀에 사용함으로써 차원축소와 함께, 추정해야할 모수의 수를 줄여준다.
- 이 사실은 고차원의 자료에 특히 유용하게 사용될 수 있다. 회귀변수로 사용될 주성분의 적절한 선택을 통해, 주성분회귀는 반응변수에 대한 효과적인 예측을 가능하게 한다.

# 13.2 주성분분석 소개

- 주성분분석(principal component analysis, 이하 PCA)은 직교변환을 통해 서로 상관된 변수들의 관측치들을 선형적으로 무상관 된 변수(이를 주성분이라 함)값으로 변환시키는 통계적 방법이다. 주성분의 수는 원 변수의 수와 작거나 같다.
- 이 변환은 다음과 같은 방법으로 정의된다. 제1 주성분은 가장 큰 분산을 가지도록(즉, 데이터가 가진 변동을 최대한 많이 설명하도록) 정해지며, 이후의 주성분은 차례대로 앞선 주성분과는 직교하면서 가능한 최대 분산을 가지도록 정해진다. 결과 벡터는 무상관의 직교 기저 집합이 된다. PCA는 원 변수들의 상대 척도에 민감하다.

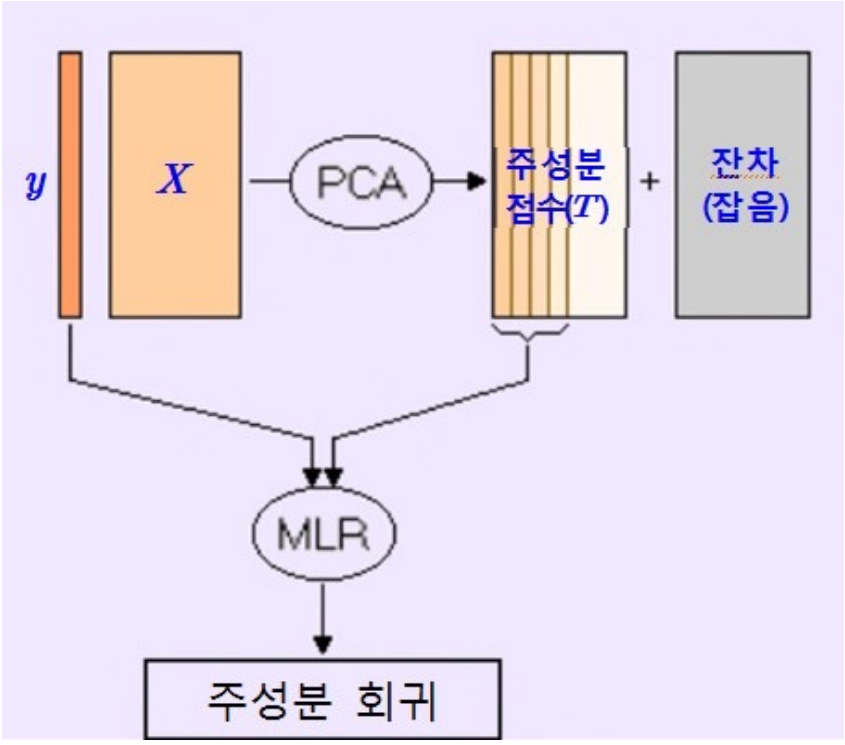
# 13.2 주성분분석 소개

- PCA는 탐색적 자료 분석과 예측모델링 방법으로 주로 사용된다. PCA는 고유차 기반의 다변량 분석 가운데 가장 단순한 방법으로, 고차원의 다변량 자료를 저차원의 공간에 사영한다. 이는 처음 몇 개의 주성분만을 사용함으로써 변환된 자료의 차원을 줄이는 방법을 사용한다.

# 13.3 주성분회귀

- 주성분회귀(Principal Component Regression, 이하 PCR)는 다중공선성이 존재하는 자료에 대한 다중회귀분석법이다. 다중공선성이 존재할 때, 최소제곱추정량(치)는 불편성은 만족하나 분산이 매우 커지게 되어 실제 값으로부터 멀리 떨어질 수 있다. 능형회귀와 마찬가지로 약간의 편향(bias)을 추가함으로써, PCR은 표준오차를 감소시킨다.
- 능형회귀에서와 같이 PCR에서도 자료의 표준화가 요구된다(독립, 반응변수 모두). 아래의 표현에서 $X$와 $y$는 모두 표준화 된 것으로 정의한다.
- PCR은 다중회귀와 주성분분석의 단순한 확장으로 그 절차는 다음의 [그림 13.1]과 같다. 첫 단계는 원 자료로부터 주성분(또는 (인자) 점수)을 계산한다. 중요한(선택된) 주성분 자료를 예측변수로 하여 반응변수($y$)와의 다중회귀를 수행한다. 주성분 변수들은 서로 직교(무상관)하므로, 다중공선성의 문제가 해결되어 다중회귀분석(OLS 추정)이 가능하다.

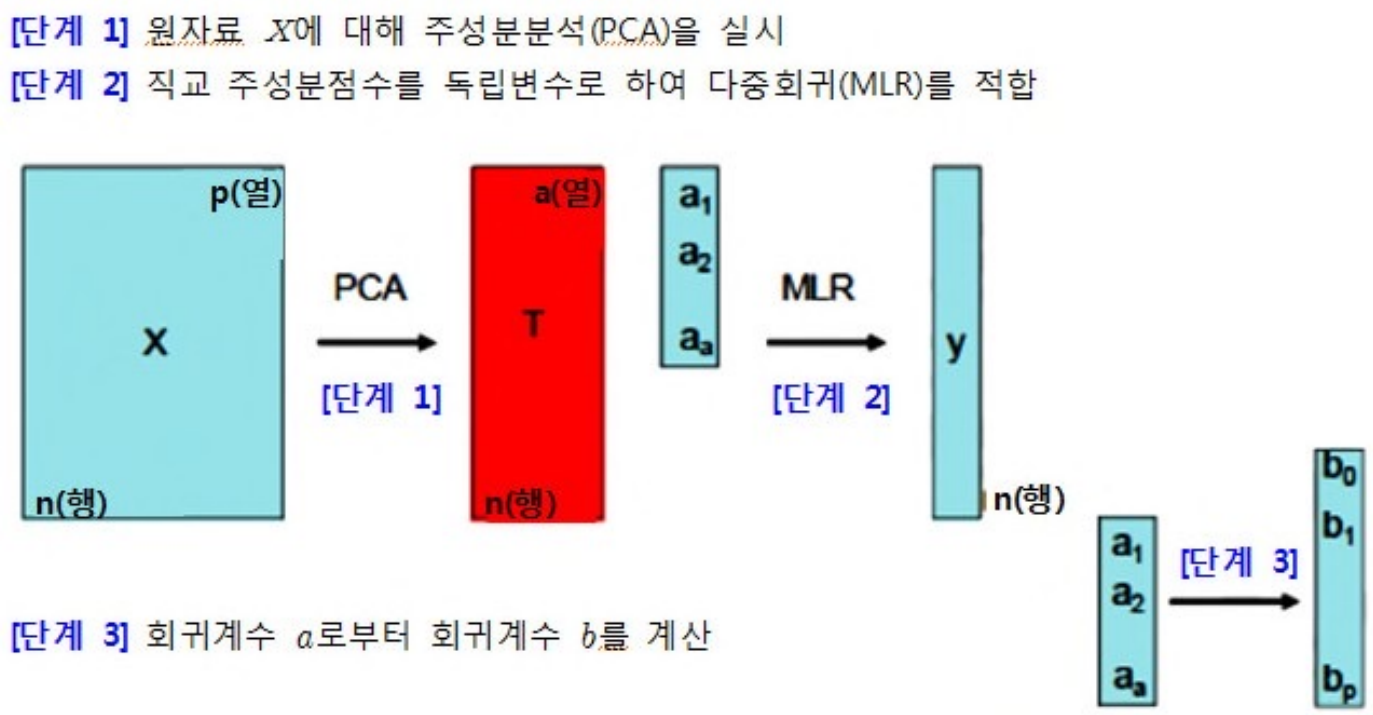
# 13.3 주성분회귀




[그림 13.1] 주성분회귀의 수행원리

# 13.3 주성분회귀

- PCR 회귀의 수행(모수 추정) 절차를 정리하면 다음의 [그림 13.2]와 같다.




[그림 13.2] 주성분회귀의 모수추정 절차

# 13.3 주성분회귀

- PCR의 특징을 소개하면 다음과 같다.

(i ) PCR에서 가장 중요한 점은 PCA에서와 마찬가지로 적절한 고유벡터를 선택하는 문제이다. 이 과정은 PCA에서와 동일하다.

(ii ) PCR의 가정은 통상적인 다중선형회귀의 가정(선형성, 상수분산(no outliers)과 독립성)과 유사하나, PCR은 신뢰구간을 제공하지 않으므로 정규성의 가정은 불필요하다.

(iii ) 다중공선성이 존재하는 경우 뿐 아니라 변수의 수가 자료의 수보다 많은 경우에도 적용될 수 있다.

# 13.3 주성분회귀

**예제 1** gasolin{pls} 자료에 대해 주성분회귀를 수행한다.

```
> library(pls)
> data(gasoline)
> str(gasoline)
'data.frame':  60 obs. of  2 variables:
 $ octane: num  85.3 85.2 88.5 83.4 87.9 ...
 $ NIR   : AsIs [1:60, 1:401] -0.0502 -0.0442 -0.0469 -0.0467 -
0.0509 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr  "1" "2" "3" "4" ...
  .. ..$ : chr  "900 nm" "902 nm" "904 nm" "906 nm" ...
```

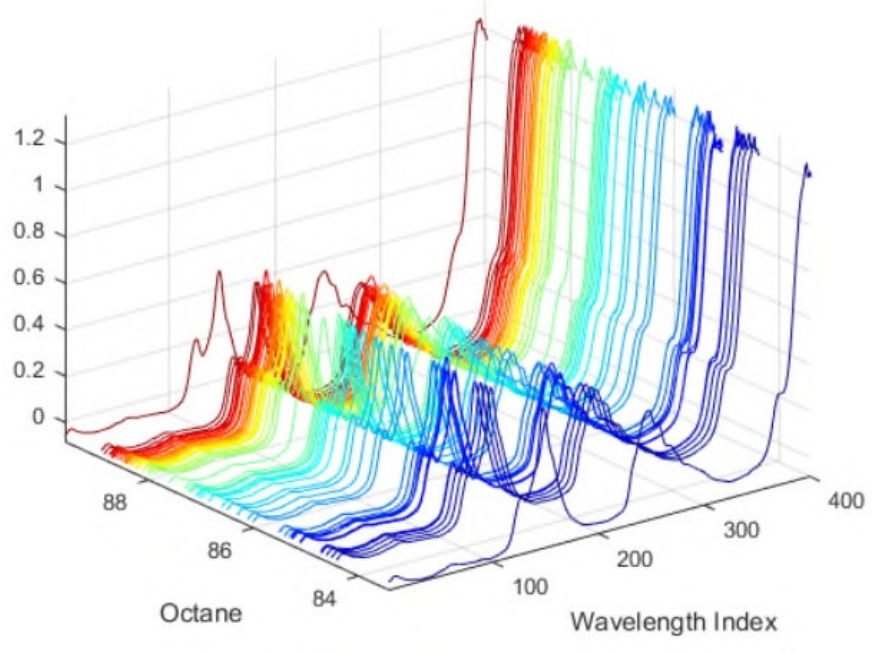
# 13.3 주성분회귀

**자료 설명**

gasoline 자료는 60개의 가솔린 표본의 근적외선(NIR) 스펙트럼과 옥탄가(옥탄)의 자료이다. 각 NIR 스펙트럼은 900~1700nm 범위에서 2nm 간격으로 총 401개의 파장 길이에서 측정되었다. 가솔린($60 \times 1$)을 반응변수로, NIR($60 \times 401$)을 예측변수로 하는 회귀모형을 적합하고자 한다. 예측변수의 수가 401개로 관측된 자료(60개)에 비해 매우 크다.

# 13.3 주성분회귀

- 다음의 [그림 13.3]은 자료의 이해를 돕기 위한 그림이다.




[그림 13.3] gasoline 자료

# 13.3 주성분회귀

```
> ## 훈련용 자료와 검증용 자료 생성
> gasTrain <- gasoline[1:50,]
> gasTest <- gasoline[51:60,]


> ## 주성분회귀 수행: pcr{pls} 함수 이용
> gas1 <- pcr(octane ~ NIR, ncomp=10, data=gasTrain,
              validation="LOO")
> summary(gas1)
Data:   X dimension: 50 401
        Y dimension: 50 1
Fit method: svdpc
Number of components considered: 10
                                  (…)
```

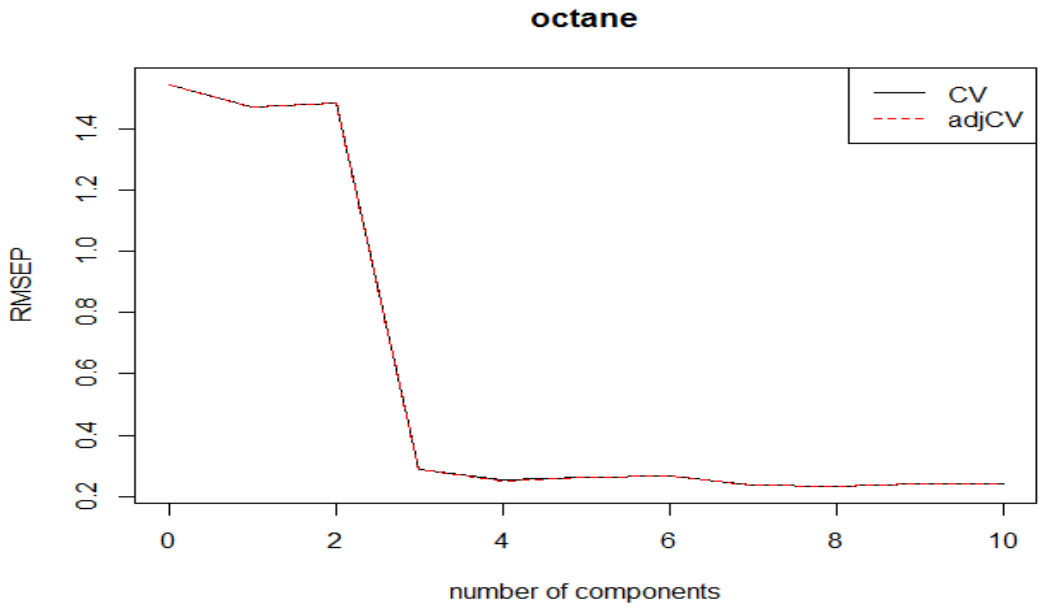
# 13.3 주성분회귀

```
                                        (…)
VALIDATION: RMSEP
Cross-validated using 50 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           1.545    1.472    1.483   0.2894   0.2522   0.2622
adjCV        1.545    1.471    1.482   0.2879   0.2518   0.2618
       6 comps  7 comps  8 comps  9 comps  10 comps
CV      0.2681   0.2386   0.2328   0.2416    0.2423
adjCV   0.2677   0.2373   0.2323   0.2411    0.2415


TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps
X         79.86    88.12    93.54    96.54    97.74
octane    16.99    21.36    97.00    97.71    97.73
        6 comps  7 comps  8 comps  9 comps 10 comps
X         98.38    98.75    99.06    99.28    99.42
octane    97.77    98.47    98.54    98.62    98.83
```

# 13.3 주성분회귀

```
> plot(RMSEP(gas1), legendpos="topright")
```

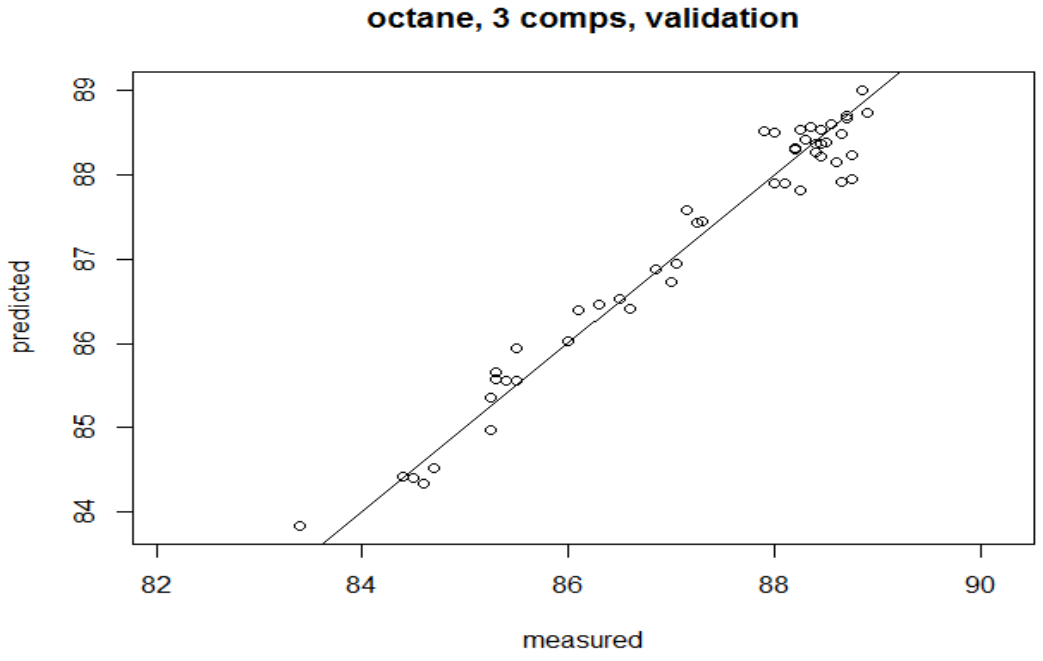



**해 석**

LOO(leave-one-out) 방법으로 RMSEP(제곱근평균제곱예측오차)를 구해본 결과 주성분의 수는 3이 적당한 것으로 보인다.

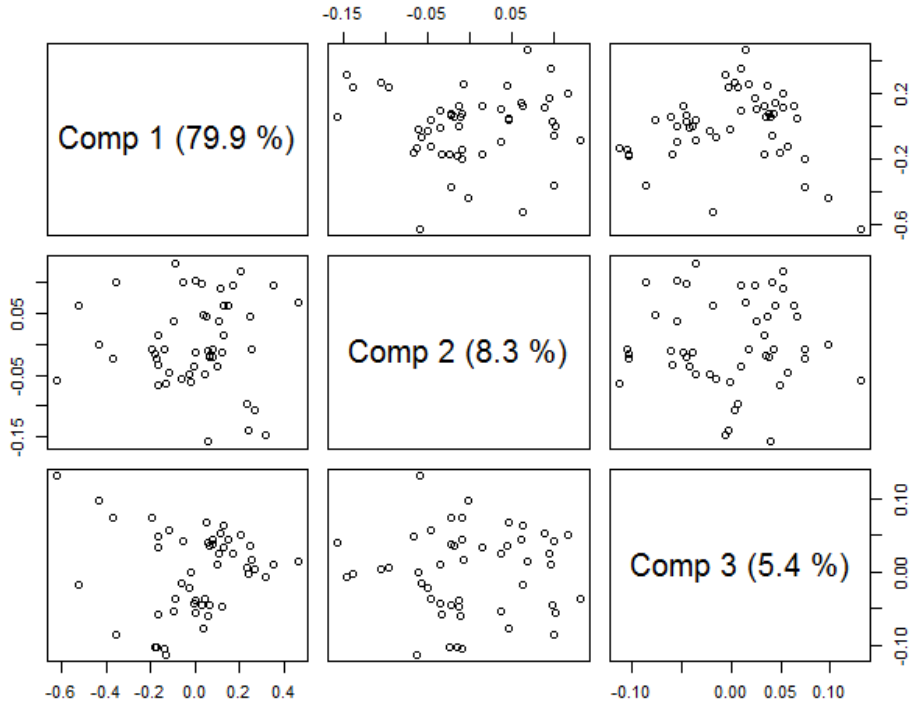
# 13.3 주성분회귀

```
> ## 3개의 주성분을 이용하여 PCR을 수행한 결과
> plot(gas1, ncomp=3, asp=1, line=TRUE)
```



# 13.3 주성분회귀

```
> plot(gas1, plottype="scores", comps=1:3)
```



# 13.3 주성분회귀

```
> explvar(gas1)
    Comp 1     Comp 2     Comp 3     Comp 4     Comp 5
79.8586603  8.2639500  5.4171903  3.0034945  1.1963215
    Comp 6     Comp 7     Comp 8     Comp 9    Comp 10
 0.6397503  0.3691514  0.3127762  0.2171267  0.1417888
```

# 13.3 주성분회귀

```
> predict(gas1, ncomp=3, newdata=gasTest)
, , 3 comps

     octane
51 87.63119
52 87.17090
53 87.84391
54 84.44888
55 84.95272
56 84.63236
57 86.88466
58 86.50888
59 88.75387
60 86.63756
```

# 13.3 주성분회귀

```
> ##  주성분의 수에 따른 RMSEP

> RMSEP(gas1, newdata=gasTest)
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps
     1.5369       1.3226       1.2568       0.4634       0.2241       0.2283
    6 comps      7 comps      8 comps      9 comps     10 comps
     0.2600       0.2795       0.2434       0.2290       0.2881
```

# 13.3 주성분회귀

- R의 {pls} 패키지에서는 PCR 분석에 사용되는 다양한 함수를 제공한다. 몇 가지를 소개하면 다음과 같다. R의 {chemometrics} 패키지도 PCR을 비롯한 다양한 다변량 분석을 제공한다.
  - validationplot() : 성분의 수에 따른 타당성(validation) 통계량(RMSEP, MSEP 또는 $R^2$ 등)을 그려준다.
  - predplot() : 측정값에 대한 적합모형의 예측값을 그려준다.
  - coefplot() : PCR과 PLSR 모형(mvr 객체)의 회귀계수를 그려준다.
  - scoreplot() : 점수(scores), 부하량(loadings)과 상관부하량(correlation loadings)을 그려준다.
  - loadingplot() : scoreplot()과 유사하다.

# 13.3 주성분회귀

- PCR의 장점과 단점을 요약하면 다음과 같다.

- PCR 분석의 장점

(i ) 차원축소

(ii ) 예측변수 간의 다중공선성 제거

(iii ) 과적합의 완화: 반응변수와 관련된 대부분의 변동과 정보가 주성분에 압축되어 있고, 적은 수의 모수 추정을 통해 과적합의 위험을 줄일 수 있다.

- PCR 분석의 단점

(i ) 변수선택 기능을 가지지 않으며, 예측변수의 영향을 파악하기 어렵다.

(ii ) 주성분의 생성과정에 반응변수가 고려되지 않으므로(비지도 학습), 반응변수의 예측을 위한 최선의 방법이라 할 수는 없다.