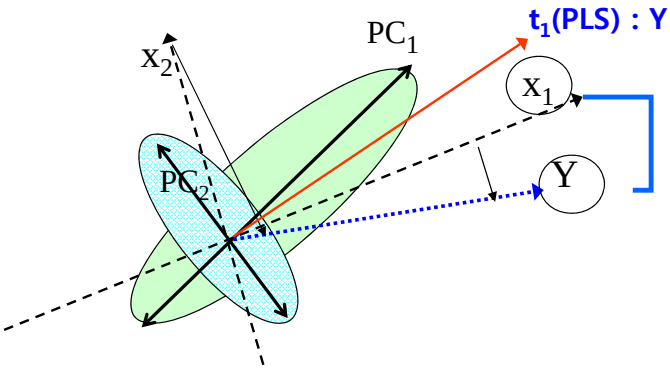
Wk15-3 : Partial Least Square Regression

1. Partial Least Square Regression (PLS)

- 주성분분석의 component 와 최소자승회귀법의 component의 비교
  - PLS는 공정변수들의 변동을 설명하는 벡터 t를 구하는데 X의 정보만을 이용하는 것이 아니라 타겟변수 *y*의 정보를 동시에 고려

Latent variable (LV)



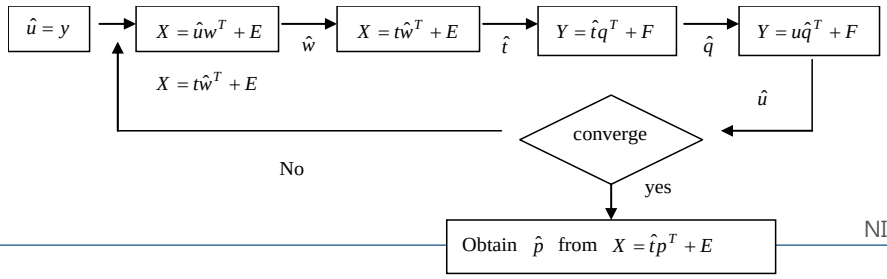
$$X = TP^T + E$$

$$Y = UQ^T + F$$

1. Partial Least Square Regression (PLS)

- PLS에서의 component는 PCR와 다르게 X의 정보만을 이용하는 것이 아니라 타겟변수(Y)와의 상관성을 고려하여 도출된다.
- Chemometrics, Marketing분야의 고차원데이터, 독립변수간 상관성 높은 데이터에 적용

- t (components)는 X들의 벡타의 선형조합으로 산출하는데 이 때 가중치로서 벡터 x_j를 벡터 y에 projection한 weight를 이용한다. 이는 NIPALS Algorithm으로 산출된다.



NIPALS : Nonlinear Iterative Partial Least Squares

1. Partial Least Square Regression (PLS)

- PLS를 수행하기 위한 추가 패키지 설치

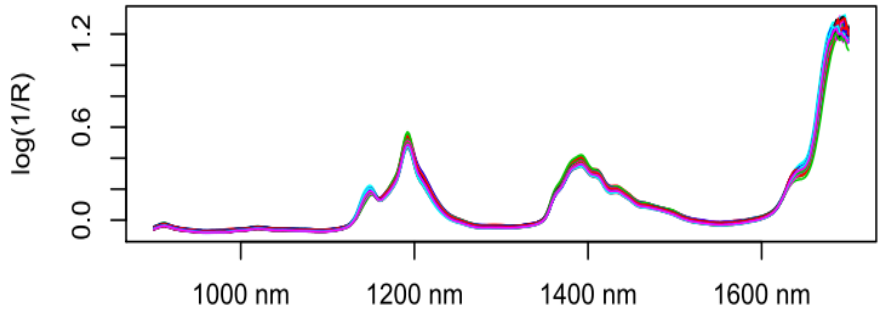
```
# lec15_3_pls.r
# Partial Least Square
# package : pls

# install package for Partial Least Square
install.packages('pls')
library(pls)

# set working directory
setwd("D:/tempstore/moocr/wk15")

# example PLS with gasoline data
data(gasoline)
help("gasoline")
```

- PLS 수행을 위한 패키지 : "pls"

- pls 패키지에 탑재된 데이터 gasoline 사용
- 여기서 data(gasoline)은 데이터를 load한다는 의미

1. Partial Least Square Regression (PLS)

- 데이터 설명 – 가솔린 데이터* (근적외선 흡광도, 60개의 가솔린 표본)
  - 독립변수의 차원 : 401
  - 타겟변수(Y) : 옥탄가(octane numbers)



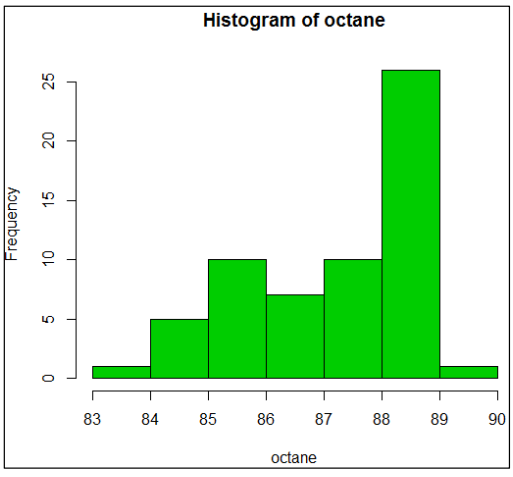
The NIR spectra were measured using diffuse reflectance as log(1/R) from 900 nm to 1700 nm in 2 nm intervals, giving 401 wavelengths.
ref : Kalivas, John H. (1997) Two Data Sets of Near Infrared Spectra *Chemometrics and Intelligent Laboratory Systems*, **37**, 255–259.

1. Partial Least Square Regression (PLS)

- 데이터 요약 설명 (타겟변수 Y : 옥탄가)

```
# descriptive statistics
par(mfrow=c(1,1))
hist(octane, col=3)
summary(octane)
```

```
> summary(octane)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  83.40   85.88   87.75   87.18   88.45   89.60
```



- 옥탄가의 최소값 83.4, 최대값 89.6
- 히스토그램은 옥탄가의 분포를 보여줌

1. Partial Least Square Regression (PLS)

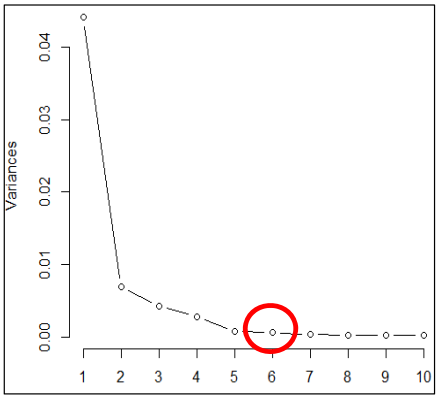
- 훈련데이터와 검증데이터 (50개 /10개)

```
# train and test set
gasTrain <- gasoline[1:50, ]
gasTest <- gasoline[51:60, ]
```

- 주성분분석에서는 최적 성분수?

```
# 1.check how many principal components
ga.pca<-prcomp(gasoline$NIR,center=T,scale.=F)
ga.pca
summary(ga.pca)
plot(ga.pca,type="l")
```

최소 5개정도의 PC는 사용




```
> summary(ga.pca)
Importance of components:
                            PC1     PC2     PC3     PC4     PC5     PC6
Standard deviation      0.2101 0.08306 0.06505 0.05291 0.02747 0.02426
Proportion of Variance  0.7257 0.11338 0.06954 0.04600 0.01240 0.00967
Cumulative Proportion   0.7257 0.83903 0.90857 0.95457 0.96698 0.97664
                            PC7     PC8     PC9     PC10    PC11     PC12
Standard deviation      0.01734 0.01485 0.01422 0.01189 0.01106 0.008421
Proportion of Variance  0.00494 0.00363 0.00332 0.00232 0.00201 0.001170
Cumulative Proportion   0.98158 0.98521 0.98853 0.99085 0.99286 0.994030
```


1. Partial Least Square Regression (PLS)

- PLS함수 : plsr

```
# pls function
help(plsr)
```

Partial Least Squares and Principal Component Regression

Description

Functions to perform partial least squares regression (PLSR), canonical powered partial least squares (CPPLS) or principal component regression (PCR), with a formula interface. Cross-validation can be used. Prediction, model extraction, plot, print and summary methods exist.

Usage

```
mvr(formula, ncomp, Y.add, data, subset, na.action,
    method = pls.options()$mvralg,
    scale = FALSE, validation = c("none", "CV", "LOO"),
    model = TRUE, x = FALSE, y = FALSE, ...)
plsr(..., method = pls.options()$plsralg)
cppls(..., Y.add, weights, method = pls.options()$cpplsalg)
pcr(..., method = pls.options()$pcralg)
```

Arguments

formula	a model formula. Most of the lm formula constructs are supported. See below.
ncomp	the number of components to include in the model (see below).
Y.add	a vector or matrix of additional responses containing relevant information about the

1. Partial Least Square Regression (PLS)

- PLS함수 : plsr(타겟변수~독립변수, ncomp= , data=

```
# pls model by training set (find LV by leave-one-out)
gas1 <- plsr(octane ~ NIR, ncomp = 10, data = gasTrain, validation = "LOO")

summary(gas1)
```

옵션사항 :
ncomp : 잠재변수의 수
validation=c("none", "CV", "LOO")

CV : cross-validation
LOO : leave-one-out

octane~NIR

NIR에 401차원의 값이 들어있음

1. Partial Least Square Regression (PLS)

- PLS결과 (1개의 잠재변수-> 10개의 잠재변수)

```
> summary(gas1)
Data:   X dimension: 50 401
        Y dimension: 50 1
Fit method: kernelpls
Number of components considered: 10

VALIDATION: RMSEP
Cross-validated using 50 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           1.545    1.357   0.2966   0.2524   0.2476   0.2398
adjCV        1.545    1.356   0.2947   0.2521   0.2478   0.2388
       6 comps  7 comps  8 comps  9 comps  10 comps
CV      0.2319   0.2386   0.2316   0.2449    0.2673
adjCV   0.2313   0.2377   0.2308   0.2438    0.2657
```

1개의 잠재변수-> 평균오차 1.357
2개의 잠재변수-> 평균오차 0.297
3개의 잠재변수-> 평균오차 0.252

```
TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X         78.17    85.58    93.41    96.06    96.94    97.89
octane    29.39    96.85    97.89    98.26    98.86    98.96
        7 comps  8 comps  9 comps  10 comps
X         98.38    98.85    99.02     99.19
octane    99.09    99.16    99.28     99.39
```

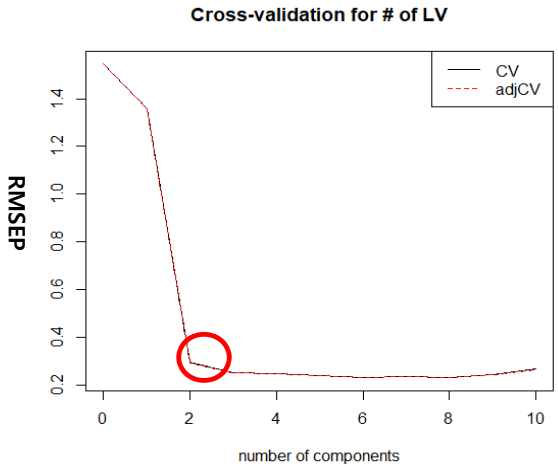
X들의 분산설명비율 : 2개의 LV로 85.58%

Y값의 변동분 설명비율 : 96.85% 설명

1. Partial Least Square Regression (PLS)

• PLS모형에서의 최적 잠재변수의 수 :

```
# 2. to choose the number of components
plot(RMSEP(gas1), legendpos = "topright", pch=46, cex=1.0, main="Cross-v
# for gasoline data, # of LV=2
```



최적 잠재변수의 수는 RMSEP가 최저이고 변화가 없는 지점에서 결정
=> 2개의 components (LV)를 추천

예측모형 평가척도 : 평균오차

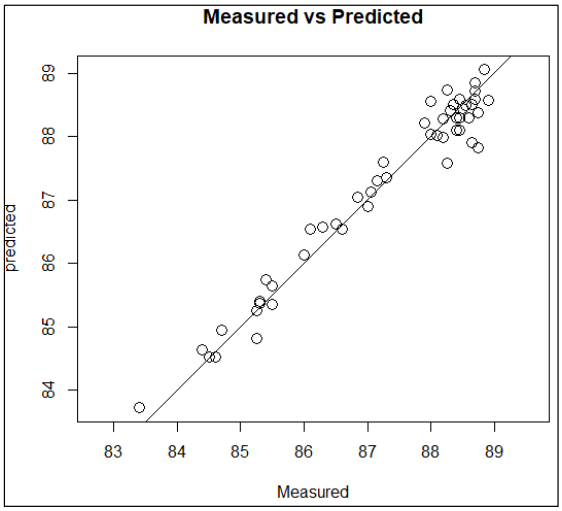
$$RMSEP=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n}}$$

root mean squared error of prediction

1. Partial Least Square Regression (PLS)

• 최적 PLS모형의 실제값과 예측값 산점도

```
# 3. Display the PLS model with LV=2
# scatterplot with true and predicted
plot(gas1, ncomp = 2, asp = 1, line = TRUE, cex=1.5,main="Measured vs
```



1. Partial Least Square Regression (PLS)

- 잠재변수 수에 따른 전체분산의(독립변수들) 설명정도

```
# Check explained variances proportion for X
explvar(gas1)
```

```
> explvar(gas1)
   Comp 1    Comp 2    Comp 3    Comp 4    Comp 5    Comp 6
78.1707683  7.4122245 7.8241556 2.6577773 0.8768214 0.9466384
   Comp 7    Comp 8    Comp 9   Comp 10
 0.4921537 0.4723207 0.1688272 0.1693770
```

2개의 잠재변수가 => 분산 전체의 85.58% 설명

1. Partial Least Square Regression (PLS)

- 검증데이터의 RMSEP계산

```
# 4. predicted Y for test data
ypred<-predict(gas1, ncomp = 2, newdata = gasTest)

y<-gasoline$octane[51:60]

# check : RMSEP for test data
sqrt((sum(y-ypred)^2)/10)
```

```
> sqrt((sum(y-ypred)^2)/10)
[1] 0.2442074
```

```
# 5. compare with the one from #4 : RMSEP for test data
RMSEP(gas1, newdata = gasTest)
```

```
> RMSEP(gas1, newdata = gasTest)
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps
     1.5369       1.1696       0.2445       0.2341       0.3287       0.2780
    6 comps      7 comps      8 comps      9 comps     10 comps
     0.2703       0.3301       0.3571       0.4090       0.6116
```

1. Partial Least Square Regression (PLS)

- PLS 예측값 내보내기

```
# output of y and predicted y
out1<-cbind(y, ypred)
# data exporting
write.csv(out1,file="out1.csv", row.names = FALSE)
```

	A	B
1	y	ypred
2	88.1	87.94125
3	87.6	87.25242
4	88.35	88.15832
5	85.1	84.96913
6	85.1	85.15396
7	84.7	84.51415
8	87.2	87.5619
9	86.6	86.84622
10	89.6	89.18925
11	87.1	87.09116

PLS 예측값 내보내기
D:/tempstore/moocr/wk15/out1.csv가 저장됨