ZIP Practical Work : PLS2

Multivariate Modeling

Martin Guy

December 31, 2016

Introduction

We have normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service in 16 x 16 grayscale images (from -1 to 1). The goal of this exercise is to recognize the right number that is written. The purpose is to continue the exercise we did for session 1 using Multivariate Regression and a **Principal Components Regression** and in session 2 using **Inter-Battery Analysis**. Now we will try **Partial Least Square 2** (PLS2) as a component based methodology to predict the digits.

1. Read the "zip_train.dat" and "zip_test.dat" files provided.

```r
train.full <- read.table("zip_train.dat")
test <- read.table("zip_test.dat")
```

Select a 5% random sample (without replacement) of the train data

```r
proportion <- 0.05
n.full <- dim(train.full)[1]  #original size of the training set
n <- floor(n.full * proportion) #new size of the training set

train <- train.full[sample(n.full,n),]
```

2. Define the response matrix (Y) and the predictor matrix (X).

```r
X.train <- as.matrix(train[,-1])
X.test <- as.matrix(test[,-1])

Y.train <- class.ind(train[,1])
Y.test <- class.ind(test[,1])
```

Center the predictor matrix.

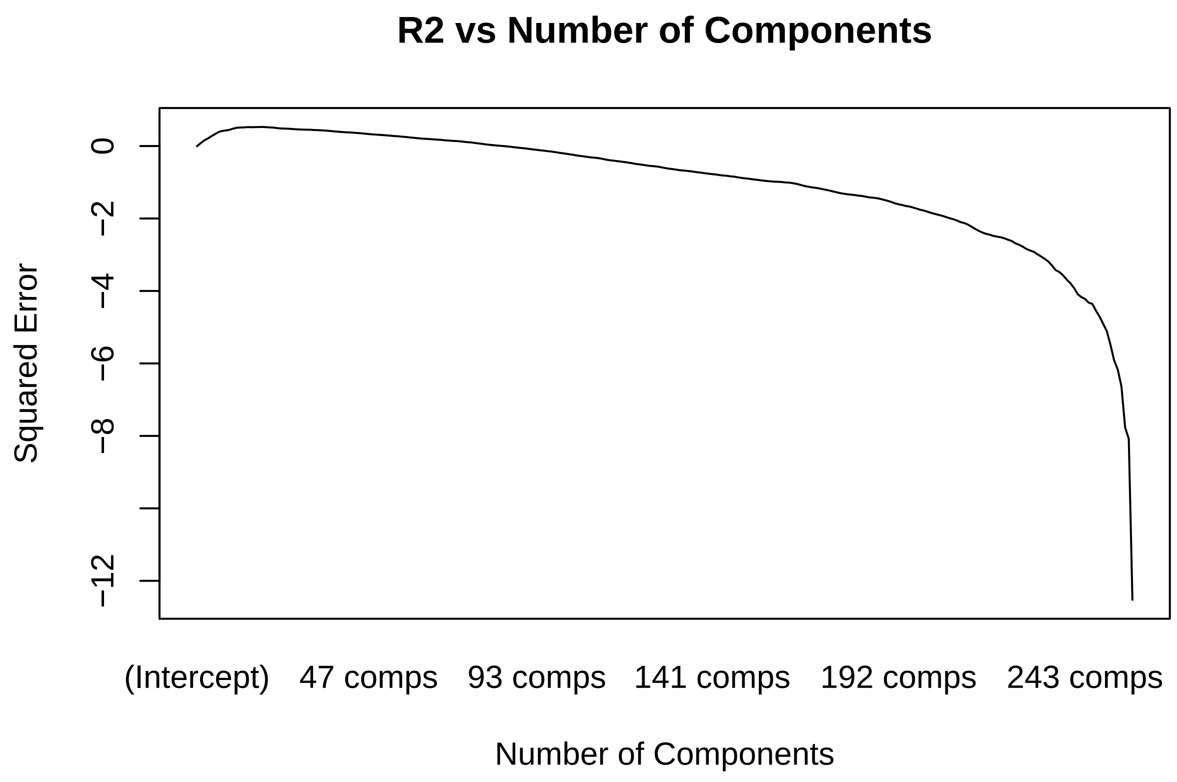
We center but not scale the data

```r
X.train.means <- colMeans(X.train)
X.train <-scale(X.train, scale=FALSE)
```

3. Perform a PLSR2 using "CV" or "LOO" for validation.

I decided to use cross-validation as it computes way faster than Leave-One-Out.

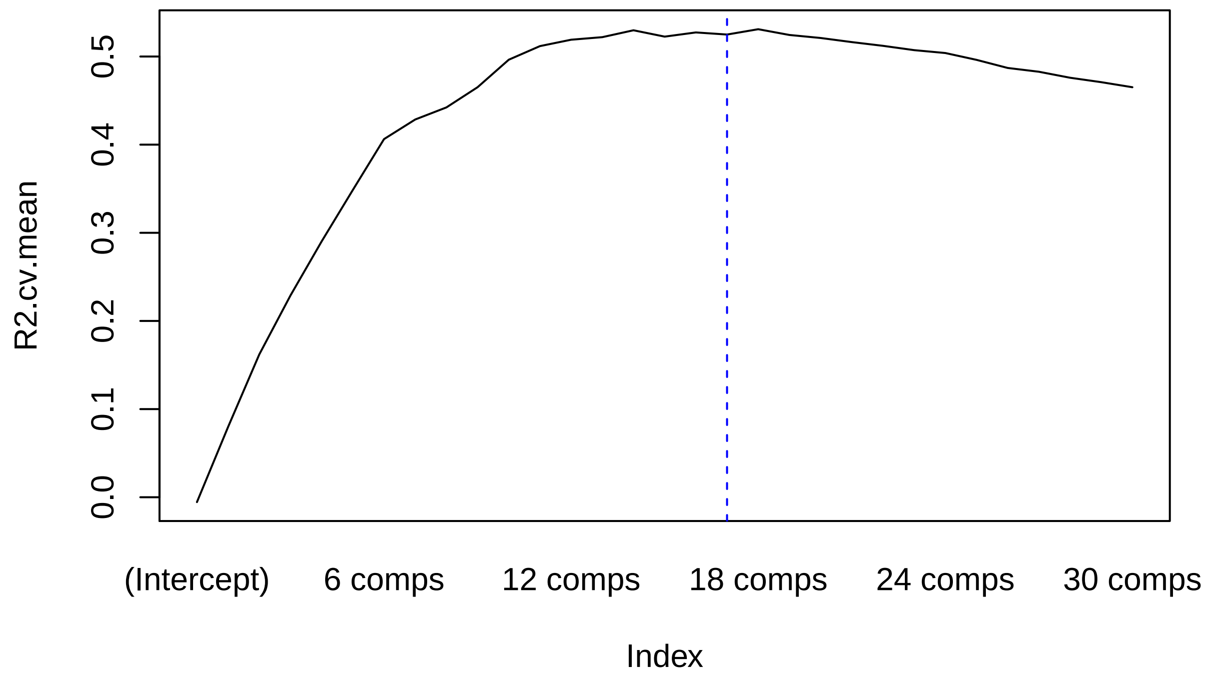
```r
pls1 <- plsr(Y.train ~ X.train, validation="CV")

# Plot the components importance
R2.cv <- R2(pls1)$val[1,,]
R2.cv.mean <- apply(R2.cv,2,mean)

plot(R2.cv.mean, main = 'R2 vs Number of Components',xlab = 'Number of Components', ylab = 'Squared Err
axis(1,at=1:ncol(R2.cv),lab=colnames(R2.cv),tick=FALSE)
```



We observe this is getting worse as we take more and more components. Then **we can zoom more**. Let's zoom to 30 components.

```r
pls2 <- plsr(Y.train ~ X.train, ncomp=30, validation="CV")



R2.cv <- R2(pls2)$val[1,,]
```

```r
R2.cv.mean <- apply(R2.cv,2,mean)
plot(R2.cv.mean,type="l",xaxt="n")
axis(1,at=1:ncol(R2.cv),lab=colnames(R2.cv),tick=FALSE)

nd <- which.max(R2.cv.mean)-1
abline(v = nd, lty=2, col="blue")
```



```r
print(paste("Number of selected components:", nd))
```

```
## [1] "Number of selected components: 18"
```

Then, **we select 18 components** here as the mean of the R2 value is the greatest. Let's see now how much of the variance is explained with this number of components:

```r
var.exp <- rep(0, nd)
curr <- 0


for(i in 1:nd)
{
  curr <- curr + pls2$Xvar[i]
  var.exp[i] = curr/pls2$Xtotvar
}
var.exp <- var.exp*100

print(paste("Variance explained with ", nd," components: ", round(var.exp[nd]*100)/100, "%", sep=""))
```

```
## [1] "Variance explained with 18 components: 69.73%"
```

4. Predict the responses in the test data, be aware of the appropriate centering.

```
X.test <- scale(X.test, center = X.train.means, scale=FALSE)
test.proj <- as.matrix(X.test) %*% pls2$projection[, 1:nd]

train.pls.data <- data.frame(pls2$scores[,1:nd])
model <- lm(Y.train~., data=train.pls.data)
pred.test.prob <- predict(model, data.frame(test.proj) , type="response")
```

Compute the average R2 in the test data.

We find this R2 value:

```
test_scale <- data.frame(test.proj)
Yhat = predict(model,test_scale)
RSS = colSums((Y.test-Yhat)^2)
TSS = apply(Y.test,2,function(x){sum((x-mean(x))^2)})
(r2 <- mean(1 - (RSS/TSS)))
```

```
## [1] 0.4735733
```

5. Assign every test individual to the maximum response and compute the error rate.

```
pred.test.numbers <- c()

n.test <- nrow(pred.test.prob)
```

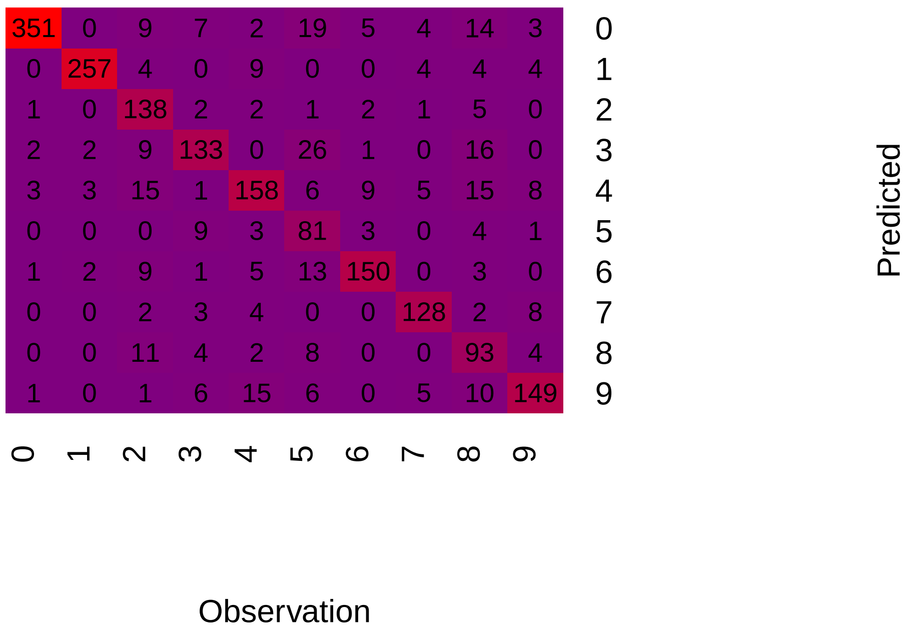
```
for(i in 1:n.test)
{
  pred.test.numbers[i] <- unname(which.max(pred.test.prob[i,])-1)
}

(pred.acc <- mean(pred.test.numbers == test[,1])) #accuracy
```

```
## [1] 0.8161435
```

```
(pred.err <- 1-pred.acc) #error rate
```

```
## [1] 0.1838565
```

We obtained this confusion matrix:



Predicted \ Observation	0	1	2	3	4	5	6	7	8	9
0	351	0	9	7	2	19	5	4	14	3
1	0	257	4	0	9	0	0	4	4	4
2	1	0	138	2	2	1	2	1	5	0
3	2	2	9	133	0	26	1	0	16	0
4	3	3	15	1	158	6	9	5	15	8
5	0	0	0	9	3	81	3	0	4	1
6	1	2	9	1	5	13	150	0	3	0
7	0	0	2	3	4	0	0	128	2	8
8	0	0	11	4	2	8	0	0	93	4
9	1	0	1	6	15	6	0	5	10	149

Comparison with previous works

Using PCR, we had around **79%** of accuracy. With IBA, we had around **80%** of accuracy. Now, we have nearly **82%**, which is better.

Conclusion

We can observe that our model here is slightly better than previous work. Moreover, PLS2 algorithm seems really powerful as it reduces here a dataset of 256 variables to 18 components with the most variability.