

Statistics 452: Statistical Learning and Prediction

Chapter 6, Part 3: Dimension Reduction Methods

Brad McNeney

2017-10-07

Reduced Dimension Regression

- ▶ Transform predictors X_1, \dots, X_p to a lower-dimension set Z_1, \dots, Z_M , for $M < p$.
 - ▶ The Z_m 's are taken to be linear combinations of the X_j 's:

$$Z_m = \sum_{j=1}^p \phi_{jm} X_j$$

- ▶ Fit a linear model to Z_1, \dots, Z_M

$$Y = \theta_0 + \sum_{m=1}^M \theta_m Z_m + \epsilon.$$

Compare to the linear model

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon.$$

- ▶ Fewer regression coefficients ($M + 1 < p + 1$).

Lower Dimension, Constraint on β 's

- ▶ As shown on pages 229,230 of the text, the lower-dimension model implies coefficients in the original model of the form

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{jm}$$

- ▶ Thus the p β s are constrained to be functions of M underlying θ s.
 - ▶ Different form of constraints from those in ridge regression and the lasso (recall the second view of these as constrained maximization).
- ▶ Introduction of a constraint is another way to view the bias/variance trade-off:
 - ▶ constraints mean lower variance, but higher bias on parameter estimates, which translates into lower variance/higher bias for predictions.

Methods for Dimension Reduction

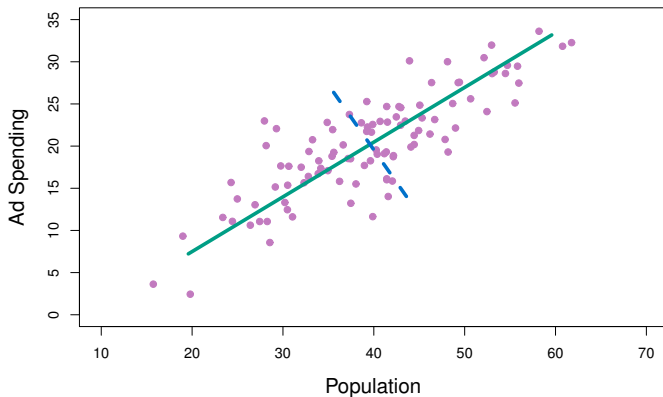
- ▶ Principal components – low-rank approximation of the X data matrix
- ▶ Partial least squares – explain X by latent variables

Principal Components Analysis (PCA)

- ▶ More details on PCA to follow in Chapter 10.
- ▶ First centre each variable by subtracting its mean.
- ▶ Then, think of principal components (PCs) as new coordinates for the data vectors.
 - ▶ The first PC is the direction of greatest variation,
 - ▶ The second PC is the direction of second-greatest variation, orthogonal to the first,
 - ▶ And so on.

PCs for Advertising Data

- Text Figure 6.14: The green line is the first PC, the blue line the second.



PCs as Linear Combinations of X 's

- ▶ We won't go into the details of how the linear combinations are derived.
- ▶ In the advertising example, the first PC is

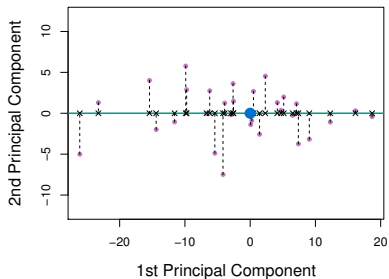
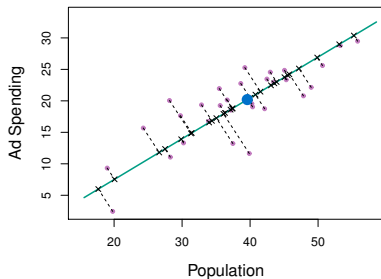
$$Z_1 = 0.838X_1 + 0.544X_2$$

where X_1 is population centred by its mean and X_2 is advertising expenditure centred by its mean.

- ▶ The coefficients of the linear combination, $\phi_{11} = 0.838$ and $\phi_{12} = 0.544$, are called the first principal component *loadings*.

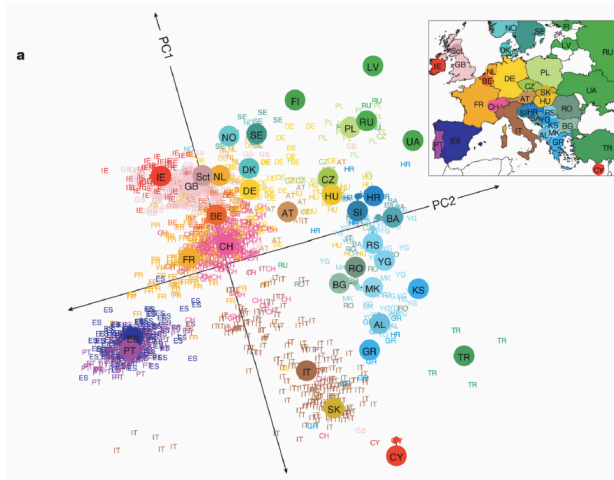
Principal Component Scores

- ▶ Projecting each point onto the PCs gives the PC scores.
 - ▶ Projecting a data vector onto a line means finding the point on the line closest to the vector.
- ▶ Text Figure 6.15: Black x's are the first PC score for each observation, distance of each purple dot from the green line is the second PC score.



High-Dimensional Example: Genes Reflect Geography

- First 2 PCs from 197,146 genetic markers on 1,387 European individuals (Novembre *et al.* 2008)



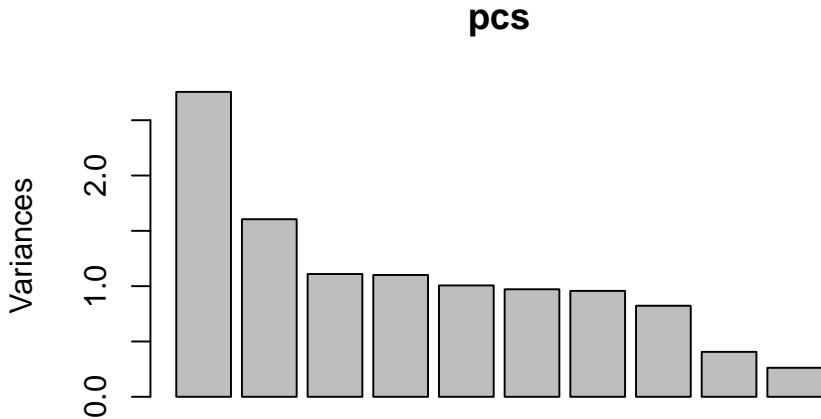
PCs and PC Scores for the Credit Data

```
uu <- url("http://www-bcf.usc.edu/~gareth/ISL/Credit.csv")
Credit <- read.csv(uu,row.names=1)
head(Credit,n=3)
```

```
##      Income Limit Rating Cards Age Education Gender Student Married
## 1  14.891  3606    283     2  34          11   Male      No      Yes
## 2 106.025  6645    483     3  82          15 Female     Yes     Yes
## 3 104.593  7075    514     4  71          11   Male      No      No
##      Ethnicity Balance
## 1 Caucasian      333
## 2    Asian      903
## 3    Asian      580
```

```
X <- model.matrix(Balance ~ ., data=Credit)
X <- X[,-1] # Remove intercept
X <- scale(X) # Centre and scale
pcs <- prcomp(X)
```

```
plot(pcs)
```

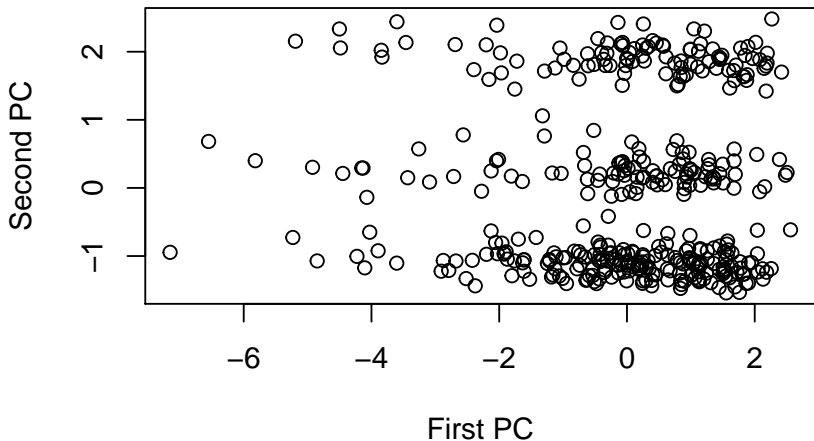


Loadings for First Two PCs

```
pcs$rotation[,1:2]
```

| ## | PC1 | PC2 |
|-----------------------|--------------|--------------|
| ## Income | -0.542206953 | 0.029036783 |
| ## Limit | -0.586332930 | 0.017502630 |
| ## Rating | -0.586751867 | 0.014971105 |
| ## Cards | -0.019086978 | -0.008549632 |
| ## Age | -0.122783390 | -0.071116603 |
| ## Education | 0.026797471 | 0.096557225 |
| ## GenderFemale | -0.002519860 | 0.052811098 |
| ## StudentYes | 0.002276904 | 0.125422970 |
| ## MarriedYes | -0.026218561 | 0.094278214 |
| ## EthnicityAsian | 0.032769895 | 0.696759512 |
| ## EthnicityCaucasian | -0.004070799 | -0.686505857 |

```
plot(pcs$x[,1],pcs$x[,2],xlab="First PC",ylab="Second PC")
```



Principal Components Regression (PCR)

- ▶ Take Z_1, \dots, Z_M to be the first M PC scores.
 - ▶ M can be chosen by cross-validation to minimize estimated test set error.
- ▶ The idea is that a handful of PCs might explain the variation in X **and** the relationship between X and Y .

PCR on the Credit Data

```
library(pls) # install.packages("pls")  
set.seed(123)  
cfit <- pcr(Balance ~ ., data=Credit, scale=TRUE,  
            validation="CV")
```

Summary

```
summary(cfit)
```

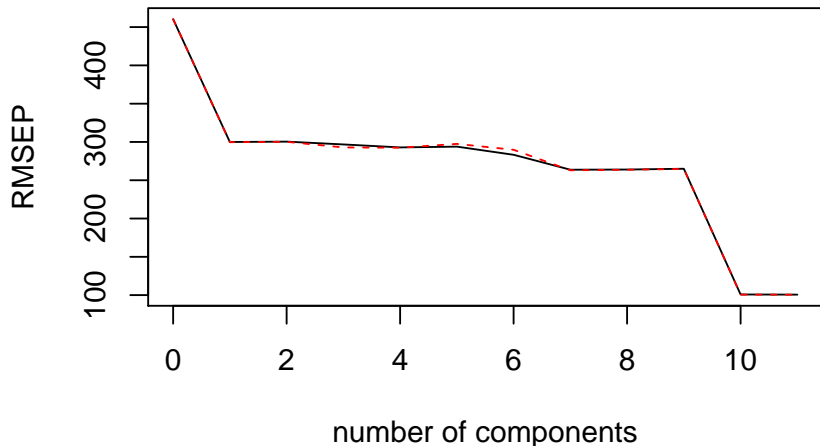
```
## Data:      X dimension: 400 11
## Y dimension: 400 1
## Fit method: svdpc
## Number of components considered: 11
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           460.3   299.9   300.4   296.8   292.9   293.9   283.3
## adjCV        460.3   299.7   300.2   292.9   292.5   297.5   289.8
##      7 comps  8 comps  9 comps 10 comps 11 comps
## CV           263.8   264.0   265.0   100.8   100.6
## adjCV        263.0   263.7   264.8   100.7   100.5
##
## TRAINING: % variance explained
##      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X           25.05   39.64   49.73   59.74   68.89   77.73   86.43
## Balance     58.07   58.37   60.78   60.90   61.46   63.11   68.70
##      8 comps  9 comps 10 comps 11 comps
## X           93.91   97.60   99.98   100.00
## Balance     68.71   68.72   95.47   95.51
```

- Note: RMSEP is root mean squared error of prediction, the square root of the MSE.

Plot the Root MSE of Prediction

```
validationplot(cfit)
```

Balance



Extract $\hat{\beta}$'s

- These are the estimates of the coefficients of the X 's,

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{jm}$$

```
coef(cfit,ncomp=10)
```

```
## , , 10 comps
##
##               Balance
## Income         -275.334437
## Limit           308.685448
## Rating          308.331638
## Cards           18.588390
## Age            -10.700222
## Education       -2.758126
## GenderFemale    -5.354857
## StudentYes      127.056873
## MarriedYes      -5.131238
## EthnicityAsian   8.004166
## EthnicityCaucasian 5.143306
```

Partial Least Squares (PLS) versus PCR

- ▶ Statistical learning methods that use the response are said to be “supervised”, while those that do not are “unsupervised”.
- ▶ PCR does unsupervised selection of the transformed features Z_1, \dots, Z_M .
- ▶ By contrast, PLS is supervised (sketch of details below).
- ▶ No clear winner between PCR and PLS.
 - ▶ Supervised dimension reduction may reduce bias by identifying features that are truly related to Y .
 - ▶ However, supervising “... has the potential to increase variance,” (text, page 238)

PLS Directions

- ▶ The loadings for the first PLS direction, Z_1 are the coefficients from the simple linear regression of Y on each X_j .
- ▶ The loadings for the second PLS direction are coefficients from the simple linear regression of the *adjusted* variable $Y - \hat{Y}$ on the adjusted $X_j - \hat{X}_j$, where \hat{Y} and \hat{X}_j are from regressions on Z_1 .
 - ▶ The residuals are the information in the variables not explained by Z_1 .
- ▶ The loadings for the third PLS direction are coefficients from the simple linear regression of the adjusted variable $Y - \hat{Y}$ on the adjusted $X_j - \hat{X}_j$, where \hat{Y} and \hat{X}_j are from regressions on Z_1 **and** Z_2 .
 - ▶ The residuals are the information in the variables not explained by Z_1 and Z_2 .
- ▶ And so on.

PLS on the Credit Data

```
cfit <- plsr(Balance ~ ., data=Credit,scale=TRUE,  
            validation="CV")
```

Summary

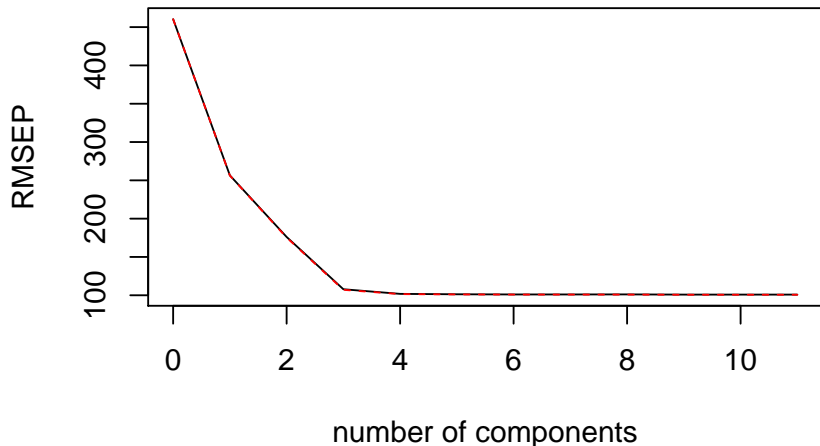
```
summary(cfit)
```

```
## Data:      X dimension: 400 11
## Y dimension: 400 1
## Fit method: kernelpls
## Number of components considered: 11
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV              460.3    256.5    176.1    107.9    101.8    101.3    101.1
## adjCV           460.3    256.3    175.3    107.1    101.5    101.1    100.9
##      7 comps  8 comps  9 comps 10 comps 11 comps
## CV          101.1    101.1    100.9    100.9    100.9
## adjCV       100.9    100.9    100.7    100.7    100.7
##
## TRAINING: % variance explained
##      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X          24.58    32.53    37.84    50.55    60.80    65.92    73.20
## Balance    69.67    86.53    94.95    95.46    95.48    95.48    95.48
##      8 comps  9 comps 10 comps 11 comps
## X          76.45    81.33    90.76    100.00
## Balance    95.50    95.51    95.51    95.51
```

Plot the Root MSE of Prediction

```
validationplot(cfit)
```

Balance



Extract $\hat{\beta}$'s

```
coef(cfit,ncomp=4)
```

```
## , , 4 comps
##
##              Balance
## Income          -274.942446
## Limit            310.143749
## Rating           306.656366
## Cards            22.106900
## Age              -11.915766
## Education        -4.175268
## GenderFemale     -7.683003
## StudentYes       125.944486
## MarriedYes       -3.676939
## EthnicityAsian   10.377071
## EthnicityCaucasian 5.060771
```