Topic # 06: Data Analytics

Dimension Reduction, Principal Components

Instructor: Prof. Arnab Bisi, Ph.D.

Johns Hopkins Carey Business School

- The setting: we have a high-dimensional matrix of data $\mathbf{X}$. We would like to reduce this to a few important factors.
- We will do this by building a simple linear model for $\mathbf{X}$ and use this model to represent $\mathbf{X}$ in a lower dimensional space.
- Factor modeling is a super useful framework, whether you get a deep understanding or just learn how they work in practice.

Principal Component Analysis

- A large set of correlated variables
- Principal Component summarizes this set with a smaller number of representative variables
- Principal Components explain most of the variability in the original set
- Principal Component Regression use principal components as predictors in the regression model

What are Principal Components?

- Wish to visualize n observations with p features $X_1, X_2, \ldots, X_p$
- Produce two-dimensional scatter plots: p(p-1)/2 plots
- Mostly likely, none of them will be informative
- We need a better method
- We need a low-dimensional representation that captures as much of the information as possible

Principal Components Analysis (PCA)

- PCA seeks a small number of dimensions that are as informative as possible
- The concept “informative” is measured by the amount of variability
- Each of the dimensions found by PCA is a linear combination of the p features
- The first principal component is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

that has the largest variance. By normalized, we mean $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

Principal Components Analysis (PCA), Cont.

- We look for linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip}$$

  that has the largest sample variance.
  - We refer $z_{11}, z_{21}, \ldots, z_{n1}$ as the scores of the first principal component.
- The PCA loading vector solves the optimization problem

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}$$

$$s.t., \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Principal Components Analysis (PCA), Cont.

- We look for the second principal component Z_2

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip}$$

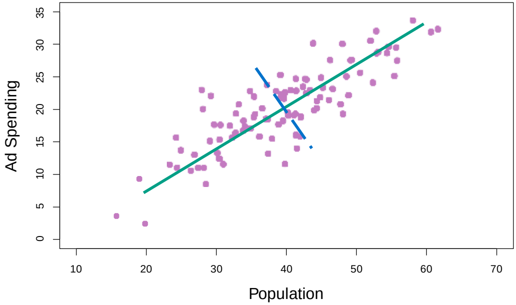
  that has largest sample variance under constraint Z_2 is uncorrelated with Z_1.

- The uncorrelated constraint is equivalent to constraining the direction ϕ_2 to the orthogonal to the direction ϕ_1.

- Each principal component loading vector is unique.

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Principal Components



- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles.
- The green solid line indicates the first principal component
- The blue dashed line indicates the second principal component.

Proportion of Variance Explained

- Question: how much of the information in a given data set is lost?
- We consider the proportion of variance explained (PVE) by each principal component
- The total variance present in a data set (assuming variables have been centered to have mean zero)

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

Proportion of Variance Explained, Cont.

- The variance explained by the *m*-th principal component is

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2$$

- Therefore, the PVE of the *m*-th principal component is given by

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

Proportion of Variance Explained, Cont.

- We want to use the smallest number of principal components to get a good understanding of the data.
- How many principal component are needed?
- No single answer!
- Produce a plot, choose the number of principal components in order to explain a sizeable amount of variation in the data.

Examples in R

- Data set USArrests

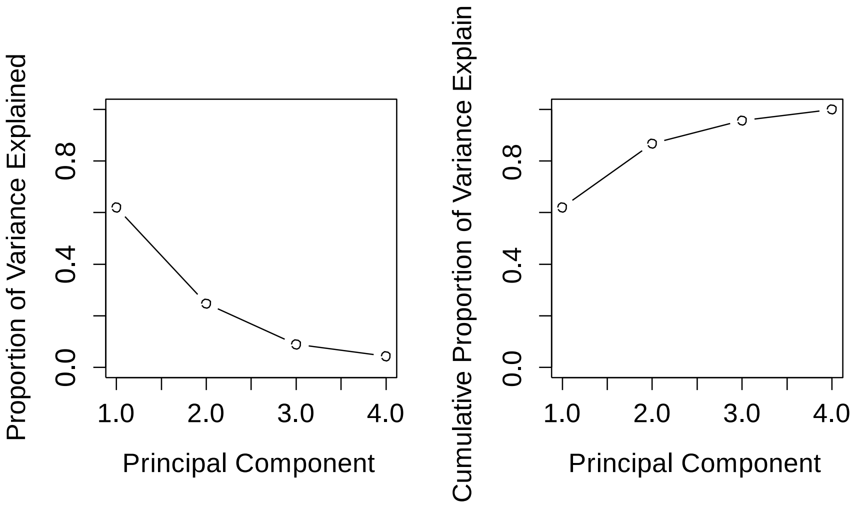
```
> states=row.names(USArrests)
> states
> names(USArrests)
> tail(USArrests)
> apply(USArrests,2,mean)
> apply(USArrests,2,var)
```

- Function `apply()`: arguments (dataset, row or column, function)
- Finding: variables are in different scales.

- Function prcomp()

```
> pr.out=prcomp(USArrests,scale=T)
> names(pr.out)
> pr.out$center
> pr.out$scale
> pr.out$rotation
> summary(pr.out)
> pr.var=pr.out$sdev^2
> pve=pr.var/sum(pr.var)
> plot(cumsum(pve))
```

- Finding: two principal variables explain 87% variability.


Proportion of Variance Explained
0.0
0.4
0.8
1.0
2.0
3.0
4.0
Principal Component
Cumulative Proportion of Variance Explain
0.0
0.4
0.8
1.0
2.0
3.0
4.0
Principal Component

- Play with other data sets: e.g., the `Boston` data set
- How much variability can be explained by the first and second principal variables?

Principal Components Regression

- A dimension reduction technique for regression
- The Principal Components Regression (PCR) approach involves constructing the first M principal components $Z_1, \ldots, Z_M$, and then use these components as the predictors in a linear regression
- The key idea: a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.
  - PCA reduces dimension, which is always good.
  - Higher variance covariates are good in regression, and we choose the top PCs to have highest variance.
  - The PCs are independent: no multicollinearity

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Cross-Validation

- Test Error rate and Training error rate can be very different.
- Hold out a subset of the data; apply statistical learning method to those hold-out data
- Randomly divide data into two parts
- Validation error rate is assessed using MSE

- *k*-fold CV randomly divide data into k folds of approximately equal size
- The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds.
- The mean squared error, MSE_1, is computed in the hold-out fold.
- Repeat *k* times: each time, a different fold is treated as a validation set.
- The *k*-fold CV estimate is computed by averaging these values

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i.$$

Principal Components Regression in R

- Principal Components Regression can be performed using function `pcr()` in library `pls`
- Determine number of principal components using eyes.
- In this data set, 7 may be a good size for principal components.
- In function `pcr()`, specify number of principal components by using ncomp=

```
> library(pls)
> library(ISLR)
> pcr.fit=pcr(Salary~.,data=Hitters,scale=T,validation="CV")
> summary(pcr.fit)

> pcr.fit2=pcr(Salary~.,data=Hitters,scale=T,validation="CV",ncomp=7)
> summary(pcr.fit2)
```

- Argument `validation="CV"` causes `pcr()` to compute the ten-fold cross-validation error for each possible M, the number of principal components used.

Note that the smallest error occurs when $M = 16$, but it is roughly the same when only one component is included.

Principal Components Regression in R, cont.

- Perform `pcr()` on a training subset
- Make predictions on test subset
- Calculate prediction errors

```
set.seed(1)
Hitters=na.omit(Hitters)
x = model.matrix(Salary~., Hitters)[,-1]
y=Hitters$Salary
train = sample(1:nrow(x), nrow(x)/2)
test=(-train)
y.test=y[test]

pcr.fit = pcr(Salary~., data= Hitters, subset=train, scale=T,
validation = "CV")
pcr.pred = predict(pcr.fit, x[test,], ncomp=7)

mean((pcr.pred - y.test)^2)
[1] 96556.22
```

Questions, Comments?

See you next time.