Data621 - Blog1

Amit Kapoor

3/28/2021

Principal Component Analysis

Sometimes we have too many predictors and if we use all of them in our regression model, we would end up with issues and explanation could be difficult due to collinearity. It could also cause prediction performance degradation by using too many predictors. Hence, it has been proven better to reduce dimension of the data to fetch meaningful, appropriate and valid results.

Principal components analysis (PCA) is one of a family of techniques to deal with high-dimensional data by using high dimensional data and its variable's dependencies to represent it in a lower dimensional form without losing too much information. PCA is one of the simplest ways of doing dimensionality reduction. Here components are independent. This is a method of extracting information from higher dimensional data by representing it to lower dimension. It does this using a linear combination (weighted average) of a set of given variables and the created index variables are called principal components.

Steps to perform PCA

- Standardize the data - make all the feature variables to follow same scale.
- Find the covariance matrix of the features - covariance matrix has coveriance between the features.
- Do perform eigen decompositon on the covariance matrix - decomposition gives the eigenvectors (principal components) and eigenvalues of the covariance matrix.
- Select pricipal components - Sort based on the magnitude of their corresponding eigenvalues to select principal components
- Find the number (m) of top principal components.
- Make the projection matrix from the selected number of top principal components.
- Find the new m-dimensional feature space.

R Application

To demonstrate the PCA, we will consider Boston housing dataset that has below variables. This dataset has 506 records and total 14 variables.

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centers
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town 12.

- B: 1000(Bk−0.63)2 where Bk is the proportion of blacks by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000s

```
# housing data
housing <- fread("https://raw.githubusercontent.com/amit-kapoor/data621/main/blog1/housing.csv", header
# assign column names
colnames(housing) <- c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B","LSTA
head(housing)
```

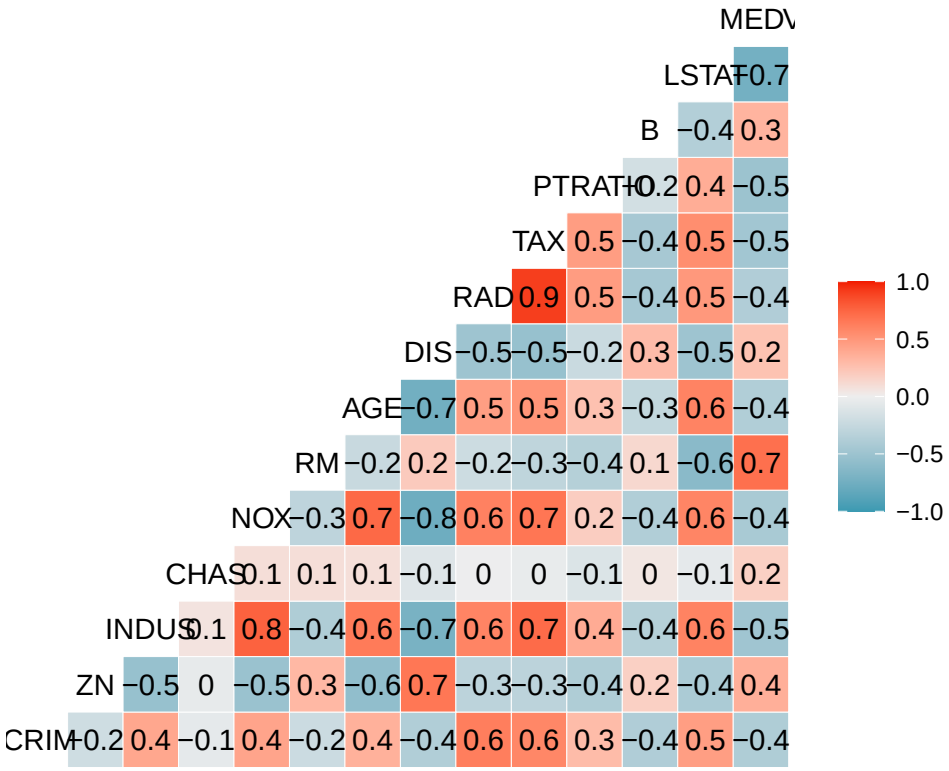
```
##       CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX PTRATIO      B LSTAT
## 1: 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2: 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3: 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4: 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5: 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6: 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##    MEDV
## 1: 24.0
## 2: 21.6
## 3: 34.7
## 4: 33.4
## 5: 36.2
## 6: 28.7
```

```
# data dimesnion
dim(housing)
```

```
## [1] 506  14
```

```
# correlation
ggcorr(housing, label = TRUE) + labs(title = "Correlation of variables")
```

Correlation of variables



We can see here that there are variables which are highly correlated,

```
# describe data
describe(housing)[-c(1)]
```

```
##            n   mean     sd median trimmed    mad    min    max  range  skew
## CRIM     506   3.61   8.60   0.26    1.68   0.33   0.01  88.98  88.97  5.19
## ZN       506  11.36  23.32   0.00    5.08   0.00   0.00 100.00 100.00  2.21
## INDUS    506  11.14   6.86   9.69   10.93   9.37   0.46  27.74  27.28  0.29
## CHAS     506   0.07   0.25   0.00    0.00   0.00   0.00   1.00   1.00  3.39
## NOX      506   0.55   0.12   0.54    0.55   0.13   0.38   0.87   0.49  0.72
## RM       506   6.28   0.70   6.21    6.25   0.51   3.56   8.78   5.22  0.40
## AGE      506  68.57  28.15  77.50   71.20  28.98   2.90 100.00  97.10 -0.60
## DIS      506   3.80   2.11   3.21    3.54   1.91   1.13  12.13  11.00  1.01
## RAD      506   9.55   8.71   5.00    8.73   2.97   1.00  24.00  23.00  1.00
## TAX      506 408.24 168.54 330.00  400.04 108.23 187.00 711.00 524.00  0.67
## PTRATIO  506  18.46   2.16  19.05   18.66   1.70  12.60  22.00   9.40 -0.80
## B        506 356.67  91.29 391.44  383.17   8.09   0.32 396.90 396.58 -2.87
## LSTAT    506  12.65   7.14  11.36   11.90   7.11   1.73  37.97  36.24  0.90
## MEDV     506  22.53   9.20  21.20   21.56   5.93   5.00  50.00  45.00  1.10
##          kurtosis   se
## CRIM        36.60 0.38
## ZN           3.95 1.04
## INDUS       -1.24 0.30
## CHAS         9.48 0.01
## NOX         -0.09 0.01
## RM           1.84 0.03
## AGE         -0.98 1.25
```

```
## DIS         0.46 0.09
## RAD        -0.88 0.39
## TAX        -1.15 7.49
## PTRATIO    -0.30 0.10
## B           7.10 4.06
## LSTAT       0.46 0.32
## MEDV        1.45 0.41
```

Next we will use prcomp function that performs a principal components analysis on the given data matrix and returns the results.

```
pca_housing <- prcomp(housing, center = TRUE, scale. = TRUE)

summary(pca_housing)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.5585 1.2843 1.16142 0.94156 0.92244 0.81241 0.73172
## Proportion of Variance 0.4676 0.1178 0.09635 0.06332 0.06078 0.04714 0.03824
## Cumulative Proportion  0.4676 0.5854 0.68174 0.74507 0.80585 0.85299 0.89123
##                            PC8    PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.63488 0.5266 0.50225 0.4613 0.42777 0.36607 0.24561
## Proportion of Variance 0.02879 0.0198 0.01802 0.0152 0.01307 0.00957 0.00431
## Cumulative Proportion  0.92003 0.9398 0.95785 0.9730 0.98612 0.99569 1.00000
```

```
# $x - principal components
dim(pca_housing$x)
```

```
## [1] 506  14
```

```
# std. deviations
pca_housing$scale
```

```
##        CRIM          ZN       INDUS        CHAS         NOX          RM
##   8.6015451  23.3224530   6.8603529   0.2539940   0.1158777   0.7026171
##         AGE         DIS         RAD         TAX     PTRATIO           B
##  28.1488614   2.1057101   8.7072594 168.5371161   2.1649455  91.2948644
##       LSTAT        MEDV
##   7.1410615   9.1971041
```

```
# means
pca_housing$center
```

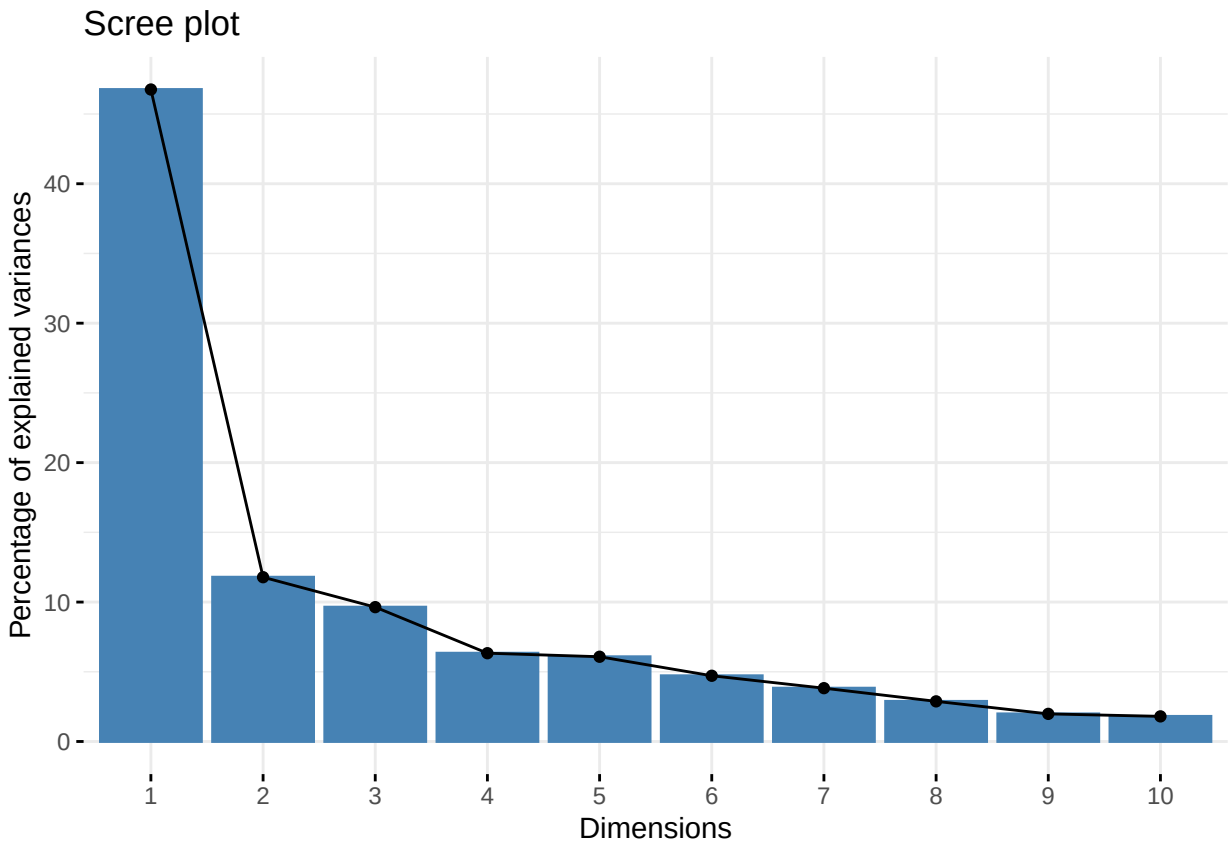
```
##         CRIM           ZN        INDUS         CHAS          NOX           RM
##   3.61352356  11.36363636  11.13677866   0.06916996   0.55469506   6.28463439
##          AGE          DIS          RAD          TAX      PTRATIO            B
##  68.57490119   3.79504269   9.54940711 408.23715415  18.45553360 356.67403162
##        LSTAT         MEDV
##  12.65306324  22.53280632
```

```
# first PCA component
round(pca_housing$rot[,1],2)
```

```
##    CRIM      ZN   INDUS    CHAS     NOX      RM     AGE     DIS     RAD     TAX
##    0.24   -0.25    0.33   -0.01    0.33   -0.20    0.30   -0.30    0.30    0.32
## PTRATIO       B   LSTAT    MEDV
##    0.21   -0.20    0.31   -0.27
```

Next we will see scree plot which is a line plot of the eigen values of principal components.

```
#scree plot
fviz_eig(pca_housing)
```



Finally we will fit the models first having full model using the original data and second using principal components (first 3) identified above.

```
set.seed(317)
# fit model using where we use all predictors
housing.fullmodel <- lm(MEDV ~ ., data = housing)
summary(housing.fullmodel)
```

```
##
## Call:
## lm(formula = MEDV ~ ., data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
## ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
## INDUS        2.056e-02  6.150e-02   0.334 0.738288
## CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
## NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
```

```
## AGE          6.922e-04  1.321e-02   0.052 0.958229
## DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
## PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## B            9.312e-03  2.686e-03   3.467 0.000573 ***
## LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
set.seed(317)
# fit model using first 3 Prinipal components
housing.pcamodel <- lm(housing$MEDV ~ pca_housing$x[,1:3])
summary(housing.pcamodel)
```

```
##
## Call:
## lm(formula = housing$MEDV ~ pca_housing$x[, 1:3])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0345  -2.1015  -0.0748   1.6409  24.3703
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              22.53281    0.17015  132.43   <2e-16 ***
## pca_housing$x[, 1:3]PC1 -2.45228    0.06657  -36.84   <2e-16 ***
## pca_housing$x[, 1:3]PC2  4.09202    0.13261   30.86   <2e-16 ***
## pca_housing$x[, 1:3]PC3  1.50086    0.14664   10.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.827 on 502 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.8268
## F-statistic: 804.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Comparing the full model with the PCA model, it is evident that PCA explains close to 83% of the variability with just three variables than the 13 significant variables from the full model which has R^2=0.73.

References

- https://www.kaggle.com/kashettivir/the-boston-housing-dataset
- https://www.youtube.com/watch?v=kw9R0nD69OU