Principal component analysis

Lucy Ng'ang'a

October 15, 2018

Principal Components Analysis

Introduction

- Sometimes data are collected on a large number of variables from a single population, this is usually known as the curse of dimensionality. With a large number of variables, the dispersion matrix may be too large to study and interpret properly.
- There would be too many pairwise correlations between the variables to consider. Graphical display of data may also not be of particular help in case the data set is very large.
- To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables(*dimensionality*) to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component.
- The principal components analysis is specifically useful in regression in **reducing the number of predictor variables** and dealing with **correlated predictor variables/Multicollinearity**.

Principal Components

Suppose that we have a random vector $X^T = (X_1, X_2, ..., X_p)$ with a population variance-covariance matrix

$$Var(X) = \sum = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}^2 \end{bmatrix}$$

?????????Consider the linear combinations

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$
$$Y_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p$$
$$\vdots$$
$$Y_p = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p$$

???

Principal Components Analysis

Each of the Y_i is a function of random data and thus it is random. With a population variance of

$$Var(Y_i) = \sum_{k=1} \sum_{l=1} e_{ik} e_{il} \sigma_{kl} = e_i^{'} \sum e_i$$

Moreover the covariance of Y_i and Y_j will have a population covariance

$$Cov(Y_i, Y_j) = \sum_{k=1} \sum_{l=1} e_{ik} e_{jl} \sigma_{kl} = e_i^{'} \sum e_j$$

And the coefficients e_{ij} are collected into a vector

$$e_i^T = (e_{i1}, e_{i2}, \cdots, e_{ip})$$

First Principal Component

- The First principal component is a linear combination of original variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component.
- It is a a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.
- Mathematically, we select $e_1^T = (e_{11}, e_{12}, \cdots, e_{1p})$ that maximizes

$$Var(Y_1) = \sum_{k=1} \sum_{l=1} e_{1k} e_{1l} \sigma_{kl} = e_1^{'} \sum e_1$$

- Subject to the constraint

$$e_1^{'} e_1 = \sum_{j=1}^{p} e_{1j}^2 = 1$$

Second Principal Component

- The second principal component Y_2 is also a linear combination of original variables which captures the remaining variance in the data set and is uncorrelated with Y_1.
- Because the two components are uncorrelated, their directions should be orthogonal.
- Mathematically, we select $e_2^T = (e_{21}, e_{22}, \cdots, e_{2p})$ that maximizes

$$Var(Y_2) = \sum_{k=1} \sum_{l=1} e_{2k} e_{2l} \sigma_{kl} = e_2^{'} \sum e_2$$

- Subject to the constraint

$$e_2^{'} e_2 = \sum_{j=1}^{p} e_{2j}^2 = 1$$

- With the additional constraint

$$Cov(Y_1, Y_2) = \sum_{k=1} \sum_{l=1} e_{1k} e_{2l} \sigma_{kl} = e_1^{'} \sum e_2 = 0$$

The ith Principal Component

- We select $e_i^T = (e_{i1}, e_{i2}, \cdots, e_{ip})$ that maximizes

$$Var(Y_i) = \sum_{k=1} \sum_{l=1} e_{ik} e_{il} \sigma_{kl} = e_i^{'} \sum e_i$$

- Subject to the constraint $e_i^{'} e_i = \sum_{j=1}^{p} e_{ij}^2 = 1$
- With the additional constraint

$$Cov(Y_1, Y_i) = \sum_{k=1} \sum_{l=1} e_{1k} e_{il} \sigma_{kl} = e_1^{'} \sum e_i = 0$$

$$Cov(Y_2, Y_i) = \sum_{k=1} \sum_{l=1} e_{2k} e_{il} \sigma_{kl} = e_2^{'} \sum e_i = 0$$

$$\vdots$$

$$Cov(Y_{i-1}, Y_i) = \sum_{k=1} \sum_{l=1} e_{i-1,k} e_{il} \sigma_{kl} = e_{i-1}^{'} \sum e_i = 0$$

- Therefore all principal components are uncorrelated with one another.

The calculation of the Coefficients

- We find the coefficients e_{ij} for a principal component by calculating the eigenvalues and eigenvectors of the variance-covariance matrix $\sum$.
- Let λ_1 through λ_p denote the eigenvalues of the variance-covariance matrix $\sum$.With the corresponding eigenvectors e_1 through to e_p.
- These are ordered so that λ_1 has the largest eigenvalue and λ_p is the smallest.

$$\lambda_1 \geq \lambda_2 \geq, \cdots \geq \lambda_p$$

- The elements for these eigenvectors will be the coefficients of our principal components.
- The variance for the ith principal component is equal to the ith eigenvalue.

$$Var(Y_i = Var(e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p) = \lambda_i$$

- Moreover, the principal components are uncorrelated with one another.

Spectral Decomposition

- The variance-covariance matrix may be written as a function of the eigenvalues and their corresponding eigenvectors. This is determined by using the Spectral Decomposition Theorem.
- Spectral Decomposition Theorem states that the variance-covariance matrix can be written as the sum over the p eigenvalues, multiplied by the product of the corresponding eigenvector times its transpose. $\sum = \sum_{i=k}^{p} \lambda_i e_i^{'} e_i$
- The total variation of X is the trace of the variance-covariance matrix, or if you like, the sum of the variances of the individual variables. This is also equal to the sum of the eigenvalues.

$$trace(\sum) = \sigma_{11}^2 + \cdots + \sigma_{11}^2 = \lambda_1 + \cdots + \lambda_p$$

- The proportion of variation explained by the ith principal component is then going to be defined to be the eigenvalue for that component divided by the sum of the eigenvalues.

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}$$

Covariance or Correlation Matrix

- Principal components analysis is not scale invariant. That is, it is influenced by the scale of measurements.
- In situations where the multivariate data has variables that are of completely different types, then the principle components from the variance covariance matrix will depend upon the choice of measurements.
- Additionally, if there are large differences between the variances of the original variables, those with large variances will tend to dorminate the early components.
- Therefore, the principle components are extracted from the correlation matrix **R**.This is equivalent with to extracting the components from hte covariance matrix after standardizing the variables.

Consider the data set iris

```
> library(MASS)
> iris<-iris
> iris_num=iris[,1:4]#removing the factor variable.
> apply(iris_num,2,var)
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.6856935    0.1899794    3.1162779    0.5810063
```

```
> library(psych)
> pairs.panels(iris[,-5],gap=0,
+              bg=c("red","yellow","blue")[iris$Species])
```

The later plot shows the multicolliearity of the four numerical variables.

Implementation in R

```
> pc=prcomp(iris_num,center = TRUE,scale. = TRUE)
> pc$rotation #prints all the principal components

                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971

> summary(pc)

Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000

> iris_pc=pc$x #calls the
```

Analysis of components

- The first componenet explains 72.96% of the variability and first and second components 95.81%.
- The variable sepal.width had a neggative correlation with all the components. This is due to the fact that it had the least variance.
- The components are now uncorrelated

```
> cor(pc$x)
```

```
              PC1           PC2           PC3           PC4
PC1  1.000000e+00  2.906325e-16 -4.776167e-16  2.153446e-15
PC2  2.906325e-16  1.000000e+00  9.341844e-17 -2.309745e-16
PC3 -4.776167e-16  9.341844e-17  1.000000e+00 -1.384981e-15
PC4  2.153446e-15 -2.309745e-16 -1.384981e-15  1.000000e+00
```

```
> pairs.panels(pc$x,gap=0,
+              bg=c("red","yellow","blue")[iris$Species])
```

Bi-plot

A biplot is a graphical representation of the variances and covariances of the variables and the distances between units.

```
> biplot(pc,col=c("red","blue"))
```

- The distance between the points representing the units reflects the generalized distance between the points.
- The length of the vector from the origin to the coordinates representing a particular variables reflects the variance of that variable.
- The correlation of two variables is reflected by the angle between the two corresponding vectors for the two variables. The greater the angle the greater the correlation.

Choice of Components

There are informal and formal ways in answering the question of how many components are needed.

1. Retain just enough components to explain some specified large percentages of the total variation of the original variables. Values between 75%and 95% are suggested.
2. Exclude the principal components whose eigenvalues (variance) are less than than the average of the eigenvalues. If the correlation matrix was used to extract the components then the average variance is 1.
3. Examine the plot of the eigenvalues(λ_i) against the variables (i). The number of components selected is the value of i that corresponds to an "elbow" in the curve. i.e. a change of slope from steep to shallow

```
> plot(pc,type = "l")#plots the scree plot
```

Example in Regression

In multiple regression,we assume that the explanatory/predictor variables are independent. If the independent variables are correlated then the regression coefficients are unstable.

```
> data=read.csv("multicollinear_pca.csv")
> str(data)
```

```
'data.frame':        20 obs. of  4 variables:
 $ x1: num  19.5 24.7 30.7 29.8 19.1 25.6 31.4 27.9 22.1 25.5 ...
 $ x2: num  43.1 49.8 51.9 54.3 42.2 53.9 58.5 52.1 49.9 53.5 ...
 $ x3: num  29.1 28.2 37 31.1 30.9 23.7 27.6 30.6 23.2 24.8 ...
 $ y : num  11.9 22.8 18.7 20.1 12.9 21.7 27.1 25.4 21.3 19.3 ...
```

The **Response (y)** is a measure for the amount of fat in a human body.

x_1 is triceps skin fold thickness
x_2 is thigh circumference
x_3 is mid arm circumference

Example in Regression

Fitting several linear regressions

```
> fit1=lm(y~x1,data=data)
> fit2=lm(y~x2,data=data)
> fit3=lm(y~x1+x2,data=data)
> fit4=lm(y~x1+x1+x3,data=data)
```

The results can be summarized in the following table

Fit	β_1	β_2	β_3
1	0.8572(0.1288)		
2		0.8565(0.1100)	
3	0.2224(0.3034)	0.6594(0.2912)	
4	4.334 (3.016)	-2.857(2.582)	-2.186(1.595)

We cannot account for the increasing standard errors.

Example in Regression

Taking a look at the correlation of the predictor variables

```
> cor(data[,-4])
```

	x1	x2	x3
x1	1.0000000	0.9238425	0.4577772
x2	0.9238425	1.0000000	0.0846675
x3	0.4577772	0.0846675	1.0000000

```
> pairs.panels(data, gap=0)
```

- All the three variables are correlated to one another leading to inflation of variances of the predictor variables.
- To multicollinearity we can fit a linear regression to the three components extracted from the data. Principal independent components are independent of one another.

Example in Regression

To extract the principle components

```
> pcr=prcomp(data[,-4],center = TRUE,scale. = TRUE)
> pcr$rotation
```

```
         PC1         PC2        PC3
x1 0.6946957 -0.05010563  0.7175565
x2 0.6294279 -0.44050902 -0.6401347
x3 0.3481645  0.89634883 -0.2744818
```

```
> summary(pcr)
```

```
Importance of components:
                          PC1    PC2     PC3
Standard deviation     1.4375 0.9658 0.02696
Proportion of Variance 0.6888 0.3109 0.00024
Cumulative Proportion  0.6888 0.9998 1.00000
```

```
> pairs.panels(pcr$x,gap=0)
```

Example in Regression

To fit a regression line to the components

```
> fit5=lm(y~pcr$x, data=data)
> summary(fit5)
Call:
lm(formula = y ~ pcr$x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1950     0.5545  36.418  < 2e-16 ***
pcr$xPC1      2.9358     0.3958   7.418 1.46e-06 ***
pcr$xPC2     -1.6498     0.5891  -2.801   0.0128 *
pcr$xPC3     27.3834    21.1066   1.297   0.2129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Principal components regression

Definition and Assumption

Principal components regression (PCR) is a regression technique based on principal component analysis (PCA).The basic idea behind PCR is to calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure.
A core assumption of PCR is that the directions in which the predictors show the most variation are the exact directions associated with the response variable.

Advantages

- Dimensionality reduction
- Avoidance of multicollinearity between predictors
- Overfitting mitigation

Principal components regression

Draw backs

- A typical mistake is to consider PCR a feature selection method. PCR is not a feature selection method because each of the calculated principal components is a linear combination of the original variables.
- Using principal components instead of the actual features can make it harder to explain what is affecting what.
- Rigorous process when making predictions on the dependent variable.

To perform a principal component regression in R

```
> library(pls)
> pcr_model <- pcr(Sepal.Length~., data = iris_num, scale = TRUE)
> summary(pcr_model)
Data:         X dimension: 150 3
        Y dimension: 150 1
Fit method: svdpc
Number of components considered: 3
TRAINING: % variance explained
              1 comps  2 comps  3 comps
```

Exercises

Correlation Matrix

Given MacDonnels correartion matrix, from the measurements of seven physical characteristics of 5000 convicted men, perform principal component analysis and interpret the derived components.

Given the US air Pollution data.

- Construct a diagram that shows the SO2 variable plotted against each of the six explanatory variables and in each of the scatter plot show the fitted linear regression. Does this diagram help in deciding on the most appropriate model for determining the variables most predictive of SO2.
- Perform a principle component regression after removing whatever cities you think should be regarded as outliers. Produce scatter plots of SO2 against each of the principle component scores. Interpret your results.