Dimension Reduction, AMSI 2021 Winter School

Tutorial 1 Solutions

Anastasios Panagiotelis

July 13, 2021

R Markdown

World Bank Data

Carry out principal components analysis on the World Bank data (`WorldBankClean.csv`) and answer the following questions. To answer the questions run the command `?prcomp` to see the help documentation of the function used to create principal components.

```
#Following code runs PCA

library(tidyverse)
wb<-read_csv('../../data/WorldBankClean.csv')

wb%>%
  select_if(is.numeric)%>% #Use only numerical variables
  scale()%>% #Standardise (See answer to Q1)
  prcomp()->pcaout # Do PCA
```

1. Do you scale the data before applying principal components? Why or why not?

The data are all measured in different units, for example FX.OWN.TOTL.ZS (Account ownership at a financial institution or with a mobile-money-service provider) is measured in percentage of population aged above 15 years, while FB.ATM.TOTL.P5 (Automated teller machines) is measured in per 100,000 adults while TX.VAL.TECH.CD (High-technology exports) is measured in $US. The variance of each variable will be sensitive to the units of measurement, for example measuring FX.OWN.TOTL.ZS in per 100,000 adults and TX.VAL.TECH.CD in millions of $US will change the relative variances.

Since PCA is a technique that explains variance, it will therefore be sensitive to changes in the units of measurement. To avoid this sensitivity, we standardise all variables, by dividing by the standard deviation so that each variable has a variance of 1.

2. What is the standard deviation of the first principal component?

All standard deviations are contained in the sdev *value of a* prcomp *object.*

```
pcaout$sdev[1]
```

```
## [1] 5.367588
```

Alternatively much information can be seen using the summary function (output not shown).

```
summary(pcaout)
```

The standard deviation of the first PC is 5.37 .

3. How many principal components are needed to explain at least 90% of the total variance?

The output can be seen from the output of the summary function. A piece of code that gives the answer is

```
#Compute proportion of explained variance
prop_exp_var<-cumsum(pcaout$sdev^2)/sum(pcaout$sdev^2)
#Compute number of PCs for cumulative to be at least 90%
ans<-sum(prop_exp_var<0.9)+1
```

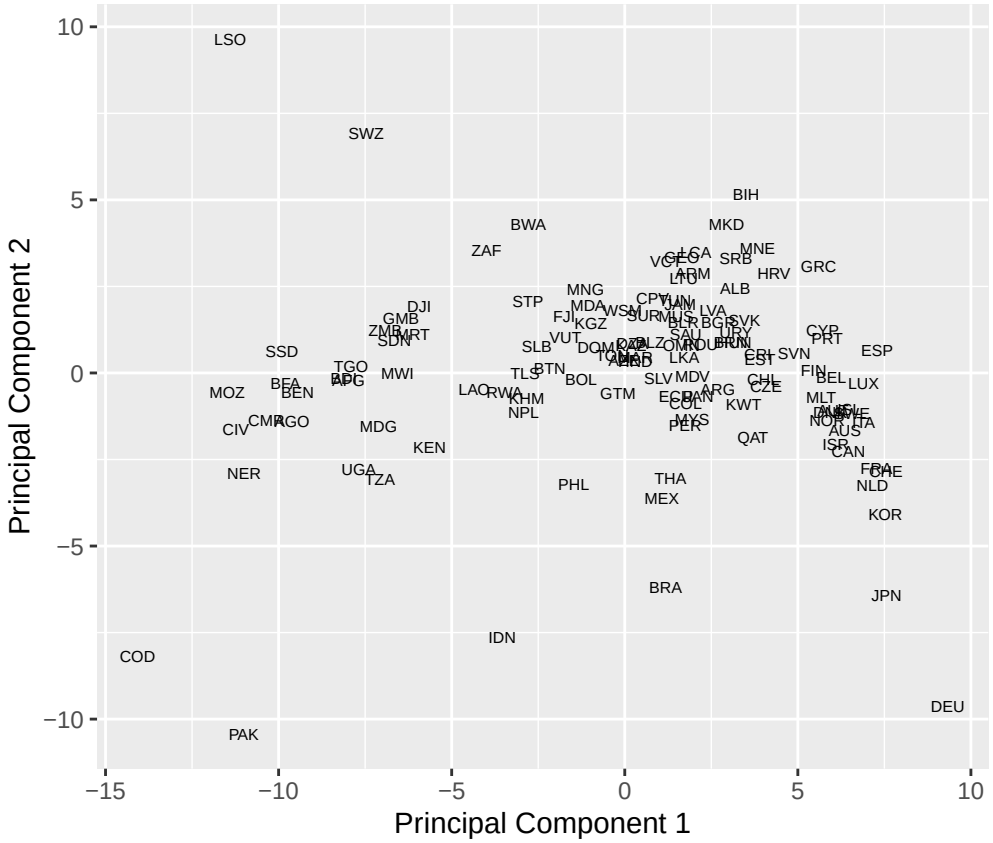
There are 11 principal components needed to explain at least 90% of the overall variance in the data

4. Plot the first two principal components of the data as a scatterplot. Use the country abbreviation rather than points.

All standard deviations are contained in the sdev *value of a* prcomp *object.*

```
#Consolidate everything into a single data frame.
library(broom)
wbPC<-augment(pcaout,wb)

wbPC%>%
  ggplot(aes(x=.fittedPC1, #First PC on x axis
             y=.fittedPC2, #Second PC on y axis
             label=`Country Code`))+ #Rather than points use country code as label.
  geom_text(size=2)+ #geom_text uses text rather than points
  xlab('Principal Component 1')+ #Better labels on axes
  ylab('Principal Component 2')+ #Better labels on axes
  coord_equal() # See discussion below
```



Note that the argument coord_equal *is used for the plot. This ensures that 1 unit on the x axis is the same length as 1 unit on the y axis. Most software that creates a scatterplot will not have this property by default.*

Why is this important? Often when we look at a PCA scatterplot we are tempted to make conclusions about points that are close to one another. What we see as close should not be distorted by a rescaling of the axes.

5. What are the loadings (weights) on the first principal component of infant mortality rate (SP.DYN.IMRT.IN- variable 78) and number of pupils in primary education (SE.PRM.ENRL - variable 123).

```
ans1<-pcaout$rotation["SP.DYN.IMRT.IN","PC1"]
ans2<-pcaout$rotation["SE.PRM.ENRL","PC1"]
```

The weight of infant mortality rate on the first principal component is -0.18 and the weight of pupils in primary education on the first principal component is -0.06 .

6. What are the loadings (weights) on the second principal component of infant mortality rate (SP.DYN.IMRT.IN- variable 78) and number of pupils in primary education (SE.PRM.ENRL - variable 123).

```
ans1<-pcaout$rotation["SP.DYN.IMRT.IN","PC2"]
ans2<-pcaout$rotation["SE.PRM.ENRL","PC2"]
```

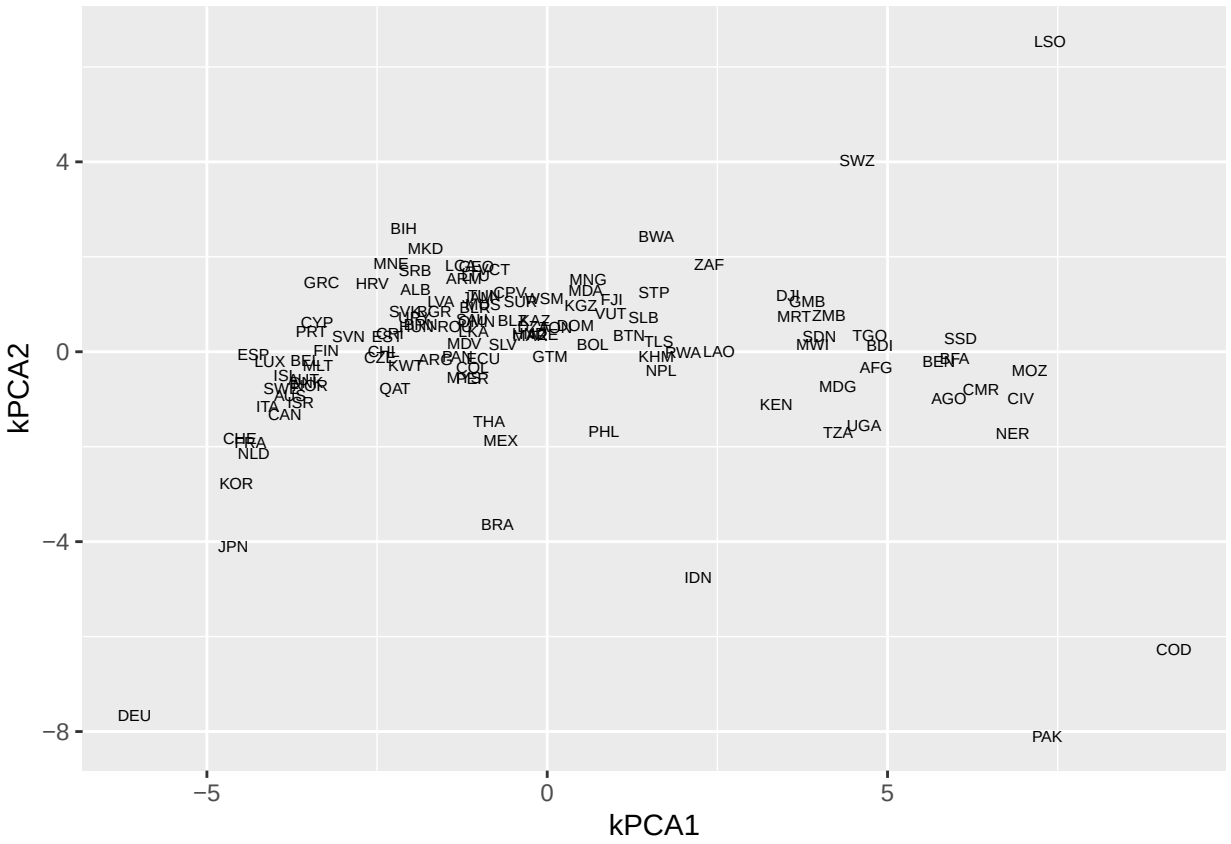
The weight of infant mortality rate on the second principal component is -0.02 and the weight of pupils in primary education on the second principal component is -0.23 .

7. Do these results make sense in light of our interpretation of the first PC measuring level of development and the second PC measuring the size of a country?

Larger values of the infant mortality rate are associated with lower values of the first principal component (since the weight is negative). This is consistent with the first PC measuring the development of a country. The weight of number of pupils in primary education is close to zero (and slightly negative) which at first might seem counter-intuitive since more students in school should be an indicator of development. The problem here is that while infant mortality is a rate, the number of pupils is a raw figure. DR Congo has a larger population that France as well as a much younger age profile therefore has more students in primary education, despite not being as economically developed as France. For this reason, the weight on pupils in primary education is high (in absolute terms) for the second PC. A larger number of students is associated with a smaller value of PC2 while the infant mortality rate has very little contribution to PC2.

8. Run kernel PCA using the `dimRed` package. Use a polynomial kernel with a degree of 3, a scale of 0.001 and an offset of 1. Use `?kPCA-class` and `?kpca` to consult the help documentation of the function in the `dimRed` package and the original function from the `kernlab` package that it wraps around.

```
library(dimRed)
wb%>%
  mutate_if(is.numeric,scale)%>% #Scale Data
  as.dimRedData(`Country Name` + `Country Code`~.,data=.)->wbdr #Convert to S4 class
kpcaout <- embed(.data = wbdr,
                 .method="kPCA",
                 kernel='polydot',
                 kpar=list(degree=3,scale=0.001,offset=1))
df<-tibble(cbind(kpcaout@data@meta,kpcaout@data@data)) # Convert back to a dataframe
ggplot(df,aes(x=kPCA1,y=kPCA2,label=`Country Code`))+geom_text(size=2) #Plot
```



Irish Smart Meter data

1. Using the household with ID=4669 for the Irish smart meter data, find the location and scale parameter from fitting electricity demand for each time of week with a lognormal distribution

HINT 1: You only need to take the log of the data then find the mean and standard deviation, but you will need to remove a small number of observations for which demand is 0.

HINT 2: The `filter` function can be used to remove zeros, the `mutate` to take the log and the `group_by` and `summarise` functions can be used to find means and standard deviations by the time of week (`tow`) variable. These are all functions from the `tidyverse`.

```
sm<-read_csv('../../data/SmartMeter.csv')
sm%>%filter(id==4669)%>% #Only consider ID 4669
  filter(demand>0)%>% #Only include demand if positive
  mutate(logd=log(demand))%>% #Take log
  group_by(tow)%>% #Group be time of week
  summarise(mu=mean(logd,na.rm=TRUE), #Mean
            sigma2=var(logd,na.rm=TRUE))->pars #Variance
```

2. Compute the Jensen Shannon distance (JSD) between all pairs. Remember that the JSD is the square root of the average of the Kullback Leibler divergence from P to Q and the Kullback Leibler divergence from Q to P. The Kullback Leibler divergence from P to Q for a lognormal distribution is given by.

$$KL(P||Q) = \log \sigma_Q^2 - \log(\sigma_P^2) + \frac{\sigma_P^2 + (\mu_P - \mu_Q)^2}{2\sigma_Q^2} - 0.5$$

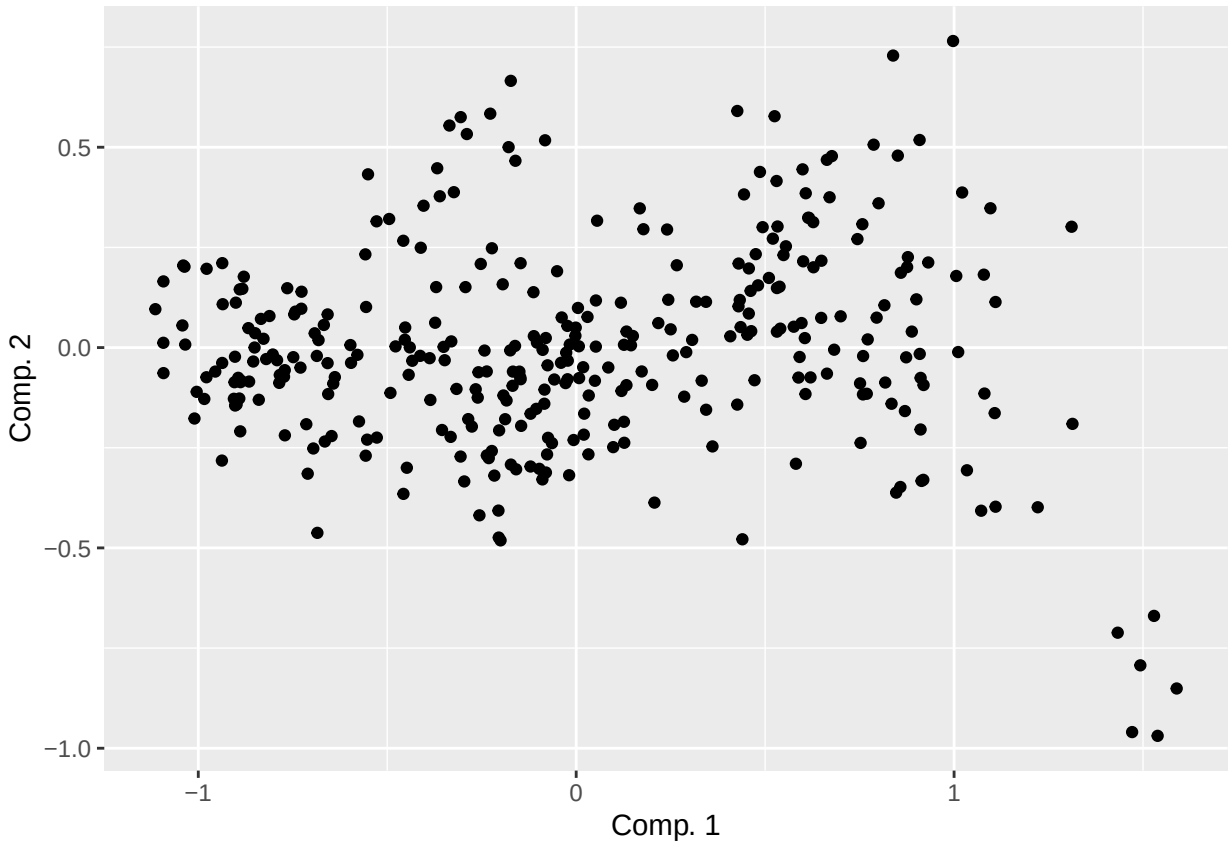
```r
n<-nrow(pars) #Number of observations
klds<-matrix(0,n,n) #Initialise dist matrix
for (i in 1:n){ #Double loop
  for (j in 1:n){
    if(i!=j){
      mui<-pars$mu[i] #Pick out mean of first distribution
      muj<-pars$mu[j] #Pick out mean of second distribution
      sigma2i<-pars$sigma2[i] #Variance of first distribution
      sigma2j<-pars$sigma2[j] #Variance of second distribution
      klds[i,j]<-log(sigma2j)-log(sigma2i)+
        ((sigma2i+(mui-muj)^2)/(2*sigma2j))-0.5
    }
  }
}
#The lower triangle will contain KL divergences from P to Q
#The upper triangle will contain KL divergences from Q to P
#Average can be found by adding matrix to its transpose and dividing by 2
jsds<-sqrt((klds+t(klds))/2)
```

3. Construct a plot similar to the one on slide 20 of Lecture 3. A simple scatterplot is sufficient, you do not need to replicate the coloring of the plot in the slides.

```r
jsds%>%
  as.dist%>%
  cmdscale()%>%
  cbind(pars)->df

colnames(df)[1:2]<-c("Comp. 1", "Comp. 2")


ggplot(df,aes(x=`Comp. 1`,y=`Comp. 2`))+geom_point()
```



4. There are six points with a very high value on the first coordinate and a low value on the second coordinate. What do the points in this underlying group have in common. HINT: To investigate use the `arrange` function

Using the arrange function we can see times of week corresponding to the highest value of Component 1.

```
arrange(df,desc(`Comp. 1`))%>%head(10)
```

```
##     Comp. 1    Comp. 2 tow          mu    sigma2
## 1  1.589058 -0.8508897  39 -0.03897928 0.5846202
## 2  1.538680 -0.9690741  87 -0.02382502 0.7097057
## 3  1.529073 -0.6696822 279 -0.13798952 0.5096594
## 4  1.492936 -0.7930853 231 -0.10925556 0.6112970
## 5  1.471169 -0.9596916 135 -0.06626386 0.7653351
## 6  1.433075 -0.7116668 183 -0.17116222 0.5941351
## 7  1.313414 -0.1904303  38 -0.44537855 0.3504907
## 8  1.311072  0.3015813  65 -0.65491728 0.1999306
## 9  1.221449 -0.3985672 134 -0.40935983 0.4963219
## 10 1.109828 -0.3972060 278 -0.48015193 0.5526908
```

The six outlying times of week are also the six times of week with the highest mean.

```
arrange(df,desc(mu))%>%head(10)
```

```
##     Comp. 1    Comp. 2 tow          mu    sigma2
## 1  1.538680 -0.9690741  87 -0.02382502 0.7097057
## 2  1.589058 -0.8508897  39 -0.03897928 0.5846202
## 3  1.471169 -0.9596916 135 -0.06626386 0.7653351
## 4  1.492936 -0.7930853 231 -0.10925556 0.6112970
```

```
## 5  1.529073 -0.6696822 279 -0.13798952 0.5096594
## 6  1.433075 -0.7116668 183 -0.17116222 0.5941351
## 7  1.221449 -0.3985672 134 -0.40935983 0.4963219
## 8  1.313414 -0.1904303  38 -0.44537855 0.3504907
## 9  1.109828 -0.3972060 278 -0.48015193 0.5526908
## 10 1.071735 -0.4072333  86 -0.50285984 0.5828020
```

The six outlying times of week occur at the same time of day.

```
#The %% is a modules operator and 48 is the number of half hour blocks in a day.
c(39,87,279,231,135,183)%%48
```

```
## [1] 39 39 39 39 39 39
```

From this investigation we can conclude that this particular household has higher than average usage at 7:30 PM.

Bonus Questions

1. The second principal component is the linear combination of weights with maximal variance while also being uncorrelated with the first principal component. Prove that the weights of this combination are given by the eigenvector corresponding to the second largest eigenvalue.
2. For the polynomial kernel

$$K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + 1)^2$$

in the case where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, show that a feature map with this kernel is given by

$$\Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \\ 1 \end{pmatrix}$$