HDDA Tutorial: Dimension Reduction : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 10

1. Load the socioeconomic data on U.S. States (used in the lecture on principal components). Isolate the 5 numeric variables and scale this data

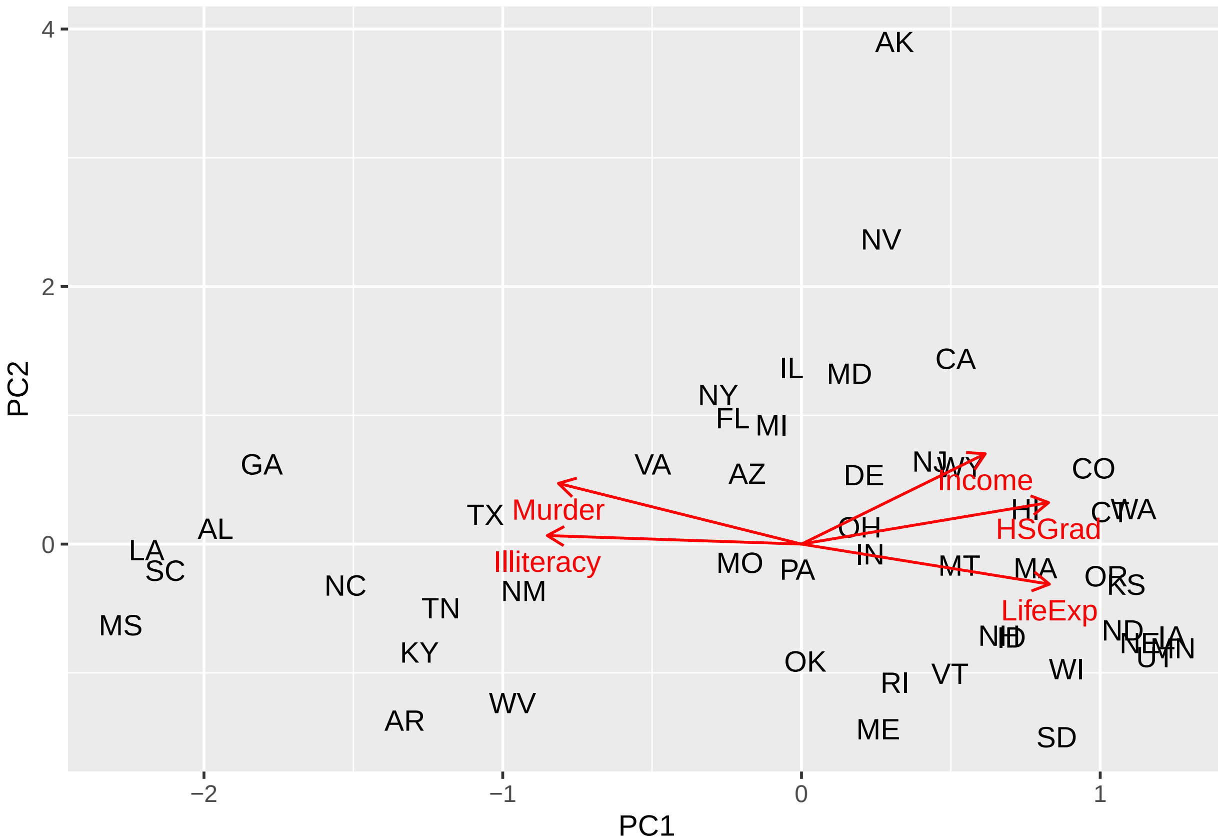
```
library(tidyverse)
States<-readRDS('StateSE.rds')
States%>%
  select_if(is.numeric)%>%
  scale->States_Scaled
```

2. Carry out the singular value decomposition on the standardised data.

```
States_SVD<-svd(States_Scaled)
```

3. Construct a correlations biplot using `ggplot`. Before doing so look at the help function for `pc.biplot` for additional instructions on how this is constructed

```
#Set scale and find sample size
scale=1
n<-nrow(States)

#Singular values can be put into a diagonal matrix using diag
#Use %*% for matrix multiplication

#Note that the observations are multiplied by root n
PC_obs<-States_SVD$u%*%diag((States_SVD$d)^(1-scale))*sqrt(n)

#Note that the variables are divided by root n
PC_var<-States_SVD$v%*%diag((States_SVD$d)^scale)/sqrt(n)

#Create dataframe for observations with PCs
df_Obs<-tibble(Label=pull(States,StateAbb), #Extract State Abbreviation
               PC1=PC_obs[,1], #Extract first PC
               PC2=PC_obs[,2]) #Extract second PC

#Create dataframe for variables
df_Vars<-tibble(Label=colnames(States_Scaled), #Extract State Abbreviation
               PC1=PC_var[,1], #Extract first loading vector
               PC2=PC_var[,2]) #Extract second loading vector

ggplot(data = df_Vars,aes(x=PC1,y=PC2,label=Label))+
  geom_text(data=df_Obs)+ #Observations
  geom_text(color='red',nudge_y = -0.2)+ #Variables (offset using nudge_y)
  geom_segment(xend=0,yend=0,color='red', #Add arrow
               arrow = arrow(ends="first",
                             length=unit(0.1, "inches"))) #Make tip smaller
```



4. Using the spectral theorem, prove that Principal Components are uncorrelated by construction.

Consider the $n \times p$ *matrix* $\mathbf{C} = \mathbf{YV}$ *where* $\mathbf{Y}$ *is the data matrix and the columns of* $\mathbf{V}$ *are the eigenvectors of the covariance matrix* $\mathbf{S}$*. The matrix* $\mathbf{C}$ *will be an* $n \times p$ *itself. The* i^{th} *row and* j^{th} *column of* $\mathbf{C}$ *is obtained by multiplying the* i^{th} *row of* $\mathbf{Y}$ *by the* j^{th} *column of* $\mathbf{V}$*. The* i^{th} *row of* $\mathbf{Y}$ *contains the values of all variables for observation* i *while the* j^{th} *column of* $\mathbf{V}$ *contains the weights for principal components* j*. This implies that The* i^{th} *row and* j^{th} *column of* $\mathbf{C}$ *is the value of principal component* j *for variable* i*.*

The matrix $\mathbf{C}$ *is essentially a data matrix but for the principal components. As such the variance covariance matrix for the principal components is found by taking* $\frac{1}{n-1}\mathbf{C}'\mathbf{C}$*. Some matrix algebra shows*

$$\frac{1}{n-1}\mathbf{C}'\mathbf{C} = \frac{1}{n-1}(\mathbf{YV})'\mathbf{YV} \tag{1}$$

$$= \frac{1}{n-1}\mathbf{V}'\mathbf{Y}'\mathbf{YV} \tag{2}$$

$$= \mathbf{V}'\frac{1}{n-1}(\mathbf{Y}'\mathbf{Y})\mathbf{V} \tag{3}$$

$$= \mathbf{V}'\mathbf{SV} \tag{4}$$

Using the eigenvalue decomposition, $\mathbf{S} = \mathbf{V\Lambda V}'$ *and substituting*

$$\frac{1}{n-1}\mathbf{C}'\mathbf{C} = \mathbf{V}'\mathbf{SV} \tag{5}$$

$$= \mathbf{V}'\mathbf{V\Lambda V}'\mathbf{V} \tag{6}$$

$$= \mathbf{\Lambda} \tag{7}$$

The above holds since $\mathbf{V}'\mathbf{V} = \mathbf{I}$. *Since* $\mathbf{\Lambda}$ *is a diagonal matrix, the covariances are all 0 and the principal components are uncorrelated.*