¡style type=”text/css”¿
body, td font-size: 12px; code.r font-size: 12px; pre font-size: 12px ¡/style¿

Homework 2 - Solutions to Homework 1

Syed Rahman

Problem 1: The data set USTATES-data.txt (available in the folder Data Sets) contains the unemployment rates of the 50 US States, plus those of the District of Columbia and Puerto Rico for the period Jan 1976- March 2007. Use PCA analysis to summarize the basic patterns in the data and carefully comment on the results. Specifically, discuss in your report possible transformations of the variables, exclusion of outliers, choice of number of principal components, fit of the solution, interpretation of the principal components.

Solution 1: In this problem, we note that $p >> n$. One approach is to transpose the data, however, this is ultimately incorrect as it would be harder to justify the $X_i \overset{iid}{\sim} \mathcal{N}_p(\mu, \Sigma)$ assumption of PCA. Instead we use an SVD based function to perform PCA. Doing the analysis with and without PR in addition to the following in Figures 1 and 2 indicates that it is an outlier and that it should be left out of the final analysis.

```
data1 = read.table("USTATES-data-1.txt", header=TRUE)
data0 = as.matrix(data1)
pc1 = prcomp(data0, center = TRUE)
summary(pc1)$importance[,1:4]
```

```
##                             PC1      PC2      PC3      PC4
## Standard deviation     35.49366 12.66850 8.843756 5.704322
## Proportion of Variance  0.76815  0.09786 0.047690 0.019840
## Cumulative Proportion   0.76815  0.86601 0.913700 0.933540
```

```
matplot(t(data0),type = "l", xlab = "Years", ylab = "Unemployment Rates")
```

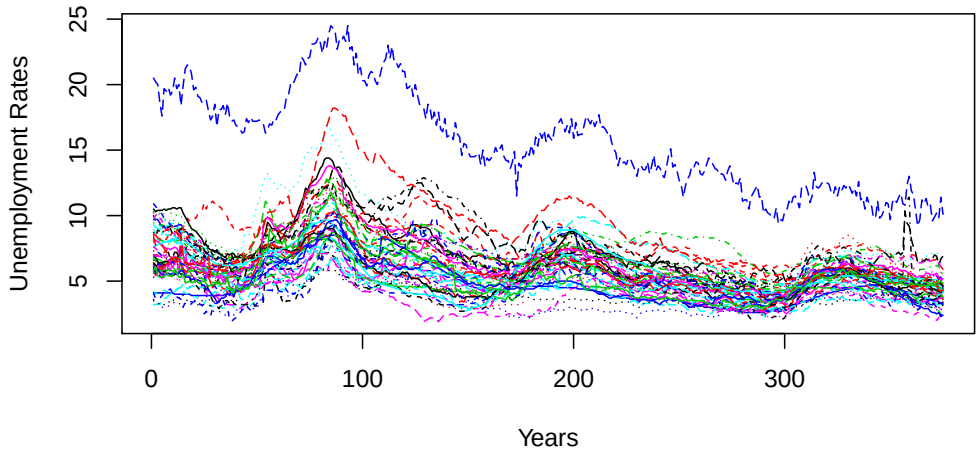

Figure 1: Unemployment Rates by state

```
plot(pc1$x[,1],pc1$x[,2] , xlab="PC1",ylab="PC2",type="n",lwd=2)
text(pc1$x[,1],pc1$x[,2],labels=abbreviate(row.names(data1)),cex=0.7,lwd=2)
```

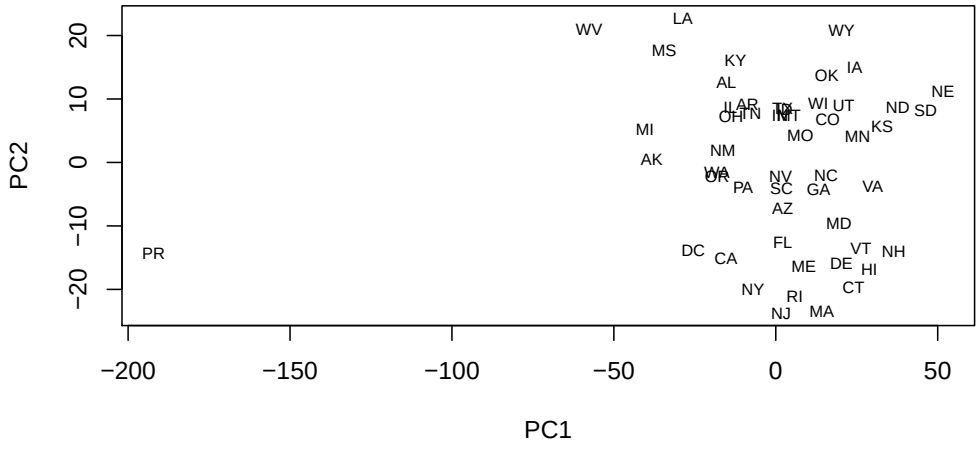

Figure 2: Bi-plot including PR

```
data = data0[-40,]
pc2 = prcomp(data, center = TRUE)
summary(pc2)$importance[,1:4]
```

```
##                             PC1      PC2      PC3      PC4
## Standard deviation     23.32671 12.33142 8.861459 5.608603
## Proportion of Variance  0.59591  0.16653 0.086000 0.034450
## Cumulative Proportion   0.59591  0.76244 0.848440 0.882880
```

The screeplot in Figure 3 indicates that it is sufficient to hold on to 2 PCs.

```
screeplot(pc2)
```

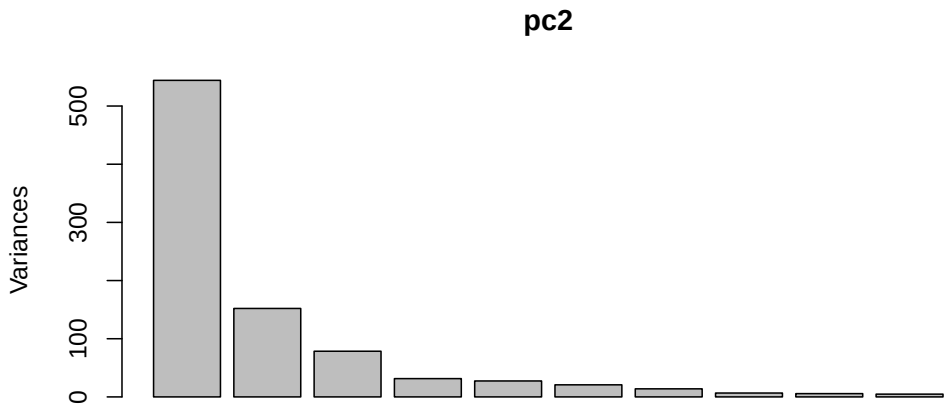

Figure 3: Screeplot excluding PR

Finally, taking a look at the factor loadings in addition to look at the bi-plot in Figure 4 suggest that PC1 is essentially an average of all the months, while PC2 is positive over certain time time periods while negative over others. In additon, the variables are all very closely related to each other due to the small angles between them on the bi-plot.

```
biplot(pc2)
```

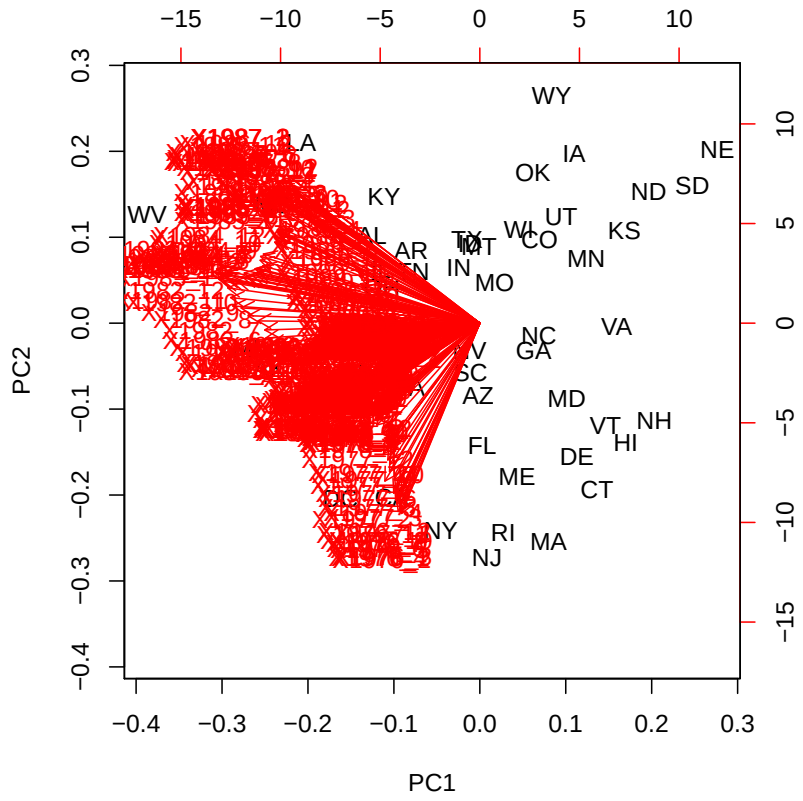

Figure 4: Bi-plot excluding PR

To check the fit of the data we do a simple bootstrap and estimate the proportion of data explained by 2 variables. The center of the histogram is near 0.76, which indicates that the PCA is a good fit for the data.

```
set.seed(12345)
niter = 1000
propexp = matrix(0,1000,1);
for(i in 1:1000){
   pc = prcomp(data[sample(nrow(data),size=51,replace=TRUE),])
   propexp[i,1] = sum(pc$sdev[1:2]^2)/sum(pc$sdev^2)
}
hist(propexp, xlab = "Proportion explained by 2 PCs", ylab = "Frequency")
```

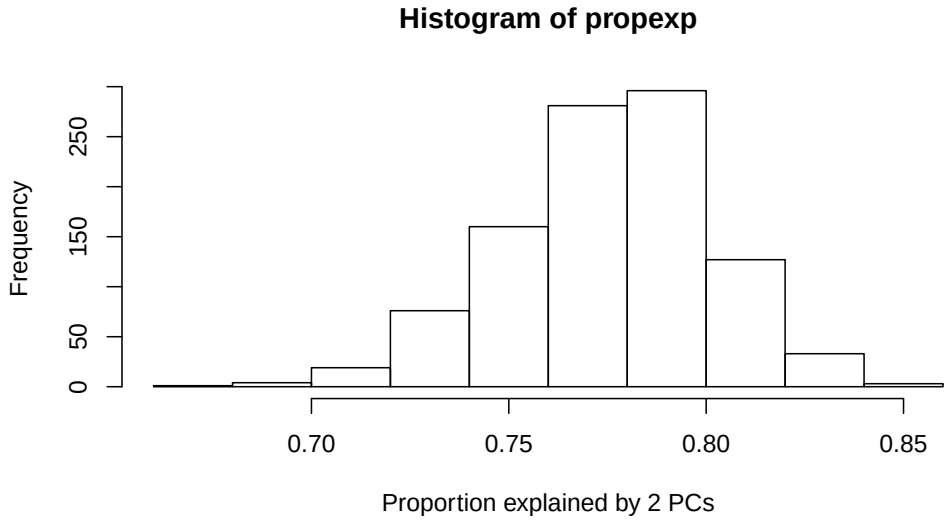

Figure 5: Histogram of Bootstrap

Problem 2: This data set contains information about the chemical composition of wines. Perform PCA and comment on the results.

Solution 2: For this dataset $n >> p$. However, the units for the variables are different, which calls for doing PCA on the correlation matrix. Based on Figure 6, we apply the log transform on the data to help stabilize it. However, a comparison with the unlog transformed data indicates that this only marginally improves the proportion explained. The screeplot in Figure 7 indicates that we should hold onto at least 3 PCs as well. Four could be appropriate too, if we wanted to explain at least 70% of the variance. By studying the factor loadings, we can see that PC1 is largely driven by Total.phenols, Falavanoids and OD280.OD315 while PC2 is Alcohol.content, Color.Intensity and Proline, PC3 is As and Alcalinity.in.ash while PC4 divies those with high Malic.acid and Proantocyanins against those with high Hue and Nonflavanoid.phenols.

```
data2 = read.csv("wine.csv", header=TRUE)
data20 = (as.matrix(data2))
```

```
matplot(t(data20),type = "l", xlab = "Variables", ylab = "Wines")
```

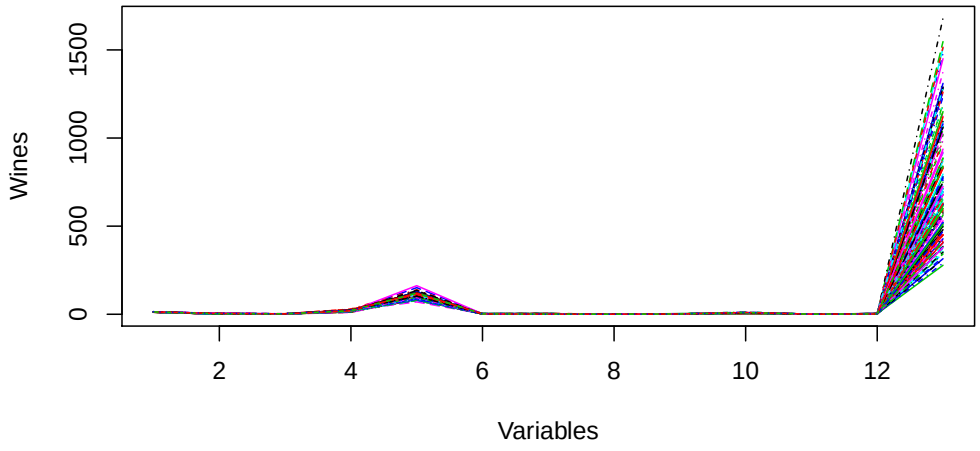

Figure 6: Screeplot

```
data20 = log(as.matrix(data2))
pc21 = prcomp(data20, center = TRUE, scale = TRUE)
summary(pc21)$importance[,1:4]

##                             PC1     PC2      PC3      PC4
## Standard deviation     2.140819 1.65417 1.218038 0.938557
## Proportion of Variance 0.352550 0.21048 0.114120 0.067760
## Cumulative Proportion  0.352550 0.56303 0.677150 0.744910

summary(pc21)$rotation[,1:4]

##                                PC1         PC2         PC3         PC4
## Alcohol.content      -0.10842850  0.47918760 -0.17865648  0.03436658
## Malic.acid            0.24914242  0.21952916  0.19286581 -0.46548550
## As                    0.01891262  0.29984440  0.60322701  0.30925797
## Alcalinity.in.ash     0.24000989 -0.06280937  0.60205834 -0.01609889
```

```
## Magnesium            -0.12779786  0.33176191  0.09148971  0.16530774
## Total.phenols        -0.39545612  0.06418244  0.15084280 -0.08133091
## Falavanoids          -0.42757871 -0.03987141  0.16118936 -0.09455543
## Nonflavanoid.phenols  0.29012478 -0.02028423  0.16120235  0.41176889
## Proantocyanins       -0.31688859  0.05267407  0.25031696 -0.43090965
## Color.Intensity       0.09029551  0.51872905 -0.11869307 -0.06840962
## Hue                  -0.33284741 -0.22538876  0.01941531  0.46783535
## OD280.OD315          -0.39359075 -0.13786792  0.16328603 -0.07315717
## Proline              -0.23285697  0.41289943 -0.13285159  0.24335554
```

The screeplot in Figure 3 indicates that it is sufficient to hold on to 2 PCs.

```
screeplot(pc21)
```

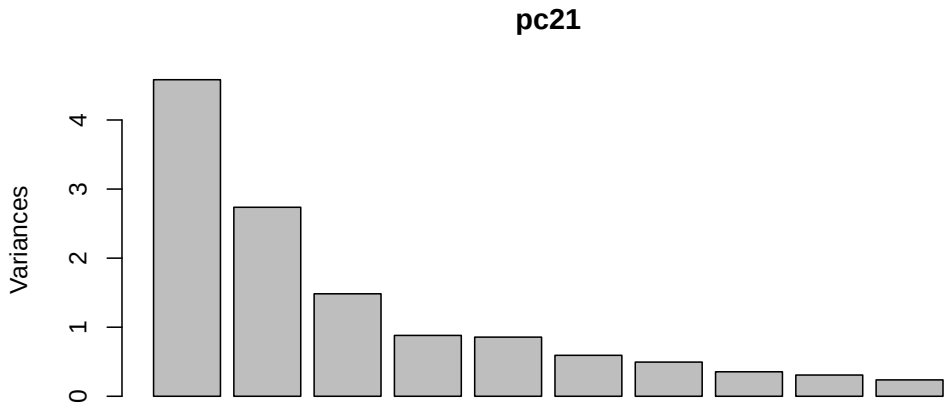

Figure 7: Screeplot

```
biplot(pc21)
```

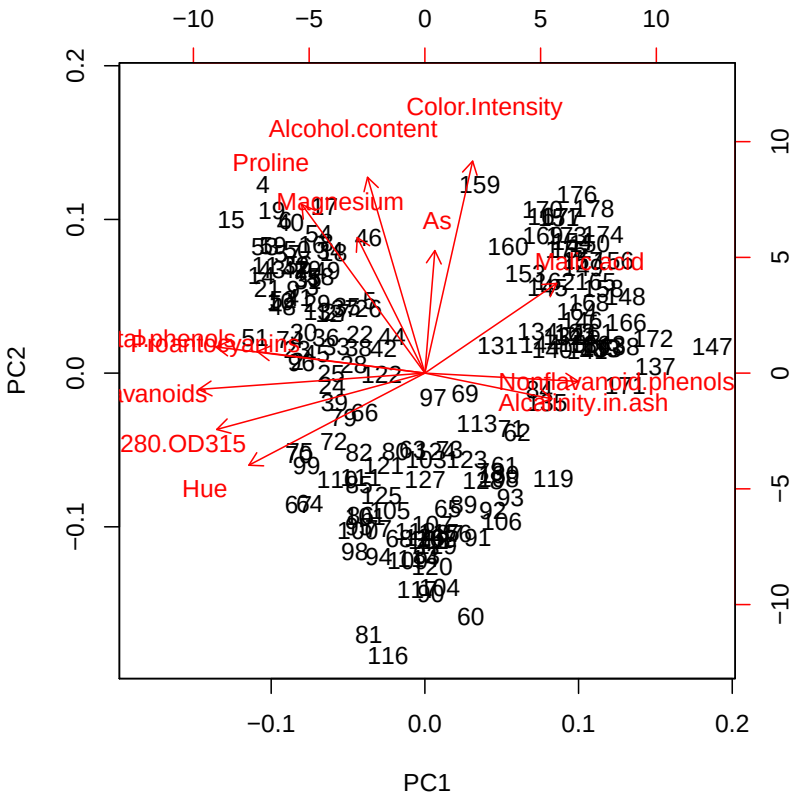

Figure 8: Bi-plot