Democracy Score Analysis

Surui Sun

March 14, 2016

Abstract

In the project, I use score data of 7 different aspects: Political System, Economics, Environment, Gender Enquality, Health, Knowledge, Gender Comprehensive to describe the degree of democracy for most countries in the world. Implementing dimension reduction models such as PCA, k-means clustering, I find that the variable Environment has very different properties with other variables, and three principal component are enough to describe the data without lossing much information. Then I take the variable 'type', which is an indicator of whether a country is democratic and has value either 0 or 1, as output variable, and apply ridge regreesion and Principal Component Regression to give prediction for type of countires based on 7 variables mentioned above and get very small prediction errors.

keyword: PCA, k-means Clustering, ridge regression, PCR

Introduction

Democracy is an interesting topic around the world. Politicians, social scientists and statisticians have used lots of data to give ranking for democracy of countries. My project is based on *http://democracyranking.org/*, which is a project carried by the Democracy Ranking Association. And in my project, I use some data transformation technique, and then implement PCA and clustering methods to describe the data, to give people more knowledge about input variables, and finally use regression methods to give accurate prediction for countries's status of democracy, which can be used further for prediction in the future.

Analysis and Results

Basic Analysis

The original score data comes from **http://democracyranking.org/**, in their analysis, each country is given a rank of democracy for the year of 2012. But in this project, I introduce a new variable: type. The value of type for an country equals to 1 if its rank is less or equal to 75(This number is arbitrary, here I choose 75 because we have 112 countries in total and 75 is about 2/3 of all), otherwise the value equals to 0.

When we apply PCA and clustering method, we are only interested in 7 variables which are described above. But when we apply ridge regression and PCR, we use output variable 'type' and those 7 features together.

Table **??** gives first ten rows of data used in the project. And here is some brief introduction of the meaning of different variables:

Annual ranking of all country-based democracies in the World Quality of Democracy is measured by freedom and other characteristics of the political system, and performance of the non-political dimensions, where the non-political dimensions are: gender, economy, knowledge, health, and the environment.

For example, PS means political System Score, and it is a combination of score of seven different aspects. Those are:

1.Political rights (aggregated scores): *Freedom House*
2.Civil liberties (aggregated scores): *Freedom House*
3.Global Gender Gap Report
4.Press Freedom: *Freedom House*
5.Corruption Perceptions Index (CPI): *Transparency International*
6.Change(s) of the head of government (last 13 years, peaceful)
7.Change(s) of the head of government (last 13 years, peaceful)

EC means Economic System Score, and it is a combination of score of six different aspects. Those are:

1.GDP per capita, PPP (constant 2005 international $)
2.GDP per capita, PPP (current international $)
3.GDP per capita, PPP (current international $)
4.Inflation, consumer prices (annual %)
5.Unemployment, total (% of total labor force)
6.Unemployment, youth total (% of total labor force ages 15-24)

For more information about the data set, please refer to http://democracyranking.org/.

In Figure **??**, if the type of a country equals to one, it has red color. Otherwise the type of a country is zero, and has yellow color. Type zero means the country is not very democratic at the moment, while Type one means the country is thought to be democratic compared to type zero countries.

	PS	EC	EN	GE	H	K	GC	rank	type
Albania	54.10	35.80	80.70	66.60	67.90	33.20	60.30	60	1.00
Argentina	66.70	75.80	80.30	77.50	70.10	47.90	74.60	34	1.00
Armenia	39.90	28.30	68.00	65.70	63.50	34.40	52.90	92	0.00
Australia	86.60	67.70	46.90	87.20	83.00	71.50	85.70	12	1.00
Austria	84.30	66.30	68.80	83.30	85.50	68.30	83.10	11	1.00
Bahrain	20.30	62.90	43.30	58.00	66.20	54.70	39.30	107	0.00
Bangladesh	50.20	37.70	73.60	61.40	50.50	19.00	56.90	80	0.00
Belgium	90.30	59.80	58.10	81.80	81.90	65.00	85.60	10	1.00
Benin	57.10	41.70	63.90	52.90	28.20	16.80	53.50	84	0.00
Bolivia	61.00	42.10	66.60	68.60	46.50	25.10	66.30	61	1.00

Table 1: First Ten Rows of Data Used in the Model

PCA

First of all, we acquired the correlation matrix for 7 variables. And give a levelplot:

Abbreviated words are used here: PS(Political System), Economics(EC), Environment(EN), Gender Enquality(GE), Health(H), Knowledge(K), Gender Comprehensive(GC).

From the plot, EN(environmrnt) has negative correlation with other variables, while other variables have high correlation with each other.(except environment). Thus we could try PCA.

Implementing PCA, the variance in the principal component is shown in figure **??**:

the 1st Principal Component has relatively much larger variance than the other PC's.

From table **??** and table**??**, first three component together account for 90% of variance of the original variables. Scores on these three components

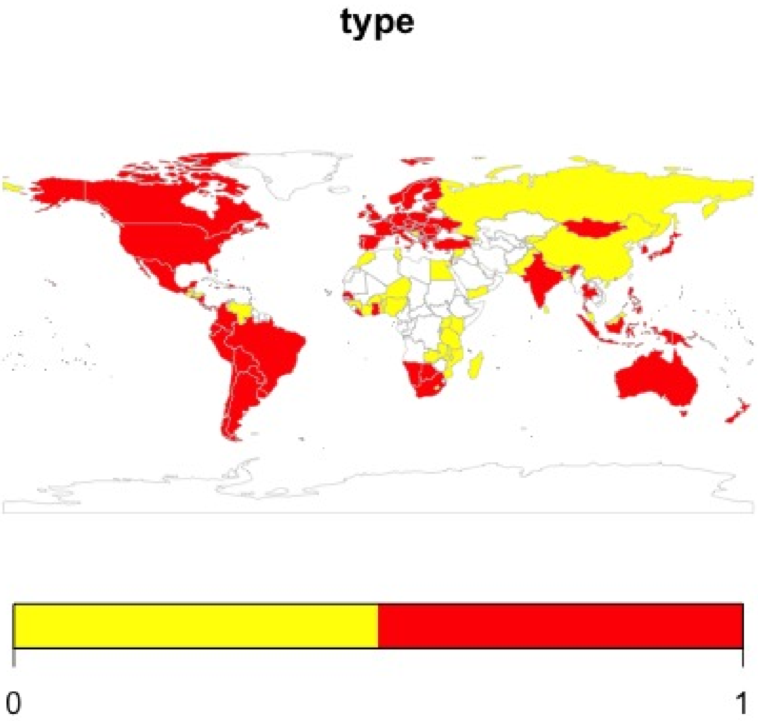

Figure 1: Country Type in the original Data

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	2.178	0.956	0.825	0.649	0.412	0.255	0.068
Proportion of Variance	0.678	0.131	0.097	0.060	0.024	0.009	0.0007
Cumulative Proportion	0.678	0.808	0.906	0.966	0.990	0.999	1.000

Table 2: Importance of Components

might be used to graph the data with little loss of information.
The first component is the overall influence, the quality of life, when a high score indicates a low quality of life. The second component is mostly the influence of environment, a high score indicates an awful environment. The third component is mainly based on political system minus economic and environment score,It means how well the political system is beyond the eco-

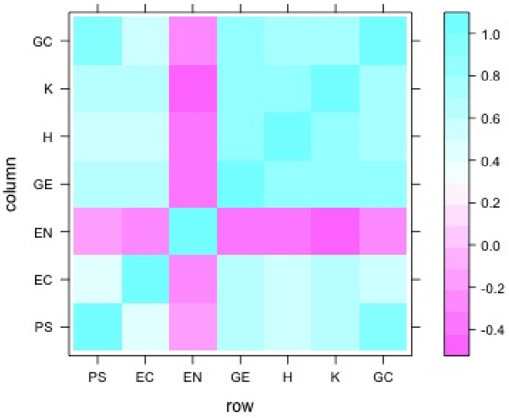

Figure 2: correlation among variables

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
PS	-0.361	-0.443	0.464	0.328	-0.174	0.280	0.492
EC	-0.331	0.120	-0.653	0.642	0	0.185	0
EN	0.206	-0.833	-0.469	-0.185	0	0	0
GE	-0.434	0	0	-0.227	0.605	-0.462	0.415
H	-0.406	0	-0.198	-0.609	0	0.645	0
K	-0.429	0	-0.104	-0.135	-0.731	-0.495	0
GC	-0.424	-0.289	0.281	0	0.249	0	-0.764

Table 3: Loadings

nomics and the environment. Besides, these three variables can be seen as macro variables, while the other variables can be seen as micro variables. So it is the measure of influence of macro situation for the country.

Figure **??** plots the scores on three major components. However, it is hard to tell any relationship between these three principle component.

Figure **??** is the bivariate boxplot for top three principle components. And the points that lie outside of the circle are throught to be candidates for

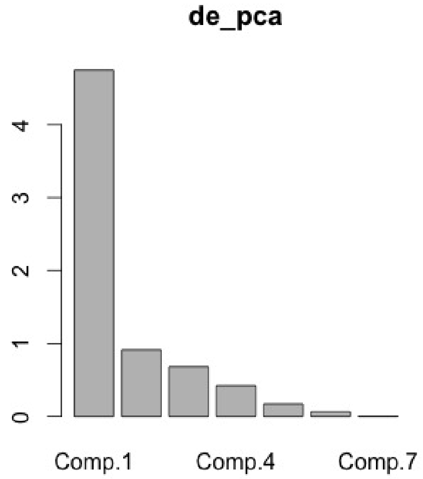

Figure 3: correlation among variables

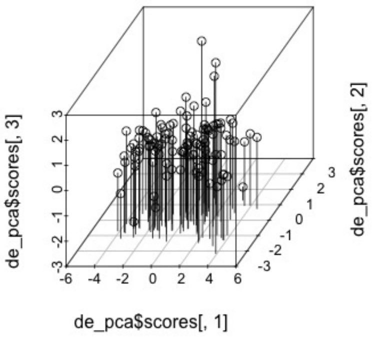

Figure 4: Scores on three major components

outliers. However, no evidence show that there are any strict outliers.

From the biplot in Figure **??**, Environment seems to be uncorrelated with

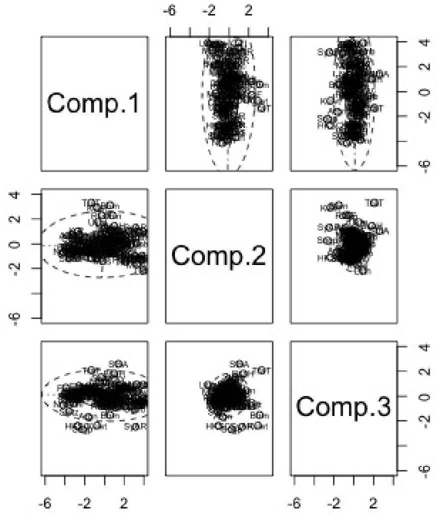

Figure 5: bivariate boxplot for top three principle components

other variables, and those other variables are similar among each other, which is confirmed in the correlation levelplot. And principal component 1 is a great way to distinguish between two types of countries.

Factor Analysis and Correspondence Analysis

Number of Factors	P-Value
1	1.15e-81
2	5.81e-17
3	7.6e-05

Table 4: Hypothesis Testing: Number of Factors

From table **??**, the total number of variables is 7, and we will reject null hypothesis that factor analysis is significant for number of factors to be 1, 2, 3, if we enhance the number of factors, then the number of factors will be too many for 7 variables. Thus we are unable to do factor analysis in this

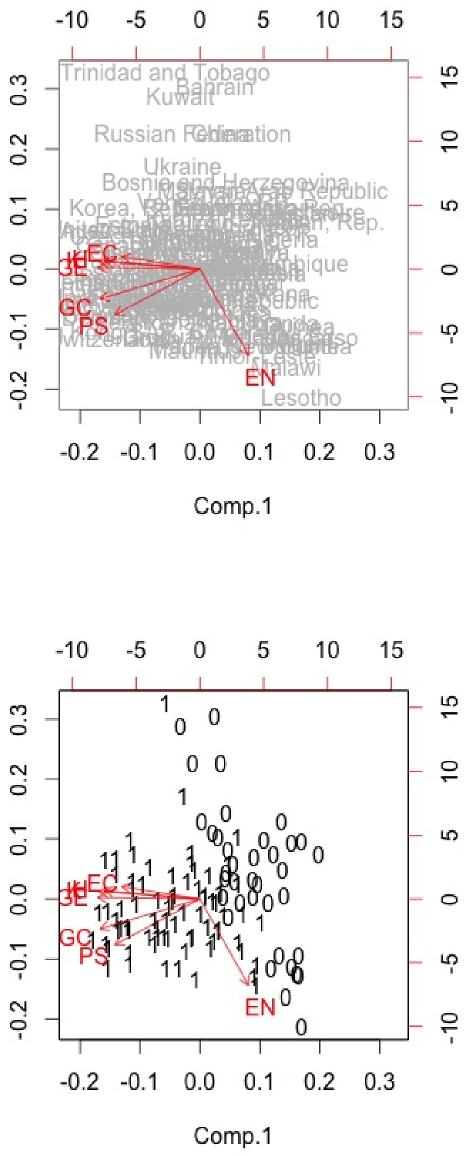

Figure 6: Biplot for Countries and Types

case.

As we do not have many category variables, we are unable to do Correspondence Analysis either.

Clustering

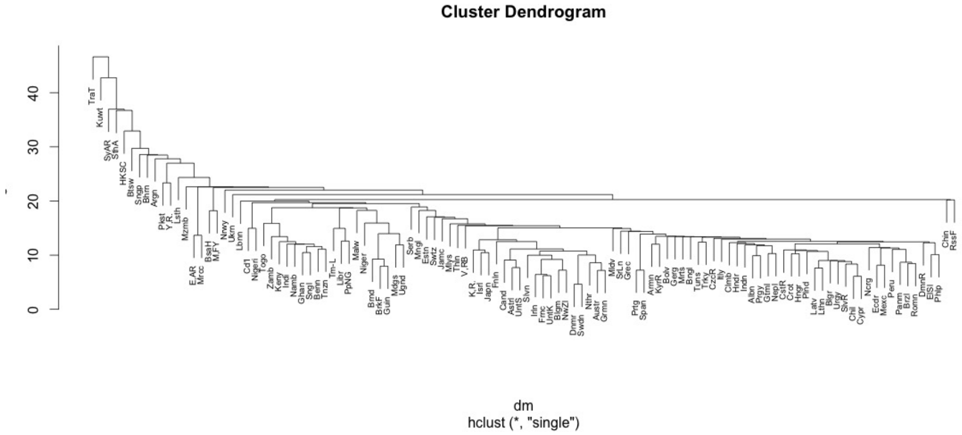

Figure 7: Clustering: single method

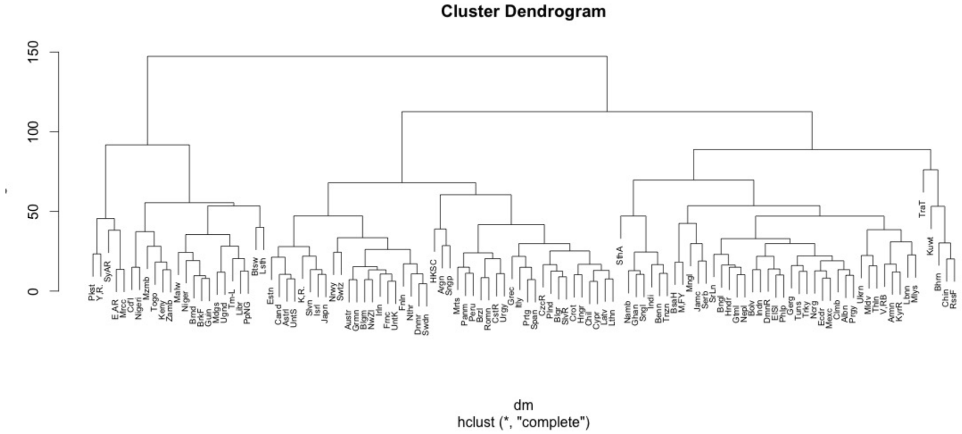

Figure 8: Clustering: complete method

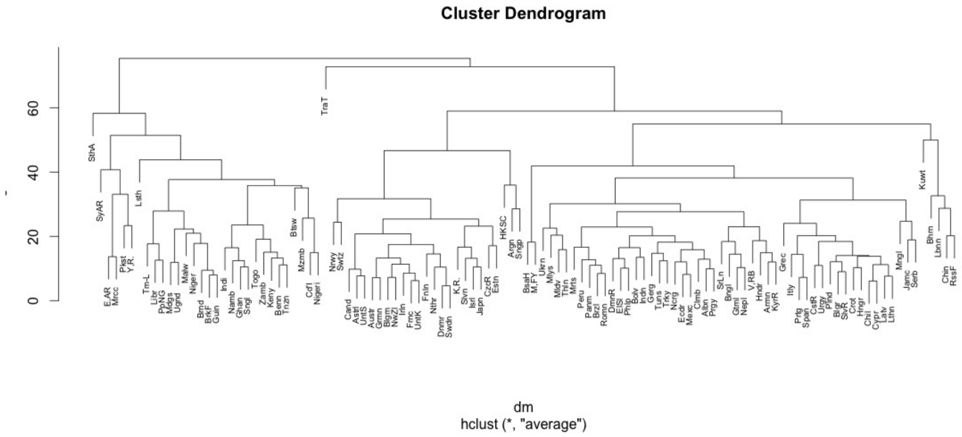

Figure 9: Clustering: average method

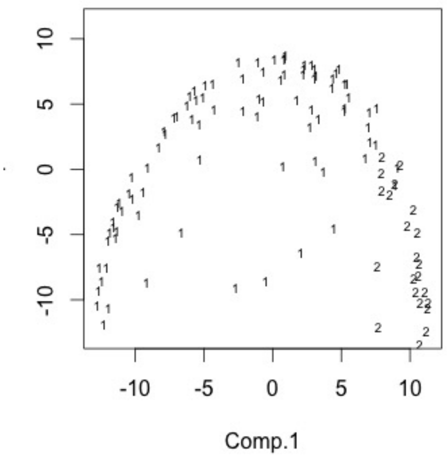

Figure 10: Complete Method: measuring data

The clustering is based on method of measuring the distance between groups, here we use single, complete, and average methods for clustering. From Figure **??**, single linkage solutions often contain long straggly clusters

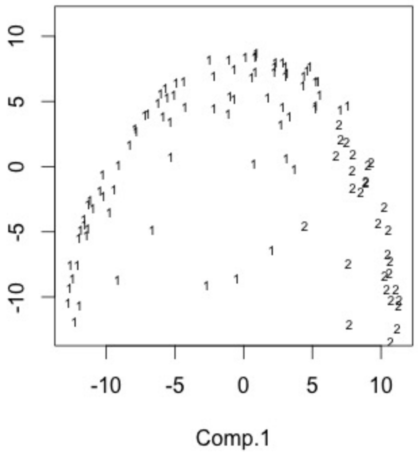

Figure 11: Average Method: measuring data

that do not give a useful description of the data. While from Figure **??** and Figure **??**, The two-group solutions from complete linkage and average linkage, are similar and have good classification results, which is further shown in Figure **??** and Figure **??**.

Then, we would like to try K-means Clustering. First of all, we compute the variance of each variable, which is shown in Table **??**. The variance for different variables are similar, so we can do k-means clustering immediately without scaling.

	PS	EC	EN	GE	H	K	GC
Variance	369.72	181.79	218.61	202.50	415.77	387.40	221.82

Table 5: Variance for Each Variables

First we plot the within-groups sum of squares for one to six-group solutions to see if we can get any indication of the number of groups(seen in Figure **??**). The largest reduction happens at n=2 and so we will look at the two-group solution.

By Figure **??**, when k=2, PC1 score is a good way to separate different

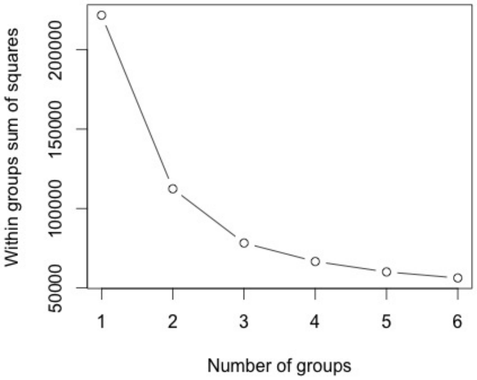

Figure 12: Within Group Sum of Squares Reduction

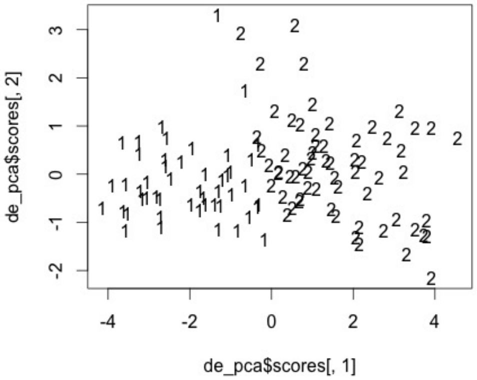

Figure 13: Two Group Solution

types. What's more, the two groups are created essentially on the basis of the first principal component score.

Ridge Regression and PCR

We random sample half of our observations as training data, and the other half as testing data. Using cross-validation to get the regularization parameter λ, which is shown in Figure **??**. The output variable is type, and I use seven features for prediction.
This is a ridge-logit problem for classification. Finally the testing error is about 4.5%. which is very small and shows that the ridge regression is a good way to predict the country type.
PCR(Principal Component Regression) is an another way. One major use of PCR lies in overcoming the multicollinearity problem. In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. This can be particularly useful in settings with high-dimensional covariates. Also, through appropriate selection of the principal components to be used for regression, PCR can lead to efficient prediction of the outcome based on the assumed model.

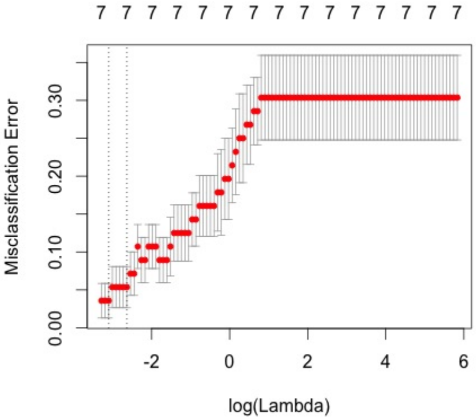

Figure 14: Cross Validation

From table **??**, when the regression is built on first three principal components, it explains 91% variance of features in the training set, but only explains 60% variance of output variables in the training set. Although not as good as expected, it can be used as a good way to characterizing the un-

	1 comps	2 comps	3 comps
X	70.55	82.92	91.31
Country Type	46.59	58.56	59.14

Table 6: Training: % variance explained

derlying model using only three parameters.

Discussion

Limitation:
1. The variable type is a constructed variable, which is somewhat arbitrary, we can construct different variables, and have different results;
2. The size of data is limited. For further study, I could use data from multiple years and for each country, I could use its historical data to predict the type in next year using ridge regression.
Advantages:
1.Among input variables, many of them are highly correlated with each other, which is a good foundation for Principal Component Analysis.
2. The ridge regression is a good method to deal with colineality of features, thus it is very suitable to implement in this case.
3. Two group clustering has good characteristics if we use two-means clustering, or complete, average clustering method.

Conclusion

Summary of results of above data analysis:
[1]. By PCA, three principal component are able to explain 90% of total variance and can be used to replace original data without losing much information;
[2]. By hypothesis testing of number of factors for factor analysis, we reject the null hypothesis. That means, it is not appropriate to implement factor analysis in this case;
[3]. By the characteristics of original data, it is hard to implement corre-

sponding analysis in this case;
[4]. By using three different clustering methods: single, complete and average, we found that single linkage solutions often contain long straggly clusters that do not give a useful description of the data. The two-group solutions from complete linkage and average linkage, are similar and have good classification results;
[5]. By k-means clustering using k equals to 2, we found that the two groups are created essentially on the basis of the first principal component score;
[6]. By ridge regression using half training data and half testing data, implementing cross-validation on training data to find the best shrinkage parameter λ, and predicting on testing data, the final result is that we get only 5% prediction error(testing error), which shows that ridge regression is a good way to classify whether a country is thought to be democratic or not so democratic;
[7]. By principal component regression, when the regression is built on first three principal components, it explains 91% variance of features in the training set, but only explains 60% variance of output variables in the training set. Although not as good as expected, it can be used as a good way to characterizing the underlying model using only three parameters.

References

[1]. http://democracyranking.org/: The Ranking of Quality of Democracy
[2]. Brian Everitt and Torsten Hothorn: An Introduction to Applied Multivariate Analysis with R (2011). Springer, **Chapter 1-6**
[3].Hastie, T., Tibshirani, R., Friedman, J. (2011), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer. ESL, **Page 61-68: Ridge Regression, Page 79-80: Principal Component Regression**
[4]Bishop, Christopher (2007), Pattern Recognition and Machine Learning. Springer. Bishop, **Page 583-586: Factor Analysis**
[5]Alan Izenman Modern Multivariate Statistical Techniques (2008). Springer.**Page 414-420: Agglomerative Nesting, Page 423-424: K-means**

Appendices: r code

```
#data transformation
demo <- read.csv("/Users/suruisun/Downloads/demo1.csv",header=TRUE)
row.names(demo) <- demo[,2]
demo <- demo[,-c(1:2)]
demo$type=0
demo[demo$rank<=75,]$type=1 #demo[demo$rank>75,]$type=0
str(demo)
de=demo[,1:7]

#==================PCA=======================#
s=cor(de)
library(lattice)
levelplot(s)
de_pca <- princomp(de, cor = TRUE)
plot(de_pca)
pca_result=summary(de_pca, loadings = TRUE)
library(scatterplot3d)
scatterplot3d(de_pca$scores[,1],de_pca$scores[,2],de_pca$scores[,3],typ
library(MVA)
pairs(de_pca$scores[,1:3], ylim = c(-6, 4), xlim = c(-6, 4),
      panel = function(x,y, ...) {
        text(x, y, abbreviate(row.names(de)),
             cex = 0.6)
        bvbox(cbind(x,y),add =TRUE)
      })
biplot(de_pca, col = c("gray", "red"))
biplot(de_pca,xlabs=demo$type)

#=============factor analysis and correspondence analysis==============
sapply(1:5, function(f) factanal(de, factors = f, method ="mle")$PVAL)

#=========clustering============#
#cluster analysis page 167
dm <- dist(de)
plot(cs <- hclust(dm, method = "single"),labels=abbreviate(row.names(de
plot(cc <- hclust(dm, method = "complete"),labels=abbreviate(row.names(d
plot(ca <- hclust(dm, method = "average"),labels=abbreviate(row.names(de
body_pc <- princomp(dm, cor = TRUE)
xlim <- range(body_pc$scores[,1])
plot(body_pc$scores[,1:2], type = "n",xlim = xlim, ylim = xlim)
lab <- cutree(cc, h = 125)
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
plot(body_pc$scores[,1:2], type = "n",xlim = xlim, ylim = xlim)
lab <- cutree(ca, h = 75)
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)

#can do k-means by k-means
sapply(de, var)
n <- nrow(de)
wss <- rep(0, 6) #in total we have 7 variables, k-means the maximum is
wss[1] <- (n - 1) * sum(sapply(de, var))
for (i in 2:6)
  wss[i] <- sum(kmeans(de, centers = i)$withinss)
plot(1:6, wss, type = "b", xlab = "Number␣of␣groups", ylab ="Within␣grou

plot(de_pca$scores[,1],de_pca$scores[,2],type='n')
text(de_pca$scores[,1],de_pca$scores[,2],labels=kmeans(de, centers = 2)$

#=======ridge & PCR===========#
library(glmnet)
train=sample(1:nrow(de),nrow(de)/2)
ridge_cv<-cv.glmnet(x=as.matrix(demo[train,1:7]),y=demo[train,9],type.m
lambda_rid<-ridge_cv$lambda.min
ridge.pred=predict(ridge_cv,s=lambda_rid,newx=as.matrix(demo[-train,1:7]
err_rid=sum(as.numeric(ridge.pred)!=demo[-train,9])/length(demo[-train,9
err_rid  #the error rate for ridge regression is rather small#
#PCR:
library(pls)
de_pcr=pcr(demo$type~as.matrix(demo[,1:7]),ncomp=3,data=demo,validation=
summary(de_pcr)
#=========draw worldmap:=========#
library(rworldmap)
demo$country=row.names(demo)
```

```
sPDF <- joinCountryData2Map(demo,joinCode="NAME",nameJoinColumn ="countr
mapCountryData(sPDF ,nameColumnToPlot='type',numCats=2,catMethod = "fixe
```