S&DS 563 / F&ES 758b - Multivariate Statistics
Homework #2 Principle Components Analysis

Lanxin Jiang, Yazhi Sun, Chenglin Lu

2018-02-08

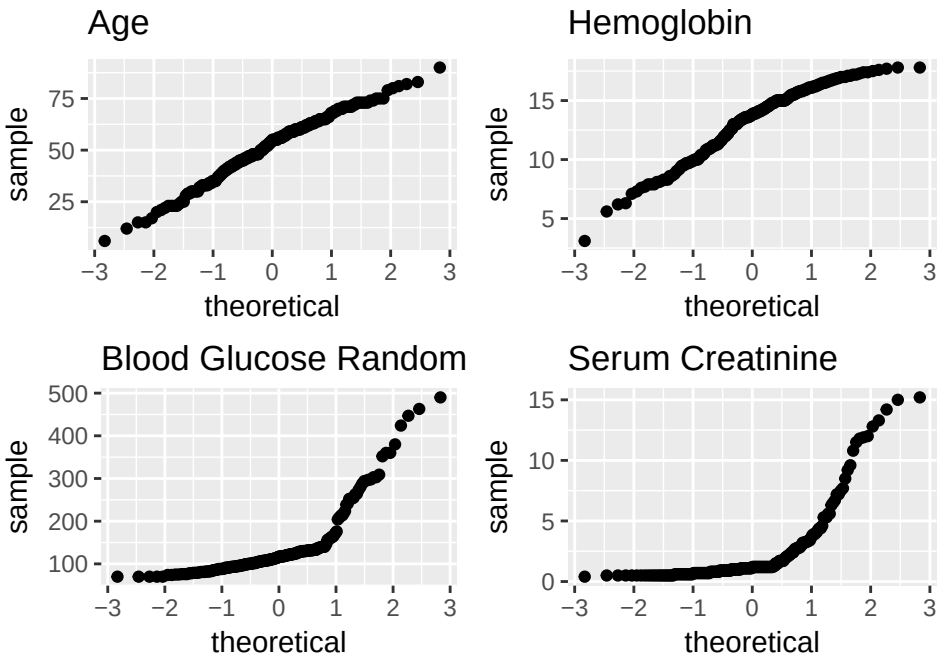
1. First, discuss whether your data seems to have a multivariate normal distribution. Make univariate plots (boxplots, normal quantile plots as appropriate). Then make transformations as appropriate. You do NOT need to turn all this in, but describe what you did.

```
library(RWeka)
library(ggplot2)

CKD <- read.arff("../Chronic_Kidney_Disease/chronic_kidney_disease_full.arff")

# Only Numeric Variables
CKD.numeric <- CKD[,c(1:2,10:18)]
# Remove Missing Observations
CKD.numeric <- CKD.numeric[complete.cases(CKD.numeric),]

gg.qqplot <- function (variable, title) {
    ggplot(CKD.numeric, aes_string(sample=variable)) + ggtitle(title) + stat_qq()
}
p1 <- gg.qqplot("age", "Age")
p2 <- gg.qqplot("bgr", "Blood Glucose Random")
p3 <- gg.qqplot("hemo", "Hemoglobin")
p4 <- gg.qqplot("sc", "Serum Creatinine")
multiplot(p1, p2, p3, p4, cols=2)
```

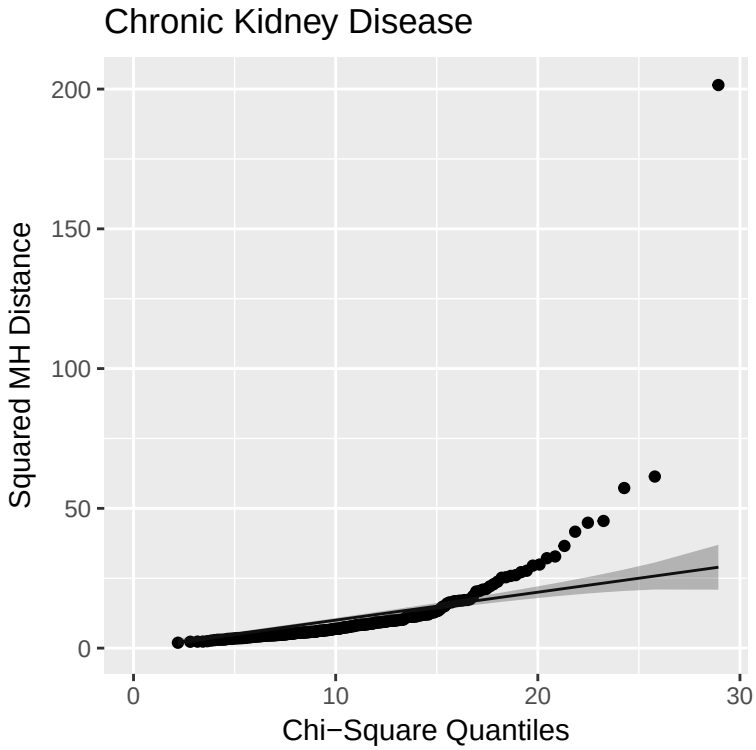


[Interpretation on QQplot, multivariate normal distribution]

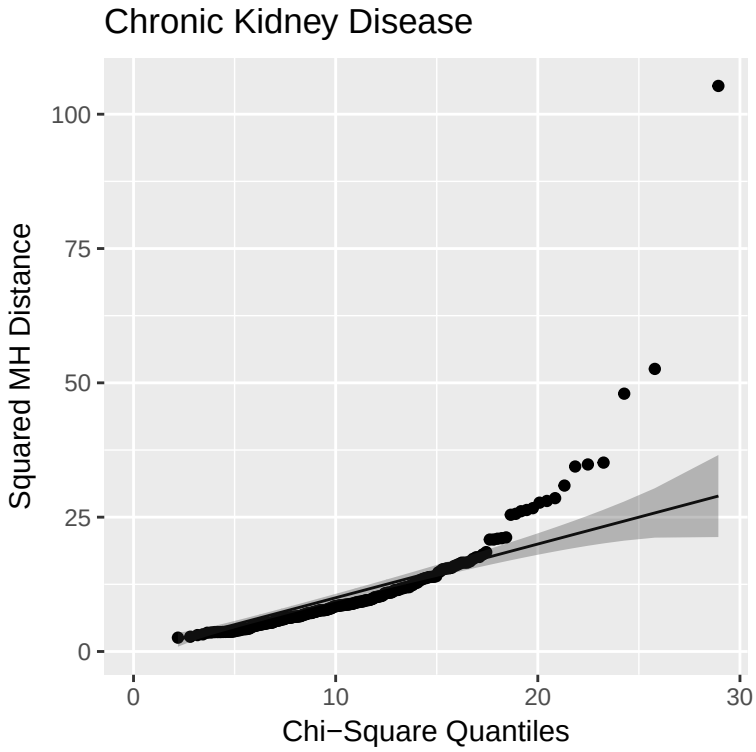
THEN make a chi-square quantile plot of the data. Turn in your chi-square quantile plot as appropriate and comment on what you see.

NOTE that multivariate normality is NOT a requirement for PCA to work!

```
gg.CSQPlot(CKD.numeric,label="Chronic Kidney Disease")
```

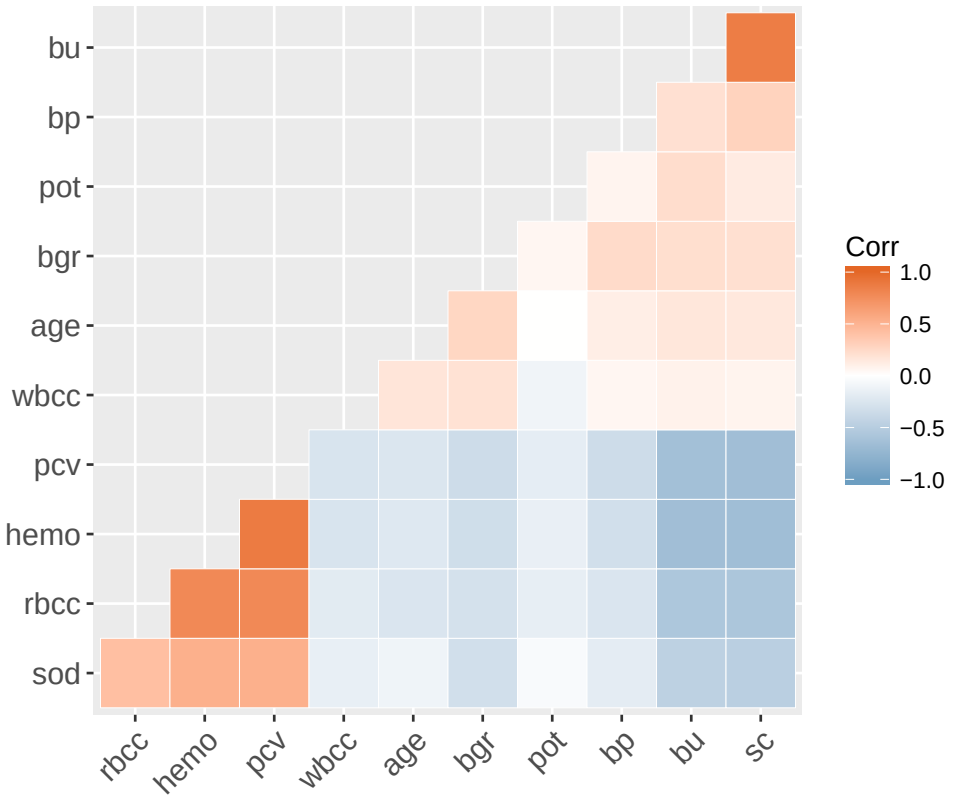



```
gg.CSQPlot(log(CKD.numeric),label="Chronic Kidney Disease")
```



2. Compute the correlation matrix between all variables (SAS and SPSS will provide this for you as part of the PCA procedure – in SPSS, click on DESCRIPTIVES. In R use the cor() function.). Comment on relationships you do/do not observe. Do you think PCA will work well?

```
library(ggcorrplot)
CKD.numeric %>%
  cor(use="pairwise.complete.obs", method = "pearson") %>%
  ggcorrplot(hc.order = TRUE, type = "lower", # lab = TRUE,
    outline.col = "white",
    ggtheme = ggplot2::theme_gray,
    colors = c("#6D9EC1", "white", "#E46726"))
```



3. Perform Principle components analysis using the Correlation matrix (standardized variables). Think about how many principle components to retain. To make this decision look at

- Total variance explained by a given number of principle components
- The 'eigenvalue > 1' criteria
- The 'scree plot elbow' method (turn in the scree plot)
- Parallel Analysis: think about whether this is appropriate based on what you discover in number 1.

```
pc1=princomp(CKD.numeric, cor=TRUE)

#print results
print(summary(pc1),digits=2,loadings=pc1$loadings,cutoff=0)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.1607179 1.1351712 0.99613107 0.94492798
## Proportion of Variance 0.4244275 0.1171467 0.09020701 0.08117172
## Cumulative Proportion  0.4244275 0.5415741 0.63178116 0.71295288
##                           Comp.5     Comp.6     Comp.7     Comp.8
## Standard deviation     0.90177870 0.87723005 0.75196840 0.72895591
## Proportion of Variance 0.07392771 0.06995751 0.05140513 0.04830697
## Cumulative Proportion  0.78688059 0.85683809 0.90824322 0.95655020
##                           Comp.9    Comp.10    Comp.11
## Standard deviation     0.47483048 0.36013316 0.35041109
## Proportion of Variance 0.02049673 0.01179054 0.01116254
## Cumulative Proportion  0.97704693 0.98883746 1.00000000

## Warning in if (loadings) {: the condition has length > 1 and only the first
```

```
## element will be used
```

```
##
## Loadings:
##      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## age  -0.15  -0.44   0.35   0.39   0.61   0.19   0.21  -0.23   0.03
## bp   -0.19  -0.07   0.41  -0.74  -0.09   0.42   0.20  -0.10  -0.08
## bgr  -0.21  -0.39   0.39  -0.14  -0.05  -0.62  -0.13   0.47   0.00
## bu   -0.37   0.29  -0.15   0.09   0.18  -0.02   0.34   0.35  -0.02
## sc   -0.38   0.26  -0.18  -0.06   0.23   0.03   0.32   0.29  -0.11
## sod   0.30   0.02   0.19   0.15   0.04   0.55  -0.26   0.68   0.11
## pot  -0.10   0.42   0.61   0.42  -0.45  -0.03   0.20  -0.11  -0.02
## hemo  0.42   0.00   0.08  -0.03   0.09  -0.11   0.29   0.06  -0.47
## pcv   0.42   0.01   0.03  -0.03   0.09  -0.12   0.27   0.09  -0.45
## wbcc -0.13  -0.56  -0.29   0.22  -0.57   0.23   0.37   0.12  -0.10
## rbcc  0.38   0.00   0.02  -0.13   0.02  -0.15   0.53   0.06   0.73
##      Comp.10 Comp.11
## age   0.01   -0.02
## bp    0.08    0.04
## bgr  -0.02   -0.02
## bu    0.53    0.44
## sc   -0.55   -0.44
## sod  -0.03   -0.01
## pot  -0.05   -0.07
## hemo -0.42    0.55
## pcv   0.47   -0.54
## wbcc -0.02   -0.02
## rbcc -0.04   -0.03
```

```
#view eigenvalues (note that R gives square-root of eigenvalues
#which is why I square them)
pc1$sdev^2
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 4.6687020 1.2886137 0.9922771 0.8928889 0.8132048 0.7695326 0.5654565
##    Comp.8    Comp.9   Comp.10   Comp.11
## 0.5313767 0.2254640 0.1296959 0.1227879
```

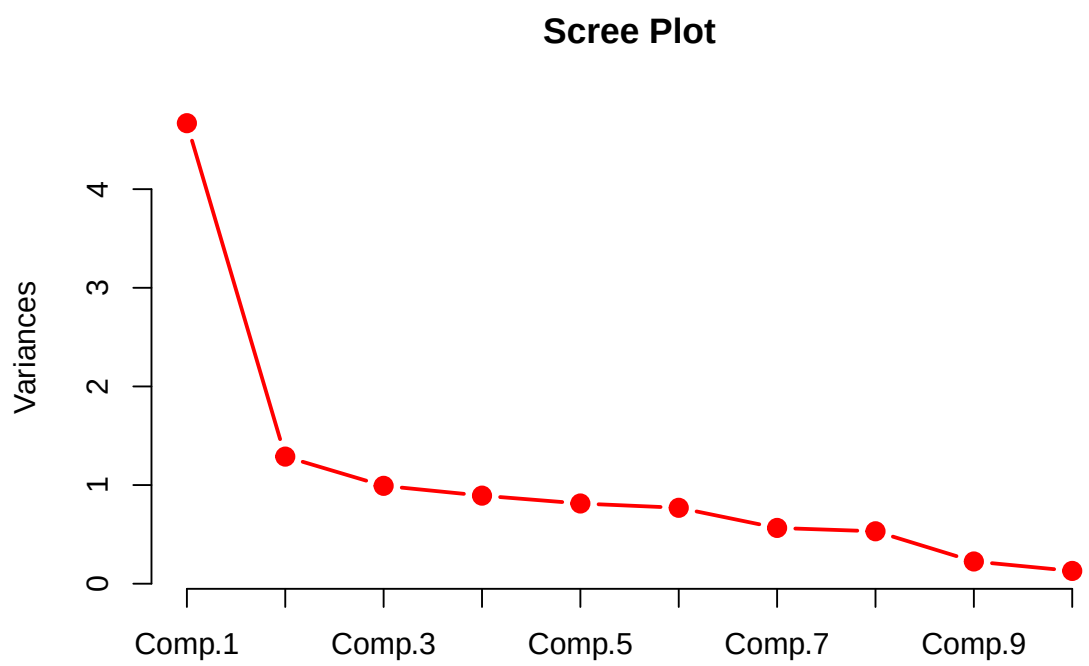
```
#Comp1 and Comp2 larger than 1

#view eigenvectors
pc1$loadings
```

```
##
## Loadings:
##      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## age  -0.150 -0.443  0.352  0.395  0.606  0.192  0.211 -0.226
## bp   -0.192         0.408 -0.737         0.417  0.203 -0.101
## bgr  -0.210 -0.394  0.389 -0.141        -0.620 -0.130  0.475
## bu   -0.372  0.288 -0.150         0.177         0.340  0.350
## sc   -0.377  0.261 -0.181         0.230         0.315  0.293 -0.112
## sod   0.300         0.194  0.152         0.549 -0.262  0.682  0.108
## pot          0.425  0.610  0.424 -0.445         0.199 -0.107
## hemo  0.418                               -0.109  0.289        -0.475
## pcv   0.420                               -0.118  0.269        -0.453
## wbcc -0.131 -0.559 -0.294  0.218 -0.566  0.228  0.373  0.115 -0.104
```

```
## rbcc  0.384               -0.132        -0.154  0.528         0.725
##      Comp.10 Comp.11
## age
## bp
## bgr
## bu    0.530   0.444
## sc   -0.554  -0.442
## sod
## pot
## hemo -0.424   0.553
## pcv   0.469  -0.542
## wbcc
## rbcc
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.091  0.091  0.091  0.091  0.091  0.091  0.091  0.091
## Cumulative Var  0.091  0.182  0.273  0.364  0.455  0.545  0.636  0.727
##                Comp.9 Comp.10 Comp.11
## SS loadings     1.000   1.000   1.000
## Proportion Var  0.091   0.091   0.091
## Cumulative Var  0.818   0.909   1.000
```

```
#make a screeplot
screeplot(pc1,type="lines",col="red",lwd=2,pch=19,cex=1.2,main="Scree Plot")
```



4. For principle components you decide to retain, examine the loadings (principle components) and think about an interpretation for each retained component if possible.

5. Make a score plot of the scores for at least two pairs of component scores (one and two, one and three, two and three, etc). Discuss any trends/groupings you observe. **As a bonus, try to make a 95% Confidence Ellipse for two of your components**. You might want to also try making a bi-plot if you're using R.

6. Write a paragraph summarizing your findings, and your opinions about the effectiveness of using principle components on this data. Include evidence based on scatterplots of linearity in higher dimensional space, note any multivariate outliers in your score plot, comment on sample size relative to number of variables, etc.