2019R2 High-Dimensional Data Analysis (STAT5103) Assignment 3

Yiu Chung WONG 1155017920

Principal Component Analysis (PCA) on uscrime Dataset

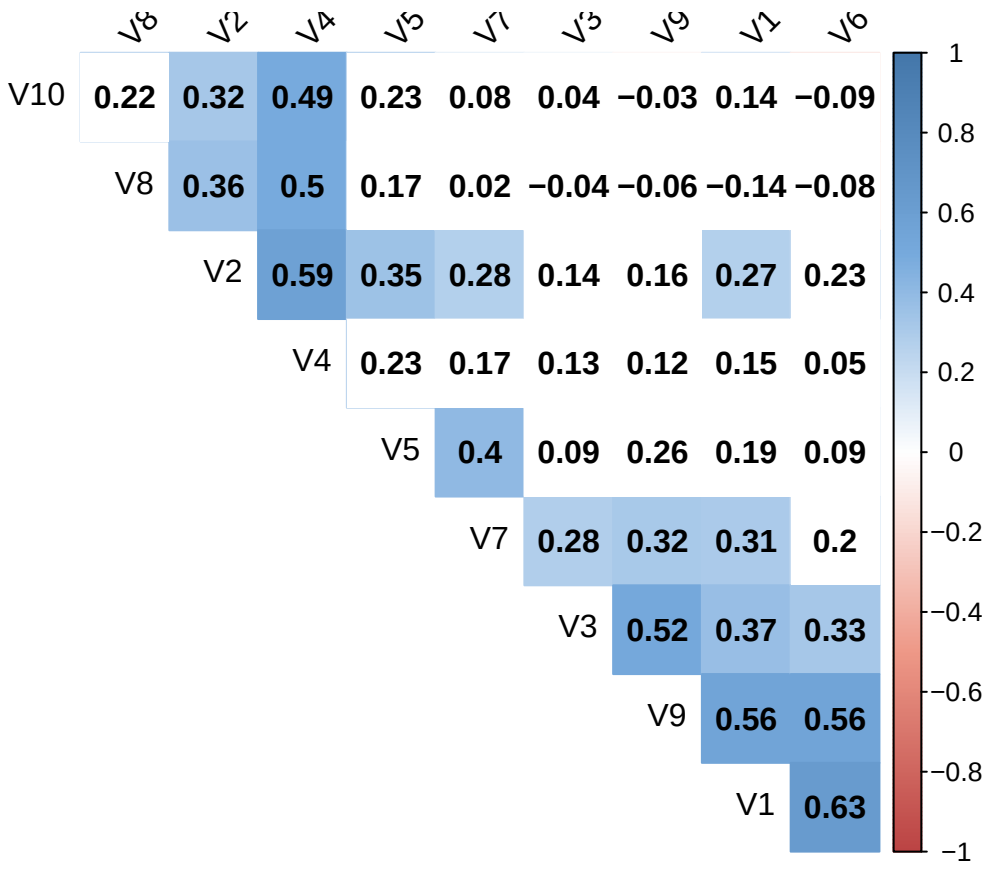
```
# import data
teachers <- read.csv('hw3(2020)a.dat', header = FALSE, sep = '')
```

a)

Correlation matrix

```
# matrix of the p-value of the correlation
p.mat <- cor.mtest(teachers)

col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot::corrplot(cor(teachers), method="color", col=col(200),
                   type="upper", order="hclust",
                   addCoef.col = "black", # Add coefficient of correlation
                   tl.col="black", tl.srt=45, #Text label color and rotation
                   # Combine with significance
                   p.mat = p.mat, sig.level = 0.01, insig = "blank",
                   # hide correlation coefficient on the principal diagonal
                   diag=FALSE
                   )
```



- The colour indicates the strength of correlation: the deeper the blue, the more positive the correlation between two items.
- Correlations that are not coloured are not statistically sigificant.
- Factor analysis assumes at least some items be correlated. From this graph we see that this assumption is valid.
- The items seem to assemble in two different groups. This suggest the possibility of two different latent factors behind the items.

Bartlett's test of sphericity

```
bartlett <- psych::cortest.bartlett(teachers)
bartlett
```

```
## $chisq
## [1] 262.3824
##
## $p.value
## [1] 1.816464e-32
##
## $df
## [1] 45
```

- H_0: All variables are independent
- p-value is $1.8164638 \times 10^{-32}$; H_0 is to be rejected
- Not all variables are independent

KMO

```
kmo <- psych::KMO(teachers)
kmo
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: psych::KMO(r = teachers)
## Overall MSA =  0.72
## MSA for each item =
##   V1   V2   V3   V4   V5   V6   V7   V8   V9  V10
## 0.74 0.76 0.77 0.68 0.68 0.71 0.78 0.69 0.75 0.64
```

- Overall MSA is 0.7237724, whcih is okay

Subject/variable ratio

```
n <- nrow(teachers)
p <- ncol(teachers)
ratio <- n/p
```

- The subject/variable ratio is 8.9. This number is low considering the suggested ratio should be around 10

Scale of data

- The data is from a 10-item, 4-point Likert scale questionnaire.
- Factor analysis assumes interval or ratio variables.
- If ordinal variable is to be used, there should be at least 5 categories.

Eventhough the data show adaquit dependency, the number of category in items is just too low. This results in the strong non-normality in data. The subject to variable ratio is also not satisfacroty. Hence factor analysis on the provided data not recommended.

b)

Principal Components Analysis

```
 #Principle Component using non-centered, non-scaled datas
teachers_pca <- prcomp(teachers)
names(teachers_pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
teachers_pca
```

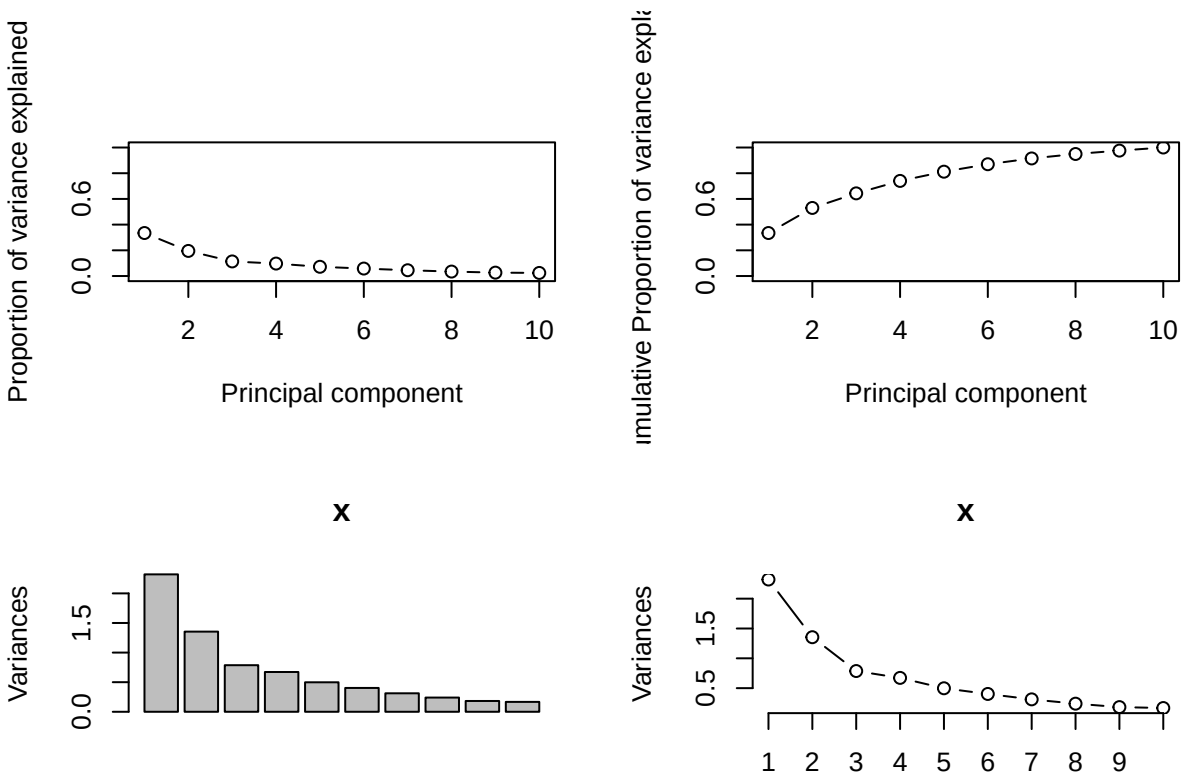
```
## Standard deviations (1, .., p=10):
##  [1] 1.5234287 1.1635773 0.8870355 0.8197833 0.7057196 0.6346637 0.5587597 0.4899041 0.4272766
## [10] 0.4088632
##
## Rotation (n x k) = (10 x 10):
##           PC1         PC2        PC3         PC4         PC5         PC6         PC7         PC8
## V1  0.3423963 -0.13030447  0.1249553 -0.30289719  0.30987420 -0.20406435  0.09114970 -0.01804107
## V2  0.3128334  0.44938280  0.1138146 -0.23557689 -0.40787942 -0.17326123  0.47024915  0.36528426
## V3  0.4578133 -0.29568886  0.2669743  0.71703209 -0.16969009 -0.03652516  0.25000586 -0.15523625
## V4  0.2064996  0.43290601  0.2291022  0.04748169 -0.15202782 -0.04744018 -0.38009599  0.09371373
## V5  0.2223380  0.21304767 -0.3086366 -0.08506694  0.11681593  0.70548931  0.39806828 -0.28486639
## V6  0.3326138 -0.25183781  0.1963398 -0.50506476 -0.01973894 -0.20178959 -0.01962777 -0.51067985
## V7  0.4073389  0.07288394 -0.7967072  0.13014144  0.01541194 -0.34723734 -0.22441612 -0.01516539
## V8  0.0436176  0.30567427  0.1050410  0.03278626 -0.38121445  0.18205301 -0.42197770 -0.49197054
## V9  0.4328727 -0.27472745  0.1007116 -0.12616628  0.06051882  0.47479270 -0.41772320  0.49403611
## V10 0.1318812  0.47244125  0.2407171  0.20027691  0.72280362 -0.08470095 -0.04263258 -0.07128761
##            PC9        PC10
## V1   0.71467021  0.31839203
## V2  -0.20071817  0.20124726
## V3   0.02041917 -0.02044159
## V4   0.34997126 -0.64351367
## V5   0.12708165 -0.18858817
## V6  -0.38675344 -0.28361945
## V7  -0.07312608  0.01823672
## V8   0.04844978  0.54003414
## V9  -0.21320770  0.13406251
## V10 -0.32738600  0.13285578
```

```
summary(teachers_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7    PC8     PC9   PC10
## Standard deviation     1.5234 1.1636 0.8870 0.81978 0.7057 0.63466 0.55876 0.4899 0.42728 0.4089
## Proportion of Variance 0.3346 0.1952 0.1134 0.09689 0.0718 0.05807 0.04501 0.0346 0.02632 0.0241
## Cumulative Proportion  0.3346 0.5298 0.6432 0.74010 0.8119 0.86997 0.91498 0.9496 0.97590 1.0000
```

```
pcaCharts(teachers_pca)
```

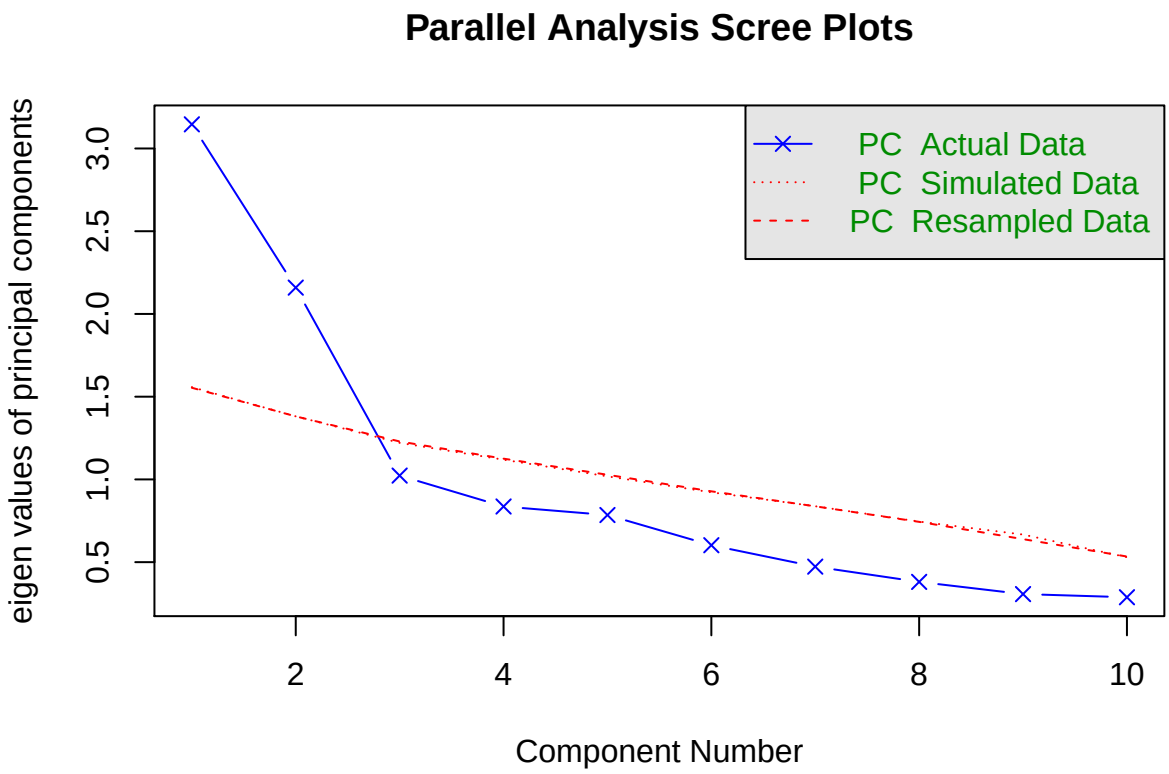
```
## [1] "proportions of variance:"
##  [1] 0.33458712 0.19518903 0.11343497 0.09688645 0.07180081 0.05807006 0.04501063 0.03460088
##  [9] 0.02631983 0.02410022
```



- The scree plot presents the portion of variabce each principle component explains.
- There is no obvious "elbow" indicating sigificant differences of variance explained.

Parallel Analysis

```
psych::fa.parallel(x = teachers, fa = "pc", nfactors = p)
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  2
```

- Parallal analysis suggests the first two principle components exhibits eigenvalues higher than random data.
- This echos the two-group seperation presented by the correlation plot

Principal Component Factor Analysis without rotation

```
pcfa <- psych::fa(r = teachers, nfactors = 2, rotate = "none", fm = "pa")
pcfa_load <- pcfa$loadings[1:p,]
pcfa_com <- pcfa$communality
pcfa_psi <- pcfa$uniquenesses
pcfa_tbl <- cbind(pcfa_load, pcfa_com, pcfa_psi)
pcfa_tbl
```

```
##           PA1        PA2  pcfa_com  pcfa_psi
## V1  0.6855630 -0.3520859 0.5939611 0.4060389
## V2  0.5952682  0.4267551 0.5364641 0.4635359
## V3  0.4785653 -0.2489213 0.2909866 0.7090134
## V4  0.5386745  0.6455960 0.7069645 0.2930355
## V5  0.4212721  0.1658706 0.2049832 0.7950168
## V6  0.5710989 -0.4196027 0.5022204 0.4977796
## V7  0.4645564 -0.0363741 0.2171357 0.7828643
```

```
## V8  0.1842495  0.5449659 0.3309357 0.6690643
## V9  0.6725257 -0.4148205 0.6243669 0.3756331
## V10 0.2839845  0.4363827 0.2710770 0.7289230
```

```
pcfa$Vaccounted
```

```
##                             PA1       PA2
## SS loadings           2.6298585 1.6492368
## Proportion Var        0.2629859 0.1649237
## Cumulative Var        0.2629859 0.4279095
## Proportion Explained  0.6145828 0.3854172
## Cumulative Proportion 0.6145828 1.0000000
```

Maximum Likelihood Factor Analysis without rotation

```
mlm <- psych::fa(teachers, nfactors = 2, rotate = "none", fm="ml")
mlm_load <- mlm$loadings[1:p,]
mlm_com <- mlm$communalities
mlm_psi <- mlm$uniquenesses
mlm_tbl <- cbind(mlm_load, mlm_com, mlm_psi)
mlm_tbl
```

```
##           ML1        ML2   mlm_com   mlm_psi
## V1  0.6364258 -0.4623392 0.6187959 0.3812046
## V2  0.6411359  0.3167015 0.5113559 0.4886449
## V3  0.4391704 -0.3023675 0.2842953 0.7157032
## V4  0.6470531  0.5665984 0.7397109 0.2602886
## V5  0.3976168  0.0723530 0.1633328 0.8366659
## V6  0.5350169 -0.5169175 0.5534472 0.4465533
## V7  0.4170593 -0.1145407 0.1870563 0.8129420
## V8  0.2747542  0.5315152 0.3579983 0.6420017
## V9  0.5841648 -0.4804438 0.5720747 0.4279252
## V10 0.3612089  0.4024712 0.2924529 0.7075451
```

```
mlm$Vaccounted
```

```
##                             ML1       ML2
## SS loadings           2.5931324 1.6873931
## Proportion Var        0.2593132 0.1687393
## Cumulative Var        0.2593132 0.4280525
## Proportion Explained  0.6057977 0.3942023
## Cumulative Proportion 0.6057977 1.0000000
```

Factor loadings can be interpreted like standardized regression coefficients, one could also say that the variable V4‘ has a correlation of 0.6470531 with Factor 1.

The precise value of each loading are not our main concern; we ar?re looking for groups of high values that hopefully make sense and lead to a descriptive factor. Without rotation, all 7 variables load on the first two axes and is currently impossible to see any patterns.

`Robbery` and `Auto_theft` have relatively high Ψ value, this is bad because a high Ψ indicates that particular variable is unique and does not load into any factor well.

If we subtract the Ψ value from 1, we get the column commonality. Commonality is the proportion of variance of the ith variable contributed by the m common factors. Looking at the commonality for the variable V4, which has a value of 4. This value can be interpreted as: 400% of the V4 variance was contributed by the two common factors. Since some of the Ψ values are high, the two factors may not be explaining the overall variance so well.

Sum of squared loadings tells us how much of all observed variance was explained by that factor. Here, the first factor is able to explain 2.5931324 units of variance. Some say a factor is worth keeping if the SS loading is greater than 1. This is the case for both factors factor.

The two factors explains roughly 43% of the total variance.

Since our factor loadings are difficult to interpret, perhaps we can get better results if we perform rotation on the loading.

Maximum Likelihood Factor Analysis with varimax rotation

```
mlmv <- psych::fa(teachers, nfactors = 2, rotate = "varimax", fm="ml")
mlmv_load <- mlmv$loadings[1:p,]
mlmv_com <- mlmv$communalities
mlmv_psi <- mlmv$uniquenesses
mlmv_tbl <- cbind(mlmv_load, mlmv_com, mlmv_psi)
mlmv_tbl
```

```
##              ML1         ML2  mlmv_com  mlmv_psi
## V1   0.783771757  0.06706129 0.6187959 0.3812046
## V2   0.277392098  0.65909691 0.5113559 0.4886449
## V3   0.529933768  0.05888093 0.2842953 0.7157032
## V4   0.118289187  0.85189144 0.7397109 0.2602886
## V5   0.253236713  0.31496866 0.1633328 0.8366659
## V6   0.742832448 -0.04057948 0.5534472 0.4465533
## V7   0.390271093  0.18640409 0.1870563 0.8129420
## V8  -0.140203165  0.58167121 0.3579983 0.6420017
## V9   0.756113401  0.01916553 0.5720747 0.4279252
## V10  0.009625355  0.54070533 0.2924529 0.7075451
```

```
mlmv$Vaccounted
```

```
##                            ML1       ML2
## SS loadings           2.3457641 1.9347613
## Proportion Var        0.2345764 0.1934761
## Cumulative Var        0.2345764 0.4280525
## Proportion Explained  0.5480084 0.4519916
## Cumulative Proportion 0.5480084 1.0000000
```

After Varimax rotation, the factors are also a little more clear to interpret. `Murder`and `Assault`, are heavily loaded onto ML1. So it's clear that this is the Violence factor. The rest load heavily onto ML2, which maybe summarised as the 'Theft' factor. The variable `Rape` exhibits cross load; it is loaded onto both factors roughly 50/50. Perhaps Theft and Rape often occur at the same time, which does not sound surprising.

Both SS loadings remain greater than 1. Also, the SS loadings are more evenly divided between both factors than before rotation. The difference between the variance explained among the two factors also narrowed, but the sum remains the same. Therefore rotation is able to better separate the latent factors using our variables, but does not improve the relationship between variables and factors. This is also evident by looking at the Ψ values, which are exactly the same as before rotation.

Principal Component Factor Analysis with varimax rotation

```
pcfav <- psych::fa(r = teachers, nfactors = 2, rotate = "varimax", fm="pa")
pcfav_load <- pcfav$loadings[1:p,]
pcfav_com <- pcfav$communality
pcfav_psi <- pcfav$uniquenesses
pcfav_tbl <- cbind(pcfav_load, pcfav_com, pcfav_psi)
pcfav_tbl
```

```
##               PA1         PA2 pcfav_com pcfav_psi
## V1   0.7667310962  0.07800366 0.5939611 0.4060389
## V2   0.2668924425  0.68207956 0.5364641 0.4635359
## V3   0.5369374360  0.05181467 0.2909866 0.7090134
## V4   0.1002589510  0.83481294 0.7069645 0.2930355
## V5   0.2630153682  0.36851884 0.2049832 0.7950168
## V6   0.7074906813 -0.04095493 0.5022204 0.4977796
## V7   0.4094485814  0.22245802 0.2171357 0.7828643
## V8  -0.1422148122  0.55741426 0.3309357 0.6690643
## V9   0.7899572954  0.01828637 0.6243669 0.3756331
## V10  0.0005639991  0.52065029 0.2710770 0.7289230
```

```
pcfav$Vaccounted
```

```
##                            PA1       PA2
## SS loadings           2.3390880 1.9400073
## Proportion Var        0.2339088 0.1940007
## Cumulative Var        0.2339088 0.4279095
## Proportion Explained  0.5466314 0.4533686
## Cumulative Proportion 0.5466314 1.0000000
```

Principal Component Factor Analysis gives even clearer separation than Maximum Likelihood Factor Analysis. Variables in our 'Theft' factor 'Violence factor have more even loadings than before.

The distance difference between the two methods can be seen in the Ψ values. Principal Component Factor Analysis gives lower Ψ values which sums up to 5.7209047; whereas Ψ values from Maximum Likelihood Factor Analysis sums up to 5.7194745. By using Principal Component Factor Analysis, latent factors explains more variation of each of our variables. Both SS loadings and Proportion Variance are higher using Principal Component Factor Analysis.

The fact that Principal Component Factor Analysis finds latent factors which explains more variation is because PCA is inherently a method for finding directions/rotations of maximum variance from data sets.

d)