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

Yiu Chung WONG 1155017920

```r
pcaCharts <- function(x) {
    x.var <- x$sdev ^ 2
    x.pvar <- x.var/sum(x.var)
    print("proportions of variance:")
    print(x.pvar)

    par(mfrow=c(2,2))
    plot(x.pvar,xlab="Principal component",
         ylab="Proportion of variance explained", ylim=c(0,1), type='b')
    plot(cumsum(x.pvar),xlab="Principal component",
         ylab="Cumulative Proportion of variance explained", ylim=c(0,1), type='b')
    screeplot(x)
    screeplot(x,type="l")
    par(mfrow=c(1,1))
}


cor.mtest <- function(data, ...) {
    data <- as.matrix(data)
    n <- ncol(data)
    p.data<- matrix(NA, n, n)
    diag(p.data) <- 0
    for (i in 1:(n - 1)) {
        for (j in (i + 1):n) {
            tmp <- cor.test(data[, i], data[, j], ...)
            p.data[i, j] <- p.data[j, i] <- tmp$p.value
        }
    }
  colnames(p.data) <- rownames(p.data) <- colnames(data)
  p.data
}
```

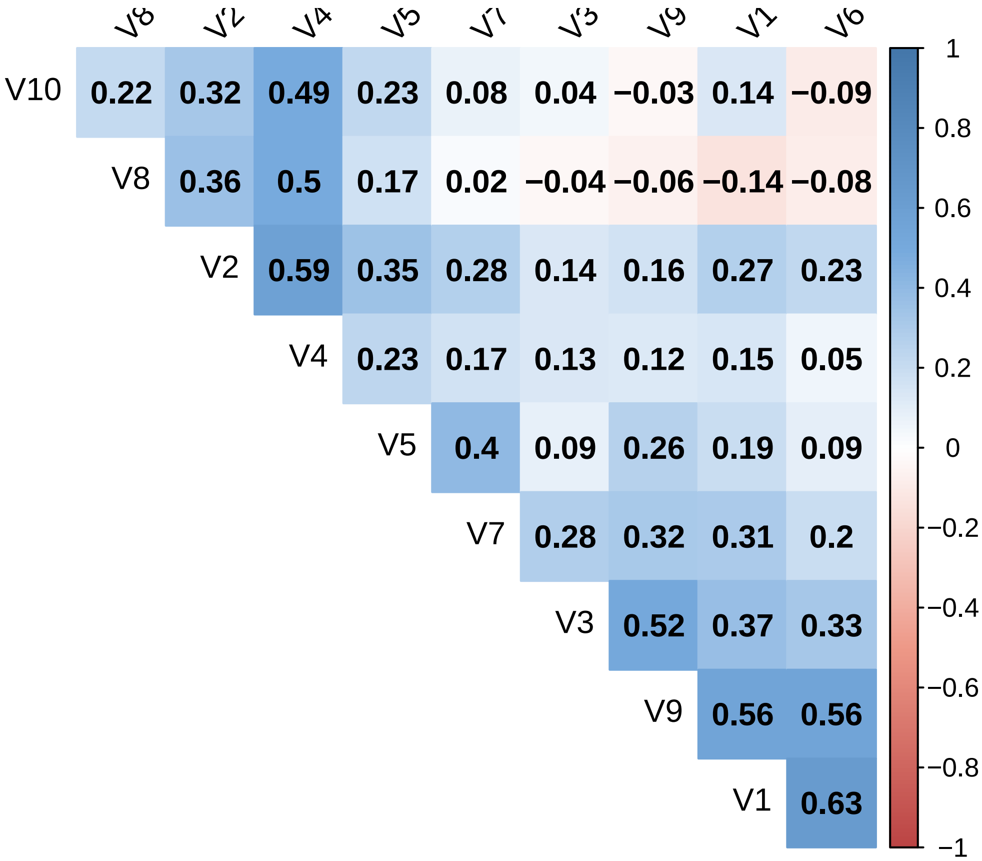
```
# import data
teachers <- read.csv('hw3(2020).dat', header = FALSE, sep = '')
```

a)

Correlation matrix

```
# matrix of the p-value of the correlation
p.data <- 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.data = p.data, 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 significant.
- 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, which 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.

Even though the data show adequate 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 satisfactory. 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
```

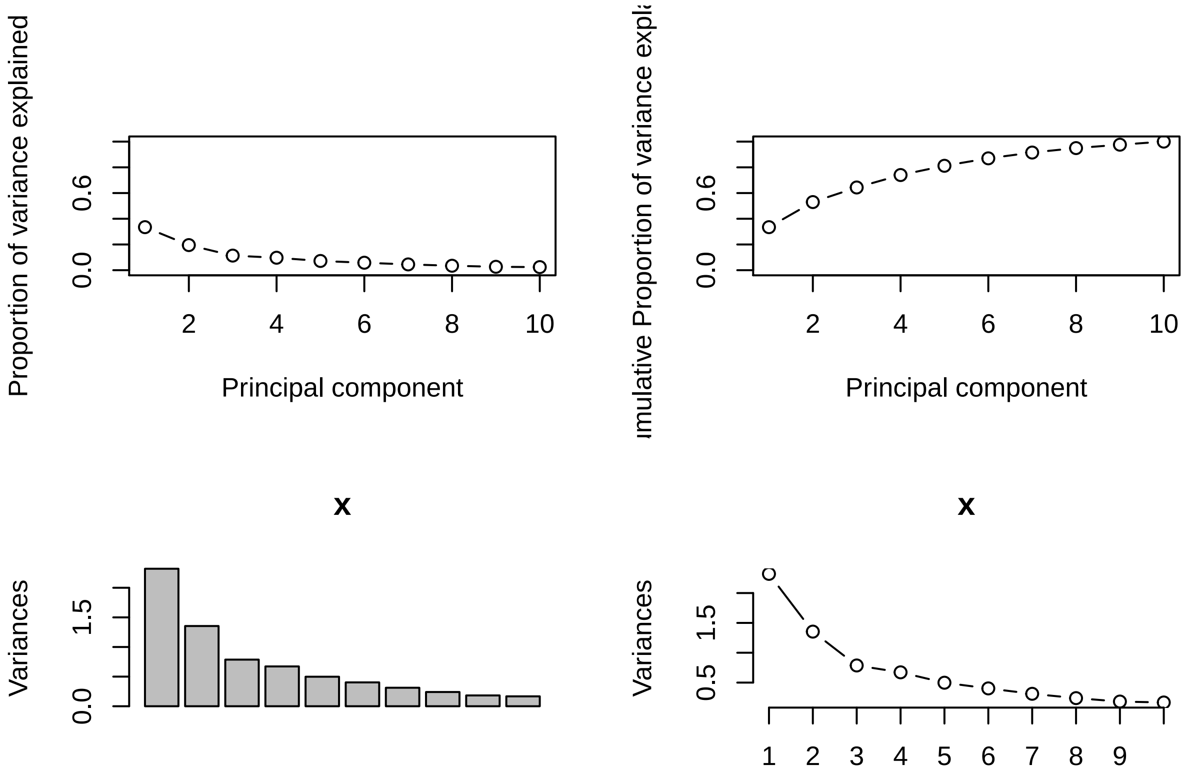
```
## Standard deviations (1, .., p=10):
##  [1] 1.5234287 1.1635773 0.8870355 0.8197833 0.7057196 0.6346637 0.5587597 0.4899041 0.4272766
## [10] 0.4088632
## 
## Rotation (n x k) = (10 x 10):
##           PC1         PC2        PC3         PC4         PC5         PC6         PC7         PC8
## V1  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)
```

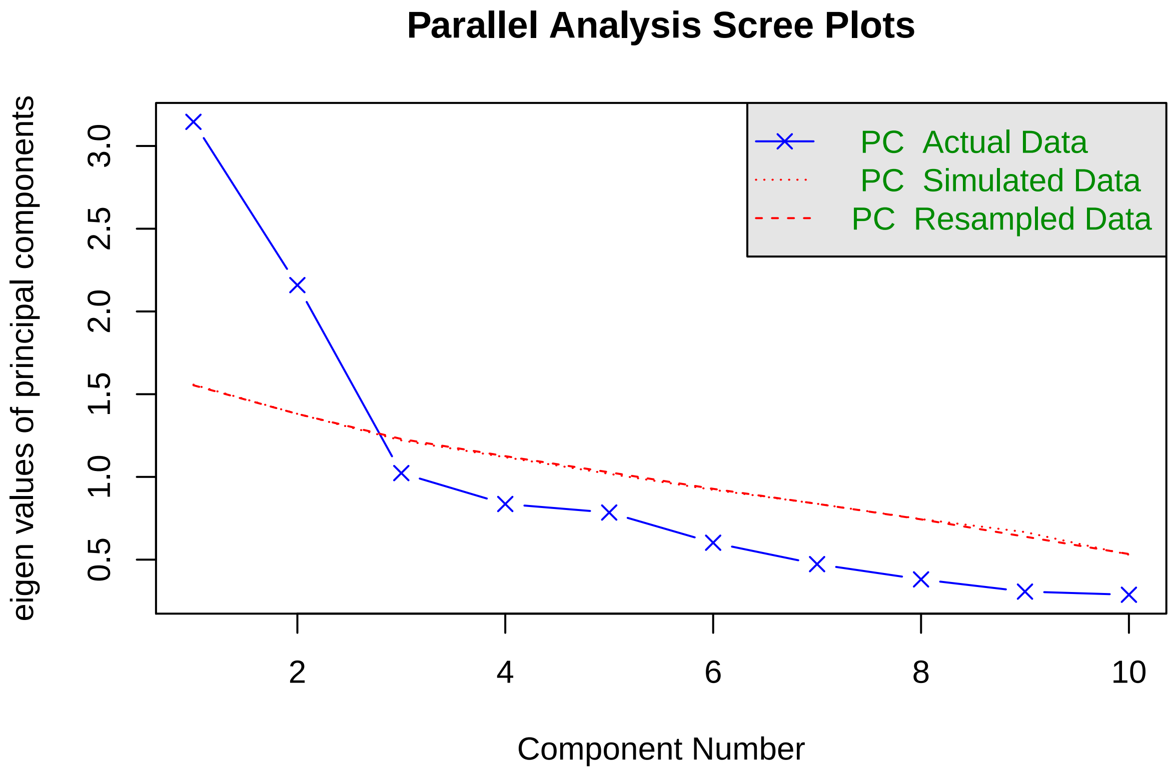
```
## [1] "proportions of variance:"
##  [1] 0.33458712 0.19518903 0.11343497 0.09688645 0.07180081 0.05807006 0.04501063 0.03460088
##  [9] 0.02631983 0.02410022
```



- The scree plot presents the portion of variance each principle component explains.
- There is no obvious "elbow" indicating significant differences in variance explained by principle components.

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
```

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

Principal Axis Factoring without rotation

Here we perform principal axis factoring with 2 latent factors

```
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
```

- Without rotation, some items such as $V1$, $V2$, $V4$, $V9$ exhibits strong cross loading.
- Communlarities for some item is as low as 0.2049832
- Both SS loadings are greater than 1. Factors are kept if this value is greater than 1 according to common practice.
- Perhaps another factor is needed.

Here we perform principal axis factoring with 3 latent factors

```
pcfa <- psych::fa(r = teachers, nfactors = 3, 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         PA3  pcfa_com  pcfa_psi
## V1  0.6679029 -0.38333152 -0.12050820 0.6075596 0.3924404
## V2  0.5908130  0.37953960 -0.02591538 0.4937820 0.5062180
## V3  0.4628661 -0.26604082 -0.06953206 0.2898574 0.7101426
## V4  0.5879984  0.68138084 -0.29761848 0.8985988 0.1014012
## V5  0.5028051  0.18784156  0.60368477 0.6525328 0.3474672
## V6  0.5603813 -0.45777959 -0.19965715 0.5634523 0.4365477
## V7  0.4799587 -0.05742989  0.28887669 0.3171083 0.6828917
## V8  0.1964952  0.52267091 -0.06323537 0.3157940 0.6842060
## V9  0.6486938 -0.43215824 -0.01028713 0.6076702 0.3923298
## V10 0.2926691  0.41536714 -0.02808550 0.2589738 0.7410262
```

```
pcfa$Vaccounted
```

```
##                             PA1       PA2        PA3
## SS loadings           2.6974112 1.7066712 0.60124665
## Proportion Var        0.2697411 0.1706671 0.06012467
## Cumulative Var        0.2697411 0.4404082 0.50053290
## Proportion Explained  0.5389079 0.3409708 0.12012130
## Cumulative Proportion 0.5389079 0.8798787 1.00000000
```

- There is no holistic improvement for the cross loading issue; some item exhibits even stronger cross loading among factors, such as $V4$.
- Communlarities for item $V2$, $V3$, $V4$, $V9$, $V10$ went down.
- SS loading for the third factor is below 1. This indicts a weak relation between item and factor.
- 2 latent factors maybe more suitable than 3.

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
```

- After rotation, SS loadings are more evenly divided between the factors. This can be interpreted as: items no longer as dependent on one particular factor.
- The difference between the variance explained among the two factors also narrowed.
- Item 1, 3, 6, 9 appear to factor into one latent variable; item 2, 4, 8, 10 factor into another.
- Item 5 and 7 does not factor into either latent variables. As is evident by their low value in communlarities and cross loading.
- Items 1, 3, 6, 9 are questions that sought to measure whether teachers relate with others in times of difficulties. This maybe labeled "Emotional support".
- Items 2, 4, 8, 10 are about the steps and approach in tackling difficulties in work. This maybe labeled as "Problem solving"
- Items 5 and 7 involves handling one's emotion but not in relation to other people. This explains the inability to load onto the first factor. Perhaps many participant do not regard medicating one's emotion as problem solving, hence the poor loading onto the second factor.

d)

- The factor solution is not a simple structure, meaning there exist cross loading and/or general factor.
- Even though Varimax rotation provides a solution that better separates items into two factors, cross loading remains.
- Many items have uniqueness well above 0.6, even as high as 0.7950168 . For these items, there are much variance not captured by the proposed factors.
- The model explains 42.8% of the total variance at most; more than half of the variance is not answered.

The factor solution is poor.