Contents

1 Question 1

1.1 Q.1a Simulate one dataset

To simulate one dataset with 100 observations containing 9 variables, the function `dataGen()` from the `rospca` package is used. The arguments of the function are specified as:

- `m=1`: generates 1 dataset;
- `n=100`: generates 100 observations;
- `p=9`: generates 9 variables;
- `bLength=3`: defines that the number of variables contained in each useful group is 3;
- `a=c(0.7,0.9,0.8)`: gives the inner correlation for each group of variables, and here we have `k=2` useful groups of variables, where the value of `k` is calculated by `k=length(a)-1`. In this case, the inner correlation between different variables is equal to `a[1]=0.7` for group 1 and equal to `a[2]=0.9` for group 2. The third group contains the remaining `p-k*bLength=3` variables, of which the inner correlation is equal to `a[3]=0.8`;
- `SD=c(20,5,10)`: gives the standard deviation for each group of variables, and the length of `SD` is equal to the length of the inner correlation vector, which is 3. Here the standard deviation for 3 groups of 3 variables each is equal to 20, 5 and 10 respectively.

The R codes used to simulate the dataset are showed as below.

```
library(rospca)
N <- 100 # number of observations
P <- 9   # number of variables
var <- 3 # number of variables for each group
corr <- c(0.7, 0.9, 0.8) # inner correlation for each group
sd <- c(20, 5, 10)   # standard deviation for each group
# simulate one dataset
simulate1 <- dataGen(m = 1, n = N, p = P, bLength = var, a = corr, SD = sd)
data1 <- data.frame(simulate1[[1]])
```

1.2 Q.1b Check the estimated correlations and standard deviations

In order to check the standard deviation of each variable for 100 observations, the function `apply()` is used. Similarly, the function `inner.outer()` from the package `multicon` is used

to calculate the average inner correlation for each group of 3 variables, as well as the average between-group correlation. The R codes and the results returned are shown as below.

```
# check the standard deviations
round(apply(data1, 2, sd), 2)
```

```
##    X1    X2    X3    X4    X5    X6    X7    X8    X9
## 20.07 20.52 20.07  4.64  4.59  4.64  8.42 10.40  9.94
```

```
# check the inner correlations
library(multicon)
# construct a list indicating the items belonging to each group
list1 <- list(data1[,1:3], data1[,4:6], data1[,7:9])
# return the average within-group and between-group correlation
round(inner.outer(list1), 2)
```

```
##         Set1 Set2 Set3
## Inner r 0.68 0.86 0.76
## Outer r 0.02 0.01 0.01
```

Table 1 is constructed to compare the estimated standard deviations and the simulated values assigned in the `dataGen()` function for each variable, as well as their differences. Similarly, the comparison among inner correlations of each group is shown in Table 2. It is seen from the table that the estimated results are close to the values that are used to generate the simulated data, with the absolute differences of standard deviations ranging from 0.1 to 1.6 and the absolute differences of inner correlations between 0.02 and 0.04. Additionally, by checking the between-group correlation matrix, we find that the correlation between variables from different groups is negligible, which corresponds to the setting of the data generating function.

Table 1: Comparisons between the estimated standard deviations and the simulated values for 9 variables.

	X1	X2	X3	X4	X5	X6	X7	X8	X9
sd.estimated	20.1	20.5	20.1	4.6	4.6	4.6	8.4	10.4	9.9
sd.simulated	20.0	20.0	20.0	5.0	5.0	5.0	10.0	10.0	10.0
difference	0.1	0.5	0.1	-0.4	-0.4	-0.4	-1.6	0.4	-0.1

Table 2: Comparisons of inner correlations for each group of 3 variables.

	Set1	Set2	Set3
corr.estimated	0.68	0.86	0.76
corr.simulated	0.70	0.90	0.80
difference	-0.02	-0.04	-0.04

1.3 Q.1c Simulate another dataset

To simulate another dataset with all the inner correlations equal to zero and the other settings the same as those of question 1a, the arguments in the `dataGen()` remain the same except the inner correlation vector for each group of 3 variables is replaced with a vector of 3 elements, which are all equal to zero. The R codes are shown below.

```
# simulate another dataset with all the inner correlations equal to zero
simulate2 <- dataGen(m = 1, n = N, p = P, bLength = var,
                     a = c(0,0,0), SD = sd)
data2 <- data.frame(simulate2[[1]])
```

Table 3 and Table 4 are constructed to compare the estimated standard deviations and inner correlations with the assigned values. As showed in the table, the differences are negligible, with the absolute differences of standard deviations ranging from 0.1 to 1.3 and the absolute differences of inner correlations between 0.02 and 0.08.

Table 3: Comparisons between the estimated standard deviations and the simulated values for 9 variables.

	X1	X2	X3	X4	X5	X6	X7	X8	X9
sd.estimated	21.1	20.6	19.9	4.8	4.4	4.5	8.7	11.1	10.7
sd.simulated	20.0	20.0	20.0	5.0	5.0	5.0	10.0	10.0	10.0
difference	1.1	0.6	-0.1	-0.2	-0.6	-0.5	-1.3	1.1	0.7

Table 4: Comparisons of inner correlation for each group of 3 variables.

	Set1	Set2	Set3
corr.estimated	-0.02	0.02	-0.08
corr.simulated	0.00	0.00	0.00
difference	-0.02	0.02	-0.08

1.4 Q.1d Perform principal component analysis

The function `prcomp` is used to perform principal component analysis (PCA) on both of the simulated datasets `data1` and `data2`. Prior to the analysis, variables in both datasets are scaled to have the unit variance, as the principal component loadings and the percentage of variance explained by each component differ when the scale of variables changes. In this study, scaling is applied to compare the influence of correlated and uncorrelated data on the PCA results, thus excluding that the impact of variables having variances of different magnitudes.

The principal component loadings are shown in Table 5 and Table 6 for `data1` and `data2` respectively. For the simulated `data1`, the first principal component is mainly given by the three variables `X4`, `X5` and `X6`, while the second one is mainly driven by `X7`, `X8` and `X9` and the third one has more weights on the other three variables `X1`, `X2` and `X3`. Similarly, Figure 1(a) indicates that the first two principal components are dominated by the two groups of variables of which the inner correlations are 0.86 and 0.76 respectively. According to the first three loadings for the simulated `data2`, where variables are approximately uncorrelated, the relationship between a particular group of variables and larger associated coefficients does not seem to exist. As indicated in Figure 1(b), variables contribute more evenly to each of the first two principal components. Therefore, we conclude that a group of variables, with high correlations within the group and negligible between-group correlations, would drive a certain principal component, which also explains large variance in the data.

Additionally, the eigenvalues, the proportion of variance explained by each component and the cumulative proportion are calculated. As shown in tables, the first three principal components explain approximately 84.8% of the total variance in `data1`, while each of the 9 components explains no more than 15.0% of the total variance in `data2`. Therefore, PCA can result in reducing dimensionality in `data1` by using the first three principal components to represent the variance of the data, which could be used in certain subsequent analysis. The conclusion is also supported by scree plots in Figure 2. The scree plot for `data1` reveals a significant decline and plateau after the first three components, while there is no obvious elbow in the scree plot for `data2`, where all the components explain small proportions of variance.

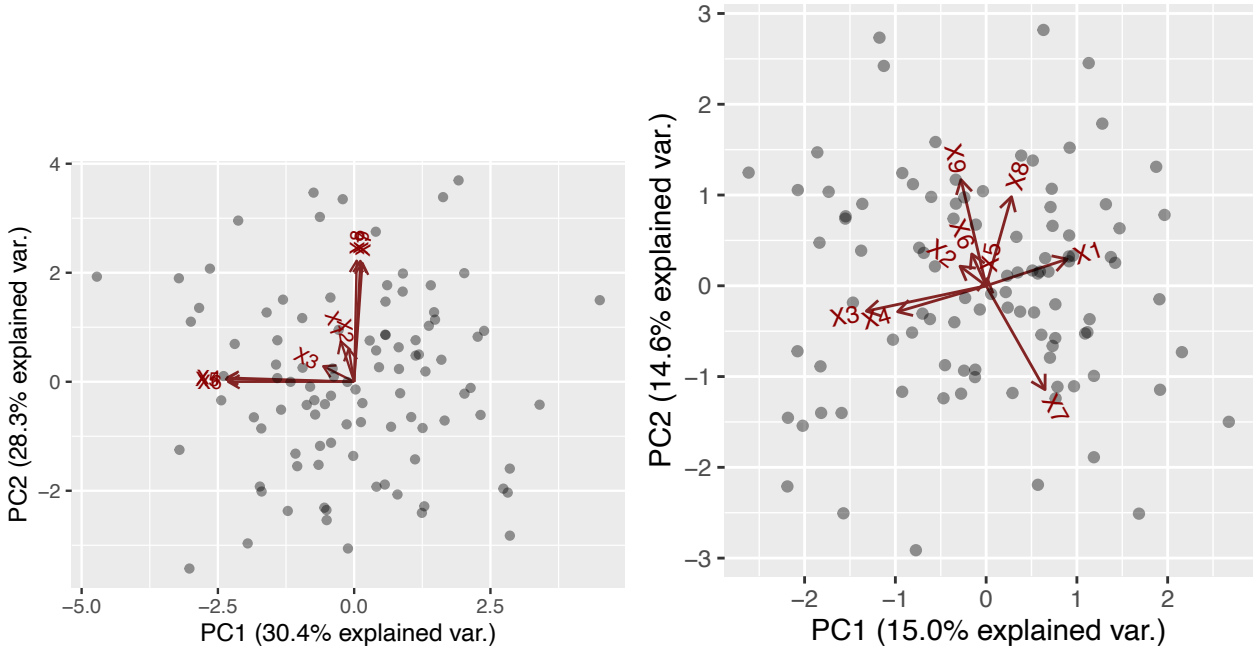
Overall, since PCA is mainly used for finding a set of uncorrelated components to replace the original variables, thus resulting in better interpretability or reduction in dimensionality of the data, the analysis is of better use when the variables are correlated. It seems pointless if all the observed variables are approximately uncorrelated.

Table 5: The loadings associated with the original variables, eigenvalues and the proportion of variance explained by each principal component, as well as the cumulative proportion for the first simulated dataset.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
X1	-0.06	0.19	-0.53	-0.80	0.08	-0.09	0.15	-0.05	0.09
X2	-0.02	0.15	-0.57	0.44	-0.59	-0.23	0.19	-0.08	0.16
X3	-0.14	0.07	-0.57	0.32	0.49	0.34	-0.34	0.15	-0.22
X4	-0.57	0.02	0.08	0.06	0.11	0.13	-0.16	-0.58	0.53
X5	-0.57	0.01	0.07	-0.08	-0.23	-0.14	-0.04	-0.18	-0.74
X6	-0.56	0.00	0.09	0.02	0.02	-0.02	0.27	0.73	0.25
X7	0.04	0.55	0.15	-0.11	-0.39	0.70	-0.06	0.08	-0.02
X8	0.01	0.56	0.13	0.00	0.03	-0.52	-0.60	0.17	0.10
X9	0.03	0.56	0.12	0.22	0.43	-0.12	0.61	-0.20	-0.13
Variance	2.74	2.55	2.35	0.37	0.27	0.25	0.22	0.16	0.10
% Total Var.	30.4%	28.3%	26.1%	4.1%	3.0%	2.8%	2.5%	1.7%	1.2%
Cumulative %	30.4%	58.7%	84.8%	88.9%	91.9%	94.6%	97.1%	98.8%	100.0%

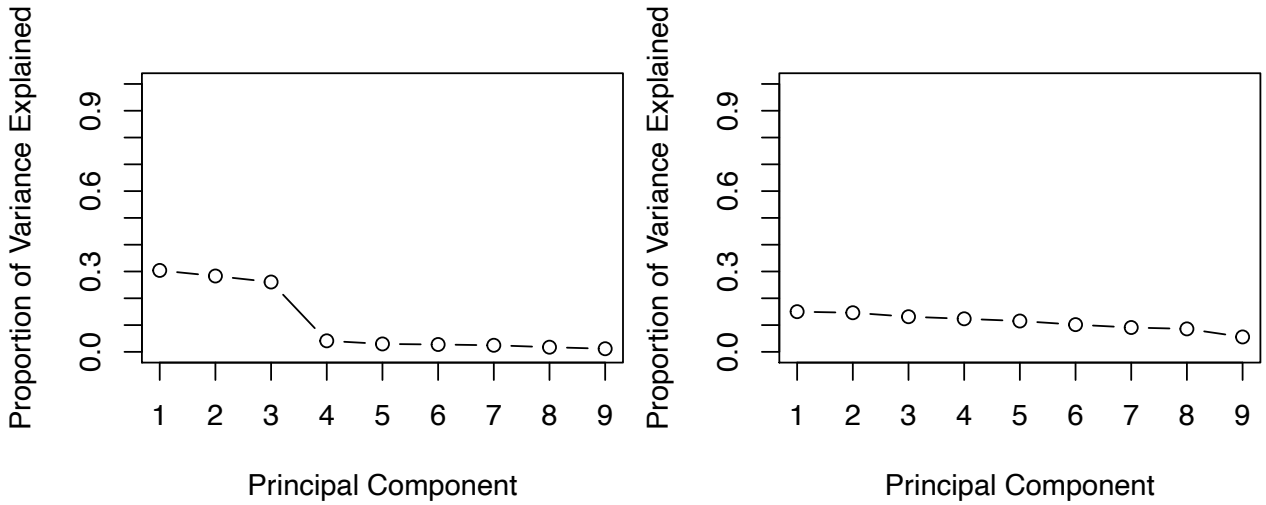
Table 6: The loadings associated with the original variables, eigenvalues and the proportion of variance explained by each principal component, as well as the cumulative proportion for the second simulated dataset.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
X1	0.44	0.15	-0.31	-0.07	-0.11	-0.75	-0.17	-0.14	0.26
X2	-0.14	0.11	0.61	0.02	-0.56	-0.18	0.34	-0.01	0.36
X3	-0.65	-0.14	-0.04	-0.10	-0.15	-0.44	-0.13	-0.30	-0.48
X4	-0.48	-0.14	-0.54	0.31	0.03	-0.04	0.29	0.17	0.50
X5	0.01	0.04	-0.35	-0.52	-0.57	0.16	-0.03	0.48	-0.15
X6	-0.08	0.17	-0.04	-0.69	0.38	-0.01	0.50	-0.26	0.13
X7	0.32	-0.57	0.00	0.12	0.02	-0.23	0.59	0.19	-0.35
X8	0.14	0.49	-0.31	0.33	-0.30	0.21	0.37	-0.42	-0.30
X9	-0.14	0.58	0.12	0.15	0.30	-0.30	0.13	0.59	-0.24
Variance	1.35	1.31	1.18	1.11	1.04	0.91	0.82	0.77	0.50
% Total Var.	15.0%	14.6%	13.1%	12.3%	11.5%	10.2%	9.1%	8.6%	5.6%
Cumulative %	15.0%	29.6%	42.7%	55.1%	66.6%	76.8%	85.9%	94.4%	100.0%



(a) First simulated data **(b)** Second simulated data

Figure 1: The first two principal component loading vectors for two simulated datasets.



(a) First simulated data **(b)** Second simulated data

Figure 2: The scree plot to distinguish the proportion of variance explained by 9 principal components for both of the two simulated datasets.

2 Question 2

2.1 Methods

Prior to multidimensional scaling (MDS) on the data set, mean and variance of each variable are calculated and the results are shown in Table 7. The mean and variance of `GDP` are significantly large, compared with those of other variables. It is recommended to standardise the data via the function `scale()` before Euclidian distances are computed, which can be realised by the function `dist()` as all the variables are numeric. MDS for one dimension and two dimensions is then performed via the function `cmdscale()`.

Table 7: Mean and variance comparison among five variables.

	Increase	Life	IMR	TFR	GDP
Mean	1.60	64.39	44.80	3.46	6349.31
Variance	1.71	79.69	1381.33	3.57	65534456.39

To visualise MDS ratio, the one-dimensional solution is presented as 25 points on a line segment while the two-dimensional solution is presented as a scatter plots with 25 scatter points in a two dimensional space. By analysing the relative location of these points, it is possible to extract meaningful insights and conclusions. Additionally, PCA is also conducted to help interpret the solutions.

To assess the goodness-of-fit of both solutions, the stress value and Shepard plot are applied. The criteria used are as below: stress value close to zero would indicate that the MDS solution is a good fit. For Shepard plot, if the MDS solution is a good fit, the plot should show a linear relationship with 45-degree slope.

2.2 Results and conclusions

The first conclusion is that one dimension solution reveals the overall development level of different countries while two dimensions solution put higher weight on economic factor. From one-dimensional solution in Figure 3, it is observed that Asian and African countries are located on the left side while US and European countries are located at the right side. It coincides with the fact that generally speaking, those US and European countries are

more developed. Thus, one-dimensional solution is an indicator of the overall economic and demographic variables.

Figure 4 supports similar observations as developed countries are clustered at the left bottom. However, Romania and Croatia stand out from other European countries in dimension 2 despite being similar in demographic characteristics. This could be resulted from that dimension 2 puts higher weight on economic factor (GDP). This is also supported in Table 8 of PCA results that GDP has the biggest coefficient (0.82) among all the loadings of five variables for PC2.

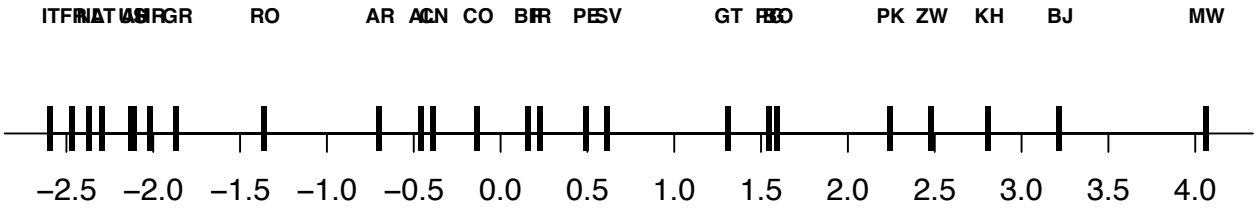


Figure 3: Configuration of countries from a one dimensional ratio MDS of economic and demographic indicators.

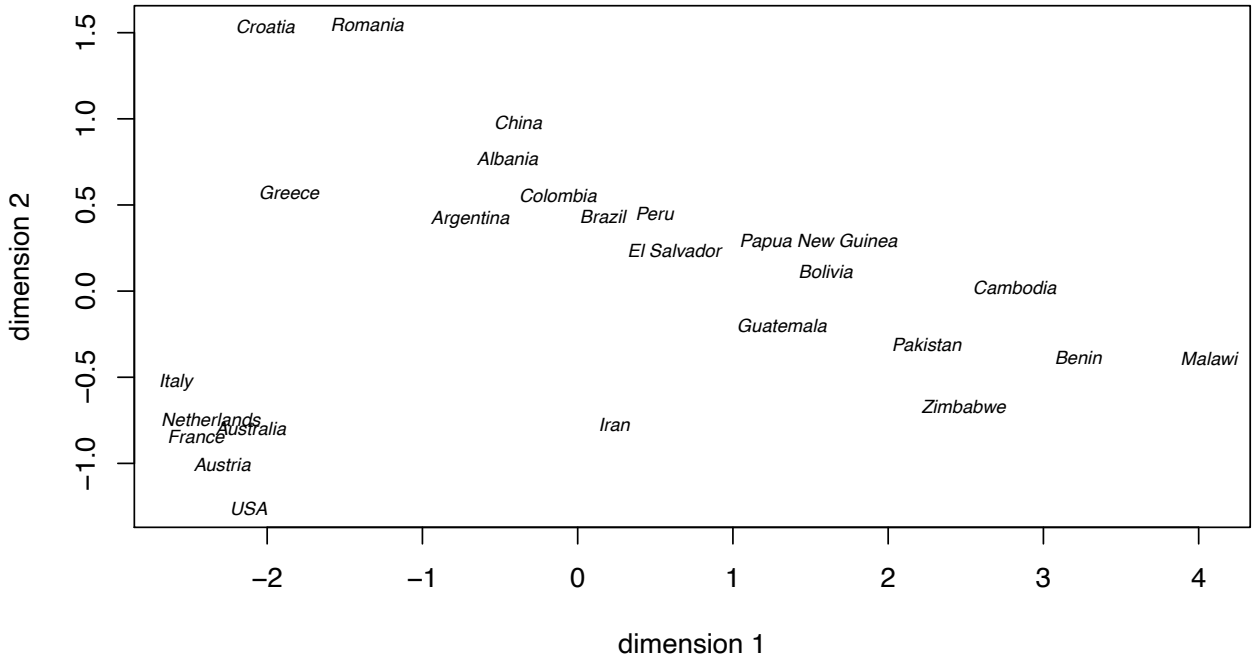


Figure 4: Configuration of countries from a two dimensional ratio MDS of economic and demographic indicators.

Table 8: Loadings associated with variables in Principla Component Analysis.

	PC1	PC2	PC3	PC4	PC5
Increase	0.43	0.52	0.64	-0.28	-0.24
Life	-0.47	0.05	0.48	-0.16	0.72
IMR	0.48	-0.02	-0.41	-0.63	0.45
TFR	0.47	0.24	-0.03	0.70	0.47
GDP	-0.38	0.82	-0.44	-0.03	-0.02

The second conclusion is that the goodness-of-fit of two-dimensional solution is much better than that of one-dimensional solution. Table 9 shows that two-dimensional solution is preferred as its stress value is closer to 0. The Shepard plot of two-dimensional solution also demonstrates a stronger linear relationship between the fitted values $\hat{d}_{ij}$ and the inter-point distances d_{ij} by comparison of Figure 5 (1-dimensional) and Figure 6 (2-dimensional).

(240 words)

Table 9: Stress value comparison for one dimension and two dimensions.

	One dimension	Two dimensions
Stress Value	0.12	0.03

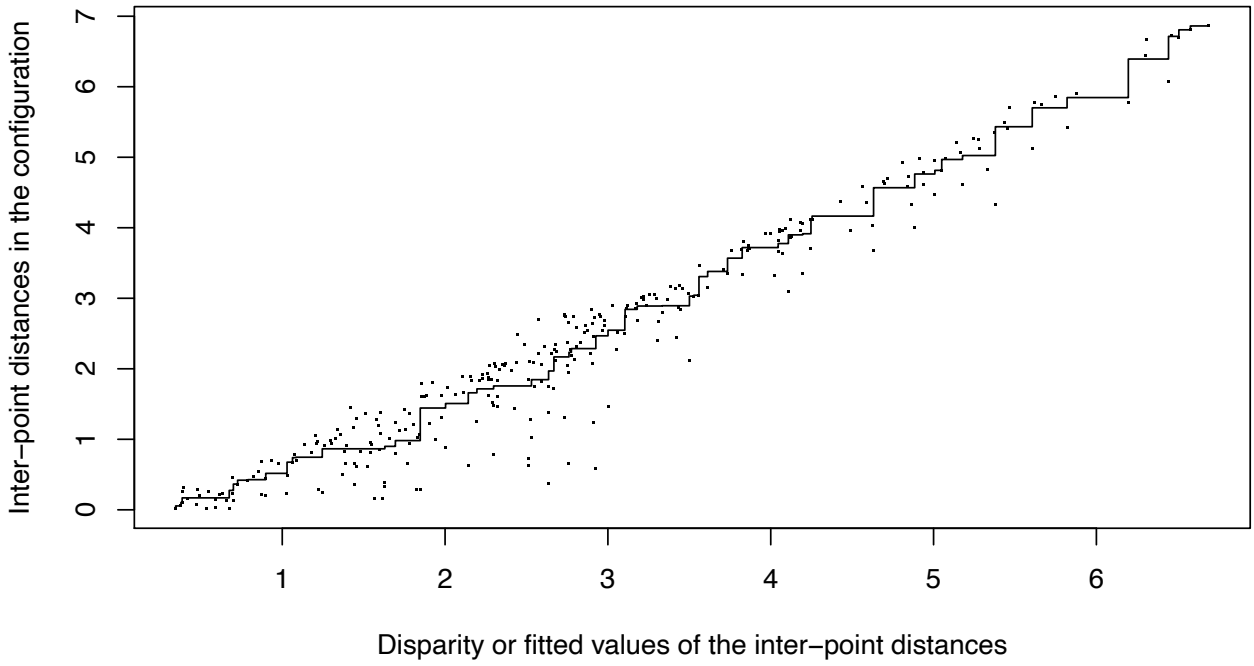


Figure 5: Shepard plot for one-dimensional solution.

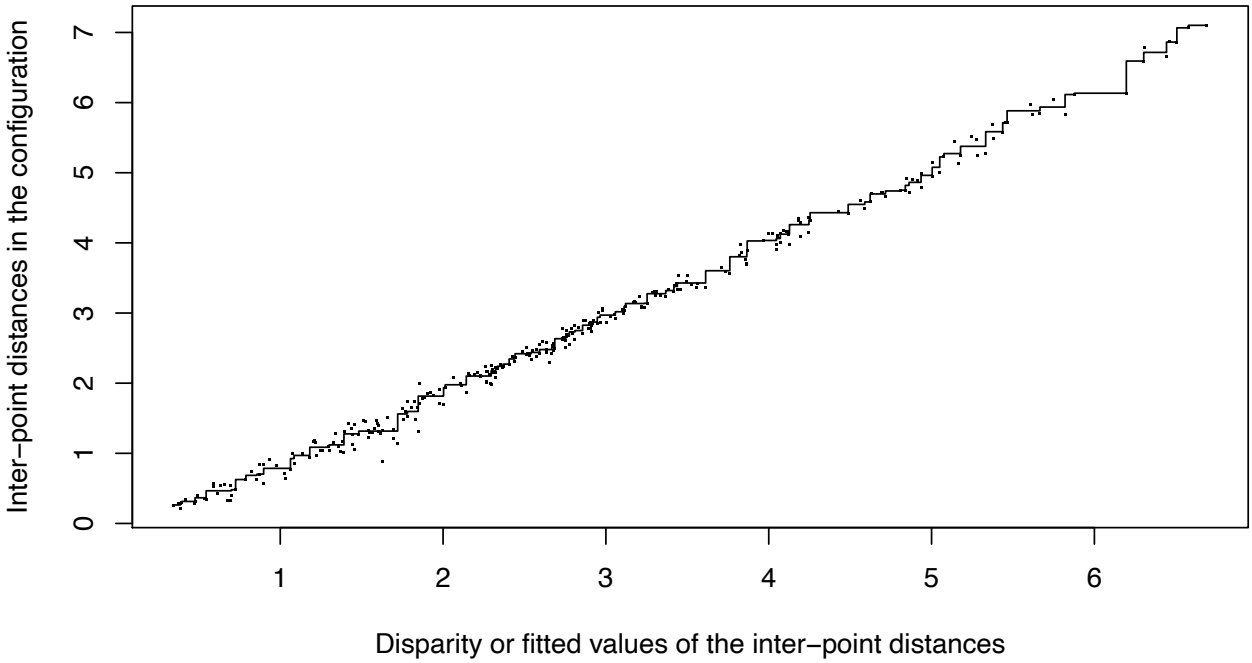


Figure 6: Shepard plot for two-dimensional solution.

3 Question 3

3.1 Purpose of study and methods adopted

The purpose of this study is to adopt an empirically aided instrument that measures online dating uses over previous instruments with vague psychometric properties using Exploratory Factor Analysis (EFA). The study uses an existing instrument (The Facebook Intensity Scale: FBI) that has been created to measure the intensity of Facebook uses among young adults, and creates the Online Intensity Scale (ODI). To test on a sample, 800 physical instrument packages were given to possible participants and 105 were invited to participate online. Moreover, Southeastern University's Psychology department had given 10,157 students the opportunity to participate through an online recruitment system. Finally, the authors were able to obtain a final sample of 494 confirmed candidates. Subsequently, there were 2 methods involved in data collection: online survey and face-to-face administration. Next, the authors started to clean their collected data by the first filtering for incomplete responses and adopted "listwise deletion" to delete such cases and create consistency. The data set was then scanned for any outliers (32) which were not removed, but rather kept in order to demonstrate a more accurate representation of the population. Moving on to the ODI, Facebook references were altered to online dating platforms, resulting in focused measurement of online dating in terms of frequency, quantity and duration, which led to a 5-point Likert-type scale with a 10-item instrument. Additionally, Marlowe-Crowne Social Desirability Scale-Short Form A (MCSDS-FA) was implemented to monitor participants' social desirability.

3.2 Results

The initial data cleaning process led the authors to discover a Kaiser-Meyer-Olkin (KMO) value equal to 0.819 and obtained significant results from Bartlett's test of sphericity with $p < 0.001$. Therefore, the authors were able to conduct an EFA. Additionally, a parallel analysis was performed to compare 100 random correlation matrices with eigenvalues. After further testing, the authors concluded the EFA to be performed on a 5-item instrument. The authors initially discovered factor loadings < 0.32, with scree plots supporting one- to three-factor solutions. Nevertheless, with actual cross-loadings and low communality issues, the

authors removed items 4,7,8,9 and 10 after individual assessment of their contribution. This was the most appropriate without harming the integrity of the study. A minimum eigenvalue of 1.12 was selected to maintain one factor and 1.05 for an additional factor based on the data set. In terms of the total variance explained, only one eigenvalue (3.04) met the criteria to maintain a single factor. The scree plot displayed similar results, with the first factor being the strongest followed by a sharp decline and then a steady plateau. Ultimately, the one-factor option was concluded to be the most appropriate for this instance, with more than 60% of the variance explained. This factor, which appeared to measure the degree of online dating uses, was determined to be consistent with the theoretical intention of the ODI.

3.3 Conclusion

This study proves the ODI to be a reliable instrument that can be used in further instances as it produces useful intensity values of the young adults. It is suggested that the ODI could be used in practice in fields such as mental health and family counselling due to the addictive nature of online platforms and how it affects such patients. Using the ODI, clinicians would be able to measure the progress of their clients towards intensity reduction of online dating activity. Essentially, the ODI could be a useful tool in tracking the improvements made in client therapy sessions or an effective method for those who wish to reduce the time spent on or to remove themselves from online dating. Overall, based on the young adult demographic, a strong instrument with the 5-item test is created to display appropriate psychometric attributes which could potentially be used for other demographic groups such as middle aged people and teenagers.

3.4 Comments

3.4.1 Robustness and generality of the results

Firstly, the assumptions necessary to conduct the exploratory factor analysis are satisfied by identifying a KMO value of 0.819 and the significant Bartlett's test of sphericity. The parallel analysis is also conducted to support the modified five-item scale. Secondly, the internal consistency reliability is examined and it is demonstrated that the model has appropriate levels of reliability with the data. Thirdly, bivariate correlations between the modified ODI

and the MCSDS-FA are conducted, and it is concluded that participants' responses were not influenced by social desirability. Finally, Spearman rank order correlations are conducted and it is revealed that statistically significant relationships between participants' scores and their demographic characteristics are not identified. Thus, it is concluded that the data collection process is rigorous to ensure heterogeneity in the sample.

Therefore, it is concluded from the robustness testing that the use of the modified five-item instrument with one factor as a measure that yields reliable scores of the online dating intensity, with principal axis factoring, oblimin rotation and Kaiser normalization, is consistent and sound with the data. The model, which demonstrates adequate psychometric properties in conjunction with emerging adult populations, can be generalizable to other populations such as adolescents and adults because of the brevity and readability of this five-item assessment. Additionally, the ODI scale, which is applied to the online dating context in this research, might be adapted to measure participants' other social media intensity.

3.4.2 Limitations of the analysis

There are several limitations arising in the research. First of all, the self-reporting biases are inherent in the data that were collected by Web-based survey and face-to-face administration, though one of the biases, social desirability, is accounted for by using MCSDS-FA to identify the social desirability aspect of the self-reported information. Secondly, there might be an artificial lower limit, below which the low-intensity level of online dating uses can't be measured, resulting in the possible floor or basement effect and a skewed data distribution. Thirdly, it is observed that there is an unequal distribution in terms of demographic characteristics, such as age, racial and ethnic background and sexual orientation. For example, more non-Hispanic, heterosexual and younger participants are represented in the data. This could result in the generality of the outcomes to all emerging adult college students. Similarly, the fact that data are mainly collected in the Southeastern US universities can also limit the generality of the analysis. Finally, the diversity of online dating uses is limited as most participants use Tinder as the preferred online dating applications. Overall, the limitations of the study can be addressed by including greater diversity in the emerging adult participants in terms of gender, age, race, ethnicity, sexual orientation and other online dating applications.