Multivariate Statistics

Assignment 1

Ana Sofia Mendes - r0925549
Ishika Jain - r0915387
Shreekar Araveti - r0919044
Sounak Ghosh - r0914328

Team_08

MSc Statistics and Data Science

November 2022

Task 1

a.

Step 1: On loading the data and computing centered variables, a CFA model with 3 correlated latent variables (`Att_organic`, `Att_packaging` and `Att_crueltyfree`) is fit, prior printing its fit measures and standardized solution. Then composite reliabilities are computed using the standardized solution.

```
#Confirmatory Factor Analysis
cfa1<- 'Att_organic=~NA*Attitude_organic1+Attitude_organic2+Attitude_organic3
Att_packaging=~NA*Attitude_packaging1+Attitude_packaging2+Attitude_packaging3
Att_crueltyfree=~NA*Attitude_crueltyfree1+Attitude_crueltyfree2+Attitude_crueltyfree3
Att_organic ~~1*Att_organic
Att_packaging ~~1*Att_packaging
Att_crueltyfree ~~1*Att_crueltyfree
Att_organic ~~Att_packaging
Att_packaging ~~Att_crueltyfree
Att_crueltyfree~~Att_organic'

#fit model on covariance matrix
fitcfa1<-cfa(cfa1,data=ccos, sample.cov=covmat,sample.nobs=150)

#summary of results
summary(fitcfa1, fit.measures=TRUE)

#print fit measures
fitmeasures(fitcfa1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

  chisq      df   pvalue     gfi     agfi     cfi     tli    rmsea    srmr
120.886  24.000   0.000    0.853   0.724   0.889   0.833   0.164   0.057
```

The `fit measures` indicate that the model is rejected by an absolute goodness of fit test, i.e. the fit of the model is significantly lower than for a perfectly fitting model (chi-square=120.886, df=24, p< 0.01). However, as the model is fitted on a rather considerable number of observations (N=150) the chi-square test is very sensitive and has high statistical power to detect a small deviation from the null hypothesis. Then, it is more appropriate to rely on descriptive fit measures. The printed descriptive measures indicate that SRMR (0.057) meets the cutoff for a good fit (SRMR < 0.08); however, CFI (0.889), TLI (0.833), GFI (0.853), AGFI (0.724) and RMSEA (0.164) do not meet the cutoff of good fit (CFI< 0.95, TLI< 0.95, GFI< 0.95, AGFI< 0.90 and RMSEA> 0.08).

Next we look for the standardized solution:

```
#ask for standardized solution
standardizedSolution(fitcfa1)

                     lhs op                   rhs est.std    se      z pvalue ci.lower ci.upper
1            Att_organic =~     Attitude_organic1   0.871 0.036 24.461      0    0.801    0.941
2            Att_organic =~     Attitude_organic2   0.726 0.048 15.272      0    0.633    0.819
3            Att_organic =~     Attitude_organic3   0.718 0.048 14.856      0    0.623    0.812
4          Att_packaging =~   Attitude_packaging1   0.843 0.033 25.698      0    0.778    0.907
5          Att_packaging =~   Attitude_packaging2   0.795 0.038 21.079      0    0.721    0.869
6          Att_packaging =~   Attitude_packaging3   0.803 0.037 21.862      0    0.731    0.876
7        Att_crueltyfree =~ Attitude_crueltyfree1   0.913 0.023 39.019      0    0.867    0.959
8        Att_crueltyfree =~ Attitude_crueltyfree2   0.790 0.036 22.100      0    0.720    0.860
9        Att_crueltyfree =~ Attitude_crueltyfree3   0.864 0.028 31.121      0    0.810    0.919
10           Att_organic ~~           Att_organic   1.000 0.000     NA     NA    1.000    1.000
11         Att_packaging ~~         Att_packaging   1.000 0.000     NA     NA    1.000    1.000
12       Att_crueltyfree ~~       Att_crueltyfree   1.000 0.000     NA     NA    1.000    1.000
13           Att_organic ~~         Att_packaging   0.739 0.054 13.756      0    0.634    0.845
14         Att_packaging ~~       Att_crueltyfree   0.725 0.051 14.242      0    0.625    0.825
15           Att_organic ~~       Att_crueltyfree   0.603 0.065  9.311      0    0.476    0.730
16     Attitude_organic1 ~~     Attitude_organic1   0.241 0.062  3.880      0    0.119    0.362
17     Attitude_organic2 ~~     Attitude_organic2   0.473 0.069  6.855      0    0.338    0.608
18     Attitude_organic3 ~~     Attitude_organic3   0.485 0.069  6.990      0    0.349    0.621
19   Attitude_packaging1 ~~   Attitude_packaging1   0.290 0.055  5.252      0    0.182    0.398
20   Attitude_packaging2 ~~   Attitude_packaging2   0.369 0.060  6.151      0    0.251    0.486
21   Attitude_packaging3 ~~   Attitude_packaging3   0.354 0.059  6.000      0    0.239    0.470
22 Attitude_crueltyfree1 ~~ Attitude_crueltyfree1   0.167 0.043  3.901      0    0.083    0.250
23 Attitude_crueltyfree2 ~~ Attitude_crueltyfree2   0.375 0.057  6.638      0    0.264    0.486
24 Attitude_crueltyfree3 ~~ Attitude_crueltyfree3   0.253 0.048  5.275      0    0.159    0.347
```

As seen in the standardized solution, all variables have significant and positive standardized loadings that exceed 0.7 (i.e. variables have a significant positive correlation with the corresponding factor). Hence, the variables have sufficient reliability and `convergent validity` is satisfied for the measurement model.

Furthermore, **`discriminant validity`** is also satisfied as the correlations between the latent factors are all significantly smaller than 1 (that can be concluded by observing that the value 1 is not in the 95% CI, and correlations are assumed to be significantly below 1) . Note that there are two rather strong correlations: between the factors “Att_organic” and “Att_packaging” (0.739), and between ”Att_packaging” and ”Att_crueltyfree” (0.725).

Finally, the composite reliability of all factor scores is good as it exceeds 0.80:

```
#function composite reliability
comp_rel<-function(x){
  A<-(sum(x))^2
  B<-sum(1-x^2)
  return(A/(A+B))
}

#Overview of composite reliability
factorscore<-c("Att_Organic","Att_packaging","Att_crueltyfree")
reliability<-round(c(comp_rel(d[1:3,4]),comp_rel(d[4:6,4]),comp_rel(d[7:9,4])),3)
data.frame(factorscore,reliability)

      factorscore reliability
1      Att_Organic       0.817
2    Att_packaging       0.855
3  Att_crueltyfree       0.892
```

Step 2: To further improve the model, a constraint of **equal residual covariances** between pairs of items that focus on the same aspect is imposed:

```
#Confirmatory Factor Analysis
cfa2<- 'Att_organic=~NA*Attitude_organic1+Attitude_organic2+Attitude_organic3
Att_packaging=~NA*Attitude_packaging1+Attitude_packaging2+Attitude_packaging3
Att_crueltyfree=~NA*Attitude_crueltyfree1+Attitude_crueltyfree2+Attitude_crueltyfree3
Att_organic ~~1*Att_organic
Att_packaging ~~1*Att_packaging
Att_crueltyfree ~~1*Att_crueltyfree
Att_organic ~~Att_packaging
Att_packaging ~~Att_crueltyfree
Att_crueltyfree~~Att_organic
Attitude_organic1  ~~c*Attitude_packaging1
Attitude_organic1  ~~c*Attitude_crueltyfree1
Attitude_crueltyfree1 ~~c*Attitude_packaging1
Attitude_organic2 ~~d*Attitude_packaging2
Attitude_organic2 ~~d*Attitude_crueltyfree2
Attitude_crueltyfree2 ~~d*Attitude_packaging2
Attitude_organic3 ~~e*Attitude_packaging3
Attitude_organic3 ~~e*Attitude_crueltyfree3
Attitude_crueltyfree3 ~~e*Attitude_packaging3'


#fit model on covariance matrix
fitcfa2<-cfa(cfa2,data=ccos, sample.cov=covmat,sample.nobs=150)

#summary of results
summary(fitcfa2, fit.measures=TRUE)

#print fit measures
fitmeasures(fitcfa2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

 chisq     df pvalue    gfi   agfi    cfi    tli  rmsea   srmr
56.736 21.000  0.000  0.922  0.833  0.959  0.930  0.107  0.042
```

Results indicate that even the new model falls short of fitting the data well (chi-square=56.736, df=21, $p< 0.01$), even though the chi-squared value is reduced to half of the previously obtained value. Therefore, the model is rejected by a goodness of fit test. Printed descriptive measures indicate that the model does not meet all the cutoff criteria for all measures, i.e. TLI (0.930) and RMSEA (0.107) (TLI< 0.95 and RMSEA> 0.08), however obtained values are close to the critical value.

As all criteria is close to showing a good fit, and the new model is still parsimonious and has a simple structure, we do not make further modifications.

Next we look for the standardized solution:

```
#print standardized solution
standardizedSolution(fitcfa2)
                  lhs op                 rhs label est.std    se      z pvalue ci.lower ci.upper
1          Att_organic =~    Attitude_organic1         0.887 0.038 23.249  0.000    0.812    0.962
2          Att_organic =~    Attitude_organic2         0.727 0.047 15.591  0.000    0.636    0.819
3          Att_organic =~    Attitude_organic3         0.718 0.047 15.195  0.000    0.626    0.811
4        Att_packaging =~  Attitude_packaging1         0.865 0.033 26.279  0.000    0.801    0.930
5        Att_packaging =~  Attitude_packaging2         0.798 0.037 21.582  0.000    0.725    0.870
6        Att_packaging =~  Attitude_packaging3         0.800 0.036 21.990  0.000    0.729    0.872
7      Att_crueltyfree =~ Attitude_crueltyfree1        0.926 0.026 35.402  0.000    0.874    0.977
8      Att_crueltyfree =~ Attitude_crueltyfree2        0.773 0.037 20.666  0.000    0.700    0.846
9      Att_crueltyfree =~ Attitude_crueltyfree3        0.833 0.032 26.201  0.000    0.771    0.896
10         Att_organic ~~          Att_organic         1.000 0.000     NA     NA    1.000    1.000
11       Att_packaging ~~        Att_packaging         1.000 0.000     NA     NA    1.000    1.000
12     Att_crueltyfree ~~      Att_crueltyfree         1.000 0.000     NA     NA    1.000    1.000
13         Att_organic ~~        Att_packaging         0.691 0.055 12.528  0.000    0.583    0.799
14       Att_packaging ~~      Att_crueltyfree         0.689 0.052 13.161  0.000    0.587    0.792
15         Att_organic ~~      Att_crueltyfree         0.570 0.066  8.636  0.000    0.441    0.699
16   Attitude_organic1 ~~  Attitude_packaging1     c   0.058 0.119  0.491  0.624   −0.175    0.292
17   Attitude_organic1 ~~ Attitude_crueltyfree1     c   0.076 0.152  0.497  0.619   −0.222    0.374
18 Attitude_packaging1 ~~ Attitude_crueltyfree1     c   0.064 0.130  0.492  0.623   −0.191    0.319
19   Attitude_organic2 ~~  Attitude_packaging2     d   0.362 0.072  5.049  0.000    0.222    0.503
20   Attitude_organic2 ~~ Attitude_crueltyfree2     d   0.282 0.059  4.774  0.000    0.166    0.397
21 Attitude_packaging2 ~~ Attitude_crueltyfree2     d   0.330 0.067  4.929  0.000    0.199    0.461
22   Attitude_organic3 ~~  Attitude_packaging3     e   0.328 0.066  4.940  0.000    0.198    0.458
23   Attitude_organic3 ~~ Attitude_crueltyfree3     e   0.343 0.069  4.990  0.000    0.208    0.477
24 Attitude_packaging3 ~~ Attitude_crueltyfree3     e   0.367 0.072  5.092  0.000    0.226    0.508
25   Attitude_organic1 ~~    Attitude_organic1         0.214 0.068  3.158  0.002    0.081    0.346
26   Attitude_organic2 ~~    Attitude_organic2         0.471 0.068  6.942  0.000    0.338    0.604
27   Attitude_organic3 ~~    Attitude_organic3         0.484 0.068  7.131  0.000    0.351    0.617
28 Attitude_packaging1 ~~  Attitude_packaging1         0.251 0.057  4.413  0.000    0.140    0.363
29 Attitude_packaging2 ~~  Attitude_packaging2         0.363 0.059  6.160  0.000    0.248    0.479
30 Attitude_packaging3 ~~  Attitude_packaging3         0.360 0.058  6.171  0.000    0.245    0.474
31 Attitude_crueltyfree1 ~~ Attitude_crueltyfree1       0.143 0.048  2.960  0.003    0.048    0.238
32 Attitude_crueltyfree2 ~~ Attitude_crueltyfree2       0.403 0.058  6.966  0.000    0.289    0.516
33 Attitude_crueltyfree3 ~~ Attitude_crueltyfree3       0.306 0.053  5.767  0.000    0.202    0.410
```

As can be seen in the standardized solution, all variables have significant and positive standardized loadings above 0.7 (variables have a significant positive correlation with the corresponding factor). Since the variables have sufficient reliability, `convergent validity` is satisfied for the measurement model.

Furthermore, `divergent validity` is also satisfied as the latent variables have moderate correlations that are significantly smaller than 1. We notice two rather strong correlations: between the factors "Att_organic" and "Att_packaging" (0.691), and between "Att_packaging" and "Att_crueltyfree" (0.689). However, the covariances added between the pairs that focus on the aspect 1 ("right thing to do", in lines 16 to 18) show a high p-value, implying that they aren't significant. Therefore, we shouldn't impose the equal constraint `c`, and in order to improve the model, it should be removed.

Finally, the composite reliability of all factor scores is good as it exceeds 0.80:

```
#Overview of composite reliability
factorscore<-c("Att_Organic","Att_packaging","Att_crueltyfree")
reliability2<-round(c(comp_rel(d2[1:3,5]),comp_rel(d2[4:6,5]),comp_rel(d2[7:9,5])),3)
data.frame(factorscore,reliability2)

      factorscore reliability2
1     Att_Organic        0.823
2   Att_packaging        0.862
3 Att_crueltyfree        0.883
```

Comparison: To compare the models obtained in step 1 and 2 we will compare the fit measures:

```
#comparing fit
fitmeasures1=fitmeasures(fitcfa1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fitmeasures2=fitmeasures(fitcfa2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fit1<-rbind(fitmeasures1,fitmeasures2)
rownames(fit1)<-c("cfa_model_Att","cfa_extended_model_Att")
chidf<-fit1[,1]/fit1[,2]

fit1<-cbind(fit1,chidf)
round(fit1, 3)
                         chisq df pvalue   gfi  agfi   cfi   tli rmsea  srmr chidf
cfa model Att          120.886 24      0 0.853 0.724 0.889 0.833 0.164 0.057 5.037
cfa extended model Att  56.736 21      0 0.922 0.833 0.959 0.930 0.107 0.042 2.702
```

By imposing the constraint of equal residual covariances, we see an improvement in the fit of the model - even though the extended model is still rejected - all the fit measures of the model in step 2 have better values than the ones obtained for the model in step 1.

b.

Step 1: The procedure is similar to what was previously done in question a. (this time applied to columns 10-18 in Table 1).

After loading the data, we compute centered variables. We fit a CFA model with 3 correlated latent variables (`BI_organic`, `BI_packaging`, and `BI_crueltyfree`), and print the fit measures and the standardized solution. We also compute the composite reliabilities using the standardized solution.

```
cfab1<- 'BI_organic=~NA*BI_organic1+BI_organic2+BI_organic3
BI_packaging=~NA*BI_packaging1+BI_packaging2+BI_packaging3
BI_crueltyfree=~NA*BI_crueltyfree1+BI_crueltyfree2+BI_crueltyfree3
BI_organic ~~1*BI_organic
BI_packaging ~~1*BI_packaging
BI_crueltyfree ~~1*BI_crueltyfree
BI_organic ~~BI_packaging
BI_packaging ~~BI_crueltyfree
BI_crueltyfree ~~BI_organic'

#fit model on covariance matrix
fitcfab1<-cfa(cfab1,data=ccos, sample.cov=covmat,sample.nobs=150)

#summary of results
summary(fitcfab1,fit.measures=TRUE)

#print fit measures
fitmeasures(fitcfab1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

  chisq      df  pvalue     gfi    agfi     cfi     tli   rmsea    srmr
147.814  24.000   0.000   0.811   0.646   0.914   0.871   0.185   0.033
```

The `fit measures` indicate that the model is rejected by an absolute goodness of fit test, i.e. the fit of the model is significantly lower than for a perfectly fitting model (chi-square=147.814, df=24, p< 0.01). However, as the model is fitted on a rather considerable number of observations (N=150), the chi-square test is very sensitive and has high statistical power to detect a small deviation from the null hypothesis. Then, it is better to rely on descriptive fit measures. The printed descriptive measures indicate that only SRMR (0.033) meets the cutoff for good fit (SRMR< 0.08); the other measures, CFI (0.914), TLI (0.871), GFI (0.811), AGFI (0.646) and RMSEA (0.185) do not meet the cutoff of good fit (CFI< 0.95, TLI< 0.95, GFI< 0.95, AGFI< 0.90 and RMSEA> 0.08).

Next we look for the standardized solution:

```
#ask for standardized solution
standardizedSolution(fitcfab1)

               lhs op             rhs est.std    se      z pvalue ci.lower ci.upper
1       BI_organic =~     BI_organic1   0.886 0.023 39.149      0    0.841    0.930
2       BI_organic =~     BI_organic2   0.897 0.021 41.980      0    0.855    0.939
3       BI_organic =~     BI_organic3   0.843 0.028 30.204      0    0.788    0.897
4     BI_packaging =~   BI_packaging1   0.875 0.023 37.407      0    0.829    0.921
5     BI_packaging =~   BI_packaging2   0.892 0.021 41.621      0    0.850    0.934
6     BI_packaging =~   BI_packaging3   0.866 0.025 35.243      0    0.818    0.914
7   BI_crueltyfree =~ BI_crueltyfree1   0.916 0.016 55.816      0    0.884    0.948
8   BI_crueltyfree =~ BI_crueltyfree2   0.918 0.016 56.707      0    0.886    0.949
9   BI_crueltyfree =~ BI_crueltyfree3   0.939 0.014 68.617      0    0.912    0.966
10      BI_organic ~~      BI_organic   1.000 0.000     NA     NA    1.000    1.000
11    BI_packaging ~~    BI_packaging   1.000 0.000     NA     NA    1.000    1.000
12  BI_crueltyfree ~~  BI_crueltyfree   1.000 0.000     NA     NA    1.000    1.000
13      BI_organic ~~    BI_packaging   0.876 0.028 30.822      0    0.820    0.932
14    BI_packaging ~~  BI_crueltyfree   0.832 0.032 25.983      0    0.770    0.895
15      BI_organic ~~  BI_crueltyfree   0.784 0.038 20.551      0    0.710    0.859
16     BI_organic1 ~~     BI_organic1   0.215 0.040  5.374      0    0.137    0.294
17     BI_organic2 ~~     BI_organic2   0.196 0.038  5.109      0    0.121    0.271
18     BI_organic3 ~~     BI_organic3   0.290 0.047  6.169      0    0.198    0.382
19   BI_packaging1 ~~   BI_packaging1   0.234 0.041  5.707      0    0.154    0.314
20   BI_packaging2 ~~   BI_packaging2   0.205 0.038  5.370      0    0.130    0.280
```

```
21    BI_packaging3 ~~    BI_packaging3   0.250 0.043   5.877       0     0.167     0.334
22 BI_crueltyfree1 ~~ BI_crueltyfree1   0.161 0.030   5.367       0     0.102     0.220
23 BI_crueltyfree2 ~~ BI_crueltyfree2   0.158 0.030   5.319       0     0.100     0.216
24 BI_crueltyfree3 ~~ BI_crueltyfree3   0.118 0.026   4.607       0     0.068     0.169
```

It can be seen from the standardized solution that all variables have significant and positive standardized loadings that exceed 0.7 (variables have a significant positive correlation with the corresponding factor). Hence, the variables have sufficient reliability so that `convergent validity` is satisfied for the measurement model.

Furthermore, `discriminant validity` is also satisfied as the correlations between the latent factors are all significantly smaller than 1. Note that there are three rather strong correlations: between the factors "BI_organic" and "BI_packaging" (0.876), between "BI_packaging" and "BI_crueltyfree" (0.832), and between "BI_organic" and "BI_crueltyfree" (0.784).

Finally, the composite reliability of all factor scores is excellent as all values exceed 0.80, being in fact all above 0.90:

```
#Overview of composite reliability
factorscoreb<-c("BI_Organic","BI_packaging","BI_crueltyfree")
reliabilityb<-round(c(comp_rel(e[1:3,4]),comp_rel(e[4:6,4]),comp_rel(e[7:9,4])),3)
data.frame(factorscoreb,reliabilityb)

    factorscoreb reliabilityb
1     BI_Organic        0.908
2   BI_packaging        0.910
3 BI_crueltyfree        0.946
```

Step 2: Once again, in order to improve the model, we will extend the model by imposing the constraint of **equal residual covariances** for all pairs of Behavior-Intention items that focus on the same aspect.

```
 cfab2<- 'BI_organic=~NA*BI_organic1+BI_organic2+BI_organic3
BI_packaging=~NA*BI_packaging1+BI_packaging2+BI_packaging3
BI_crueltyfree=~NA*BI_crueltyfree1+BI_crueltyfree2+BI_crueltyfree3
BI_organic ~~1*BI_organic
BI_packaging ~~1*BI_packaging
BI_crueltyfree ~~1*BI_crueltyfree
BI_organic ~~BI_packaging
BI_packaging ~~BI_crueltyfree
BI_crueltyfree ~~BI_organic
BI_organic1  ~~c*BI_packaging1
BI_organic1  ~~c*BI_crueltyfree1
BI_crueltyfree1 ~~c*BI_packaging1
BI_organic2 ~~d*BI_packaging2
BI_organic2 ~~d*BI_crueltyfree2
BI_crueltyfree2 ~~d*BI_packaging2
BI_organic3 ~~e*BI_packaging3
BI_organic3 ~~e*BI_crueltyfree3
BI_crueltyfree3 ~~e*BI_packaging3'

#fit model on covariance matrix
fitcfab2<-cfa(cfab2,data=cosmetics, sample.cov=covmat,sample.nobs=150)

#summary of results
summary(fitcfab2, fit.measures=TRUE
)
#print fit measures
fitmeasures(fitcfab2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

 chisq     df pvalue    gfi   agfi    cfi    tli  rmsea   srmr
26.779 21.000  0.178  0.961  0.916  0.996  0.993  0.043  0.020
```

The results of chi-square, goodness of fit test indicate that the model fits the data well (chi-square=26.779, df=21, p=0.178). Besides, it has excellent descriptive goodness of fit, as can be seen from the `fit measures`: GFI=0.961, AGFI=0.916, CFI=0.996, TLI=0.993, RMSEA=0.043, and SRMR=0.020 (all of them meet the requirements of good fit).

Next we look for the standardized solution:

```
#ask for standardized solution
standardizedSolution(fitcfab2)
```

```
              lhs op             rhs label est.std    se      z pvalue ci.lower ci.upper
1      BI_organic =~     BI_organic1         0.885 0.023 38.303  0.000    0.840    0.930
2      BI_organic =~     BI_organic2         0.886 0.023 39.317  0.000    0.841    0.930
3      BI_organic =~     BI_organic3         0.853 0.027 31.715  0.000    0.800    0.906
4    BI_packaging =~   BI_packaging1         0.876 0.024 36.773  0.000    0.829    0.922
5    BI_packaging =~   BI_packaging2         0.896 0.021 42.344  0.000    0.855    0.938
6    BI_packaging =~   BI_packaging3         0.852 0.027 31.910  0.000    0.800    0.905
7   BI_crueltyfree =~ BI_crueltyfree1         0.921 0.016 58.144  0.000    0.890    0.952
8   BI_crueltyfree =~ BI_crueltyfree2         0.916 0.016 57.023  0.000    0.885    0.948
9   BI_crueltyfree =~ BI_crueltyfree3         0.941 0.014 67.666  0.000    0.913    0.968
10     BI_organic ~~      BI_organic         1.000 0.000     NA     NA    1.000    1.000
11   BI_packaging ~~    BI_packaging         1.000 0.000     NA     NA    1.000    1.000
12  BI_crueltyfree ~~  BI_crueltyfree         1.000 0.000     NA     NA    1.000    1.000
13     BI_organic ~~    BI_packaging         0.841 0.030 28.067  0.000    0.782    0.900
14   BI_packaging ~~  BI_crueltyfree         0.806 0.033 24.445  0.000    0.742    0.871
15     BI_organic ~~  BI_crueltyfree         0.753 0.040 18.827  0.000    0.675    0.832
16    BI_organic1 ~~   BI_packaging1     c   0.317 0.074  4.280  0.000    0.172    0.462
17    BI_organic1 ~~ BI_crueltyfree1     c   0.357 0.081  4.395  0.000    0.198    0.516
18  BI_packaging1 ~~ BI_crueltyfree1     c   0.361 0.082  4.425  0.000    0.201    0.520
19    BI_organic2 ~~   BI_packaging2     d   0.505 0.072  6.974  0.000    0.363    0.647
20    BI_organic2 ~~ BI_crueltyfree2     d   0.507 0.073  6.936  0.000    0.364    0.651
21  BI_packaging2 ~~ BI_crueltyfree2     d   0.538 0.074  7.241  0.000    0.392    0.683
22    BI_organic3 ~~   BI_packaging3     e   0.223 0.065  3.434  0.001    0.096    0.350
23    BI_organic3 ~~ BI_crueltyfree3     e   0.314 0.085  3.689  0.000    0.147    0.481
24  BI_packaging3 ~~ BI_crueltyfree3     e   0.323 0.087  3.723  0.000    0.153    0.493
25    BI_organic1 ~~     BI_organic1         0.217 0.041  5.313  0.000    0.137    0.297
26    BI_organic2 ~~     BI_organic2         0.216 0.040  5.407  0.000    0.138    0.294
27    BI_organic3 ~~     BI_organic3         0.273 0.046  5.947  0.000    0.183    0.363
28  BI_packaging1 ~~   BI_packaging1         0.233 0.042  5.591  0.000    0.151    0.315
29  BI_packaging2 ~~   BI_packaging2         0.197 0.038  5.189  0.000    0.122    0.271
30  BI_packaging3 ~~   BI_packaging3         0.273 0.046  6.001  0.000    0.184    0.363
31 BI_crueltyfree1 ~~ BI_crueltyfree1         0.152 0.029  5.229  0.000    0.095    0.210
32 BI_crueltyfree2 ~~ BI_crueltyfree2         0.161 0.029  5.457  0.000    0.103    0.218
33 BI_crueltyfree3 ~~ BI_crueltyfree3         0.115 0.026  4.415  0.000    0.064    0.167
```

As seen from the standardized solution, all variables have significant and positive standardized loadings that exceed 0.7 (variables have a significant positive correlation with the corresponding factor). In fact, all the standardized loadings are above 0.90. Hence, the variables have sufficient reliability, and `convergent validity` is satisfied for the measurement model.

Furthermore, `divergent validity` is also satisfied as all latent variables have moderate correlations that are significantly smaller than 1. Note that there are three rather strong correlations: between the factors “BI_organic” and “BI_packaging” (0.841), between ”BI_packaging” and ”BI_crueltyfree” (0.806), and between “BI_organic” and ”BI_crueltyfree” (0.753).

Finally, the composite reliability of all factor scores is good as it exceeds 0.80 (being all values around 0.90):

```
#Overview of composite reliability
factorscoreb<-c("BI_Organic","BI_packaging","BI_crueltyfree")
reliabilityb2<-round(c(comp_rel(e2[1:3,5]),comp_rel(e2[4:6,5]),comp_rel(e2[7:9,5])),3)
data.frame(factorscoreb, reliabilityb2)

    factorscoreb reliabilityb2
1     BI_Organic         0.907
2   BI_packaging         0.907
3 BI_crueltyfree         0.947
```

Comparison: To compare the models obtained in step 1 and 2 we will compare the fit measures:

```
#comparing fit
fitmeasuresb1=fitmeasures(fitcfab1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fitmeasuresb2=fitmeasures(fitcfab2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fit2<-rbind(fitmeasuresb1,fitmeasuresb2)
rownames(fit2)<-c("cfa_model_BI","cfa_extended_model_BI")
chidf<-fit2[,1]/fit2[,2]

fit2<-cbind(fit2,chidf)
round(fit2, 3)
                         chisq df pvalue   gfi  agfi   cfi   tli rmsea  srmr chidf
cfa model BI           147.814 24  0.000 0.811 0.646 0.914 0.871 0.185 0.033 6.159
cfa extended model BI   26.779 21  0.178 0.961 0.916 0.996 0.993 0.043 0.020 1.275
```

By imposing the constraint of equal residual covariances, we see an improvement in the model fit - all the fit measures of the model in step 2 have a better value as compared to those in step 1. With this modification, all the measures meet the cutoff criteria, and the extended model fits the data well.

c.

Step 1: The `sem()` function is used to fit the structural equation model on the covariance matrix, and print fit measures and model output (including the standardized solution). Combined measurement models in step 2 of questions `a.` and `b.` will be used:

```
#specify structural equation model
sem1<-'#measurement model
Att_organic=~NA*Attitude_organic1+Attitude_organic2+Attitude_organic3
Att_packaging=~NA*Attitude_packaging1+Attitude_packaging2+Attitude_packaging3
Att_crueltyfree=~NA*Attitude_crueltyfree1+Attitude_crueltyfree2+Attitude_crueltyfree3
Att_organic ~~Att_packaging
Att_packaging ~~Att_crueltyfree
Att_crueltyfree~~Att_organic
Attitude_organic1  ~~c*Attitude_packaging1
Attitude_organic1  ~~c*Attitude_crueltyfree1
Attitude_crueltyfree1 ~~c*Attitude_packaging1
Attitude_organic2 ~~d*Attitude_packaging2
Attitude_organic2 ~~d*Attitude_crueltyfree2
Attitude_crueltyfree2 ~~d*Attitude_packaging2
Attitude_organic3 ~~e*Attitude_packaging3
Attitude_organic3 ~~e*Attitude_crueltyfree3
Attitude_crueltyfree3 ~~e*Attitude_packaging3

BI_organic=~1*BI_organic1+BI_organic2+BI_organic3
BI_packaging=~1*BI_packaging1+BI_packaging2+BI_packaging3
BI_crueltyfree=~1*BI_crueltyfree1+BI_crueltyfree2+BI_crueltyfree3
BI_organic ~~BI_packaging
BI_packaging ~~BI_crueltyfree
BI_crueltyfree ~~BI_organic
BI_organic1  ~~f*BI_packaging1
BI_organic1  ~~f*BI_crueltyfree1
BI_crueltyfree1 ~~f*BI_packaging1
BI_organic2 ~~g*BI_packaging2
BI_organic2 ~~g*BI_crueltyfree2
BI_crueltyfree2 ~~g*BI_packaging2
BI_organic3 ~~h*BI_packaging3
BI_organic3 ~~h*BI_crueltyfree3
BI_crueltyfree3 ~~h*BI_packaging3

#structural model
BI_organic ~Att_organic
BI_packaging~ Att_packaging
BI_crueltyfree~ Att_crueltyfree

#variances latent variables
Att_organic ~~1*Att_organic
Att_packaging ~~1*Att_packaging
Att_crueltyfree ~~1*Att_crueltyfree
BI_organic ~~BI_organic
BI_packaging ~~BI_packaging
BI_crueltyfree ~~BI_crueltyfree'


#print fit measures
fitmeasures(fitsem1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

  chisq      df  pvalue     gfi    agfi     cfi     tli   rmsea    srmr
167.696 120.000   0.003   0.893   0.847   0.981   0.976   0.051   0.085


#print model output
fitsem1<-sem(sem1,ccos)
summary(fitsem1,std=TRUE)

lavaan 0.6-12 ended normally after 65 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        63
  Number of equality constraints                    12

  Number of observations                           150
```

```
Model Test User Model:

  Test statistic                               167.696
  Degrees of freedom                               120
  P-value (Chi-square)                           0.003

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  Att_organic =~
    Attitude_rgnc1     0.719    0.059   12.277    0.000    0.719    0.857
    Attitude_rgnc2     0.612    0.062    9.879    0.000    0.612    0.723
    Attitude_rgnc3     0.760    0.074   10.232    0.000    0.760    0.741
  Att_packaging =~
    Attitd_pckgng1     0.764    0.063   12.188    0.000    0.764    0.841
    Attitd_pckgng2     0.655    0.058   11.356    0.000    0.655    0.791
    Attitd_pckgng3     0.916    0.075   12.144    0.000    0.916    0.827
  Att_crueltyfree =~
    Atttd_crltyfr1     0.847    0.061   13.869    0.000    0.847    0.904
    Atttd_crltyfr2     0.801    0.069   11.526    0.000    0.801    0.794
    Atttd_crltyfr3     0.980    0.076   12.867    0.000    0.980    0.853
  BI_organic =~
    BI_organic1        1.000                               0.914    0.873
    BI_organic2        0.967    0.063   15.340    0.000    0.884    0.881
    BI_organic3        0.915    0.067   13.732    0.000    0.836    0.842
  BI_packaging =~
    BI_packaging1      1.000                               0.859    0.868
    BI_packaging2      1.012    0.065   15.542    0.000    0.870    0.888
    BI_packaging3      0.927    0.069   13.434    0.000    0.796    0.834
  BI_crueltyfree =~
    BI_crueltyfre1     1.000                               0.961    0.913
    BI_crueltyfre2     0.985    0.052   19.036    0.000    0.946    0.901
    BI_crueltyfre3     0.973    0.049   19.804    0.000    0.935    0.929

Regressions:
                Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  BI_organic ~
    Att_organic      0.619    0.067    9.195    0.000    0.677    0.677
  BI_packaging ~
    Att_packaging    0.591    0.062    9.585    0.000    0.689    0.689
  BI_crueltyfree ~
    Att_crueltyfre   0.685    0.066   10.347    0.000    0.713    0.713


[...]
```

As indicated by the `fit measures`, the model is rejected by an absolute goodness of fit test (chi-square=167.696, df=120, p< 0.05). This is be expected since the test is very sensitive due to the large sample size. The printed descriptive measures indicate only that CFI (0.981), TLI (0.976) and RMSEA (0.051) meets the cutoff for good fit (CFI> 0.95, TLI> 0.95, RMSEA< 0.08); the others, GFI (0.893), AGFI (0.847), and SRMR (0.085) do not meet the cutoff of good fit (GFI< 0.95, AGFI< 0.90, SRMR> 0.08).

The results of the measurement model are rather similar as for the CFA model. All variables have positive and significant loadings, and all have a standardized loading that exceeds 0.7, which means that they have sufficient reliability. The standardized regression coefficients (which are partial correlations) indicate that the effects are strong. For instance, after controlling for other variables, if `Att_organic` increases one SD, `BI_organic` increases by 0.677 SDs.

Looking at the regression coefficients, we can conclude that the attitude towards sustainable cosmetics products has a significant effect on the intention to purchase or recommend them.

Step 2: A procedure similar to that of step 1 is adapted, by imposing the constraint- **3 population regression coefficients of the structural model are equal** on the structural equation model :

```
#specify structural equation model
sem2<-'#measurement_model
Att_organic=~NA*Attitude_organic1+Attitude_organic2+Attitude_organic3
Att_packaging=~NA*Attitude_packaging1+Attitude_packaging2+Attitude_packaging3
```

```
Att_crueltyfree=~NA*Attitude_crueltyfree1+Attitude_crueltyfree2+Attitude_crueltyfree3
Att_organic␣~~Att_packaging
Att_packaging␣~~Att_crueltyfree
Att_crueltyfree~~Att_organic
Attitude_organic1␣␣~~c*Attitude_packaging1
Attitude_organic1␣␣~~c*Attitude_crueltyfree1
Attitude_crueltyfree1␣~~c*Attitude_packaging1
Attitude_organic2␣~~d*Attitude_packaging2
Attitude_organic2␣~~d*Attitude_crueltyfree2
Attitude_crueltyfree2␣~~d*Attitude_packaging2
Attitude_organic3␣~~e*Attitude_packaging3
Attitude_organic3␣~~e*Attitude_crueltyfree3
Attitude_crueltyfree3␣~~e*Attitude_packaging3

BI_organic=~1*BI_organic1+BI_organic2+BI_organic3
BI_packaging=~1*BI_packaging1+BI_packaging2+BI_packaging3
BI_crueltyfree=~1*BI_crueltyfree1+BI_crueltyfree2+BI_crueltyfree3
BI_organic␣~~BI_packaging
BI_packaging␣~~BI_crueltyfree
BI_crueltyfree␣~~BI_organic
BI_organic1␣␣~~f*BI_packaging1
BI_organic1␣␣~~f*BI_crueltyfree1
BI_crueltyfree1␣~~f*BI_packaging1
BI_organic2␣~~g*BI_packaging2
BI_organic2␣~~g*BI_crueltyfree2
BI_crueltyfree2␣~~g*BI_packaging2
BI_organic3␣~~h*BI_packaging3
BI_organic3␣~~h*BI_crueltyfree3
BI_crueltyfree3␣~~h*BI_packaging3

␣#structural␣model
␣BI_organic␣~z*Att_organic
␣BI_packaging~z*Att_packaging
␣BI_crueltyfree~z*Att_crueltyfree

␣#variances␣latent␣variables
Att_organic␣~~1*Att_organic
Att_packaging␣~~1*Att_packaging
Att_crueltyfree␣~~1*Att_crueltyfree
BI_organic␣~~BI_organic
BI_packaging␣~~BI_packaging
BI_crueltyfree␣~~BI_crueltyfree'

#print fit measures
fitmeasures(fitsem2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))

  chisq      df  pvalue     gfi    agfi     cfi     tli   rmsea    srmr
169.756 122.000   0.003   0.891   0.848   0.981   0.976   0.051   0.088

#print model output
fitsem2<-sem(sem2,sample.cov=covmatc,sample.nobs=150)
summary(fitsem2,std=TRUE)

lavaan 0.6-12 ended normally after 58 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        63
  Number of equality constraints                    14

  Number of observations                           150

Model Test User Model:

  Test statistic                               169.756
  Degrees of freedom                               122
  P-value (Chi-square)                           0.003

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  Att_organic =~
    Attitude_rgnc1    0.726    0.057   12.673    0.000    0.726    0.859
    Attitude_rgnc2    0.617    0.061   10.129    0.000    0.617    0.726
    Attitude_rgnc3    0.768    0.073   10.524    0.000    0.768    0.744
  Att_packaging =~
    Attitd_pckgng1    0.782    0.062   12.707    0.000    0.782    0.846
    Attitd_pckgng2    0.670    0.057   11.808    0.000    0.670    0.796
    Attitd_pckgng3    0.939    0.074   12.700    0.000    0.939    0.834
  Att_crueltyfree =~
```

```
    Atttd_crltyfr1       0.828     0.058    14.292     0.000     0.828     0.903
    Atttd_crltyfr2       0.780     0.067    11.642     0.000     0.780     0.786
    Atttd_crltyfr3       0.956     0.073    13.125     0.000     0.956     0.846
  BI_organic =~
    BI_organic1          1.000                                   0.924     0.875
    BI_organic2          0.965     0.060    16.177     0.000     0.891     0.883
    BI_organic3          0.913     0.064    14.328     0.000     0.843     0.844
  BI_packaging =~
    BI_packaging1        1.000                                   0.891     0.876
    BI_packaging2        0.993     0.059    16.912     0.000     0.885     0.891
    BI_packaging3        0.910     0.063    14.343     0.000     0.811     0.838
  BI_crueltyfree =~
    BI_crueltyfre1       1.000                                   0.930     0.907
    BI_crueltyfre2       0.995     0.053    18.756     0.000     0.926     0.897
    BI_crueltyfre3       0.984     0.050    19.530     0.000     0.915     0.927

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  BI_organic ~
    Att_organc (z)    0.635    0.053   12.085    0.000    0.687    0.687
  BI_packaging ~
    Att_pckgng (z)    0.635    0.053   12.085    0.000    0.712    0.712
  BI_crueltyfree ~
    Att_crltyf (z)    0.635    0.053   12.085    0.000    0.682    0.682


[...]
```

As indicated by the fit measures, the model is rejected by an absolute goodness of fit test (chi-square=169.756, df=122, p< 0.05). This could be expected since the test is very sensitive due to the large sample size. The printed descriptive measures indicate that CFI (0.981), TLI (0.976) and RMSEA (0.051) meet the cutoff for a good fit(CFI> 0.95, TLI> 0.95, RMSEA< 0.08); the other measures, GFI (0.893), AGFI (0.847), and SRMR (0.085) do not meet the cutoff of for a good fit (GFI< 0.95, AGFI< 0.90, SRMR> 0.08)

The results of the measurement model are rather similarto that of the CFA model. All variables have positive and significant loadings, and all have a standardized loading that exceeds 0.7, which means that they have sufficient reliability. The standardized regression coefficients (which are partial correlations) indicate that effects are strong. For instance, after controlling for other variables, if `Att_organic` increases one SD (standard deviation), `BI_organic` increases by 0.687 SDs.

Comparison: Compare the models obtained in step 1 and 2 by fit measures and performing LR test:

```
#comparing fit
fitmeasuresc1=fitmeasures(fitsem1,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fitmeasuresc2=fitmeasures(fitsem2,c("chisq","df","pvalue","gfi","agfi","cfi","tli","rmsea","srmr"))
fit3<-rbind(fitmeasuresc1,fitmeasuresc2)
rownames(fit3)<-c("sem","adapted_sem")
semdf<-fit3[,1]/fit3[,2]
fit3<-cbind(fit3,semdf)
round(fit3, 3)

                 chisq  df pvalue   gfi  agfi   cfi   tli rmsea  srmr semdf
sem            167.696 120  0.003 0.893 0.847 0.981 0.976 0.051 0.085 1.397
adapted sem    169.756 122  0.003 0.891 0.848 0.981 0.976 0.051 0.088 1.39

#LR test
anova(fitsem1,fitsem2)
Chi-Squared Difference Test

         Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fitsem1 120 5273.9 5427.4 167.70
fitsem2 122 5272.0 5419.5 169.76     2.0597       2     0.3571
```

For both models, the fit measures obtained are similar. However, the second model, with equal regression coefficients, has more degrees of freedom and therefore is more parsimonious. Furthermore, the LR test shows a p-value=0.36 meaning that the constraints imposed are supported by the data. Besides, the AIC value obtained for the second model is lower.

We then select the <u>second model</u> as the best and final model.

Task 2

To perform canonical correlation analysis on the given data set we will install and call the library `candisc`. Canonical correlation analysis works with a standardized data set, this program has initialized the standardized values of data set `benefit` to `C_Ben`:

```
library(candisc)

#standardize variables
C_Ben<-benefits
C_Ben[,2:14]<-scale(C_Ben[,2:14],scale=TRUE,center=TRUE)
```

a.

The standardized data set has 9 X variables and 4 Y variables, this would imply that a maximum of 4 canonical variates can be extracted. On performing CCA we receive the following output:

```
#conduct canonical correlation analysis
cancor.out<-cancor(cbind(SL_pensioners, SL_unemployed, SL_old_gvntresp, SL_unemp_gvntresp)~
            SB_strain_economy+SB_prevent_poverty+SB_equal_society+
            SB_taxes_business+SB_make_lazy+SB_caring_others+unemployed_notmotivated+
            SB_often_lessthanentitled+SB_often_notentitled, data= C_Ben)

summary(cancor.out)

Canonical correlation analysis of:
        9   X  variables:  SB_strain_economy, SB_prevent_poverty, SB_equal_society, SB_taxes_business, SB_make_lazy,
  SB_caring_others, unemployed_notmotivated, SB_often_lessthanentitled, SB_often_notentitled
  with    4   Y  variables:  SL_pensioners, SL_unemployed, SL_old_gvntresp, SL_unemp_gvntresp

     CanR   CanRSQ    Eigen  percent    cum                          scree
1 0.48323 0.233515 0.30466 79.8465  79.85 ******************************
2 0.22817 0.052061 0.05492 14.3939  94.24 *****
3 0.13741 0.018883 0.01925  5.0442  99.28 **
4 .05218  0.002723 0.00273  0.7155 100.00

Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR LR test stat approx F numDF   denDF   Pr(> F)
1 0.48323      0.71092   32.719    36 12357.1 < 2.2e-16 ***
2 0.22817      0.92751   10.477    24  9565.8 < 2.2e-16 ***
3 0.13741      0.97845    5.163    14  6598.0 8.545e-10 ***
4 0.05218      0.99728    1.501     6  3300.0    0.1735
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1        1

Raw canonical coefficients

   X  variables:
                               Xcan1      Xcan2     Xcan3     Xcan4
SB_strain_economy         -0.0909717  0.4172121  0.564470 -0.059128
SB_prevent_poverty         0.0779679 -0.0254661 -0.329579 -0.125299
SB_equal_society           0.1279718  0.3828047 -0.585296 -0.097459
SB_taxes_business         -0.0850983  0.0972611 -0.067364 -0.947887
SB_make_lazy              -0.3819813  0.0411048 -0.206351  0.231770
SB_caring_others           0.0069064  0.0060264  0.128499 -0.149934
unemployed_notmotivated   -0.4933957 -0.1393655 -0.333507  0.134556
SB_often_lessthanentitled  0.2525276 -0.6831611  0.127790 -0.360191
SB_often_notentitled      -0.1393188 -0.4867982 -0.255268  0.146316

   Y  variables:
                      Ycan1     Ycan2    Ycan3    Ycan4
SL_pensioners      0.220475  0.651836 -0.28265  0.78198
SL_unemployed     -0.526682  0.156985 -0.64871 -0.63976
SL_old_gvntresp   -0.098433 -0.599184 -0.55693  0.72377
SL_unemp_gvntresp  0.764899  0.057483 -0.33698 -0.71784
```

As expected 4 canonical variates are extracted. The null hypothesis can be rejected for the first 3 as the p-value is significantly smaller than 5%; however the fourth canonical variate has a p-value = 0.1735, greater than 5% and, therefore, not significant. This will imply that the null hypothesis for the fourth canonical correlation ($H_0 : \rho(u_4, t_4) = 0$) cannot be rejected, rendering this pair insignificant (i.e. we can ignore it).

The canonical correlation between the first pair of variates is 0.48323, between the second pair is 0.22817, and the third pair is 0.13741. It can be inferred that u_1 explains 23.351% of variance in t_1, u_2 explains 5.206% of variance in t_2, and u_3 explains 1.888% of variance in t_3.

```
#compute redundancies
R2tu<-cancor.out$cancor^2
VAFYbyt<-apply(cancor.out$structure$Y.yscores^2,2,sum)/4
redund<-R2tu*VAFYbyt
round(cbind(R2tu,VAFYbyt,redund,total=cumsum(redund)),5)

         R2tu  VAFYbyt   redund    total
Ycan1 0.23351 0.28496 0.06654 0.06654
Ycan2 0.05206 0.31995 0.01666 0.08320
Ycan3 0.01888 0.27265 0.00515 0.08835
Ycan4 0.00272 0.12244 0.00033 0.08868
```

It is to be noted that we cannot directly calculate the variance of Y variables that is explained by X variables; however, the variance of Y variables can be explained by t canonical variates witch in turn can be explained by u canonical variates (which are a linear combinations of x variables). From the output generated for the redundancies, we can see that the first 3 pairs of canonical variates explain the 8.835% of the variation in Y. The major chunk of variance is explained by the first 2 pairs of canonical variates u_1 and u_2 with 6.654% and 1.666% respectively. u_3 accounts for only 0.515% of the variance in Y variables (a small increase of the variance). In summary, the first two pairs of canonical variates are particularly important for interpretation.

b.

We validate the results of the CCA using the split-half approach:

```
#validation analysis
#split data
train<-benefits[seq(2,3310,by=2),]
valid<-benefits[seq(1,3310,by=2),]


#standardize
train[,2:14]<-scale(train[,2:14],center=TRUE,scale=TRUE)
valid[,2:14]<-scale(valid[,2:14],center=TRUE,scale=TRUE)

#conduct CCA on training data
cancor.train<-cancor(cbind(SL_pensioners, SL_unemployed, SL_old_gvntresp,SL_unemp_gvntresp)~SB_strain_economy+
                 SB_prevent_poverty+SB_equal_society+ SB_taxes_business+SB_make_lazy+SB_caring_others+
                 unemployed_notmotivated+SB_often_lessthanentitled+SB_often_notentitled,data=train)

summary(cancor.train)
cancor.train$structure$X.xscores
cancor.train$structure$Y.yscores


#conduct CCA on validation data
cancor.valid<-cancor(cbind(SL_pensioners, SL_unemployed,SL_old_gvntresp,SL_unemp_gvntresp)~SB_strain_economy+
                SB_prevent_poverty+SB_equal_society+ SB_taxes_business+SB_make_lazy+SB_caring_others+
                unemployed_notmotivated+SB_often_lessthanentitled+SB_often_notentitled, data= valid)


# canonical variates calibration set
train.X1<-cancor.train$score$X
train.Y1<-cancor.train$score$Y

# compute canonical variates using data of calibration set and coefficients estimated on validation set
train.X2<-as.matrix(train[,c(6:14)])%*%cancor.valid$coef$X
train.Y2<-as.matrix(train[,c(2:5)])%*%cancor.valid$coef$Y
```

The following comparisons can be made to assess the validity of the solution:

```
#R(T,T*) and R(U,U*)
round(cor(train.Y1,train.Y2),3)                round(cor(train.X1,train.X2),3)
       Ycan1  Ycan2  Ycan3  Ycan4                    Xcan1  Xcan2  Xcan3  Xcan4
Ycan1 -0.985  0.121 -0.148  0.044            Xcan1 -0.985 -0.013 -0.058 -0.100
Ycan2 -0.057 -0.989 -0.116 -0.036            Xcan2  0.040 -0.893 -0.219  0.283
Ycan3  0.146  0.083 -0.973 -0.145            Xcan3  0.031  0.027 -0.557 -0.206
Ycan4  0.069  0.006 -0.130  0.988            Xcan4 -0.091  0.100  0.072  0.257
```

The absolute value of the diagonal elements of R(T, T^*) and R(U, U^*) represent the reliabilities of the canonical variates for Y and X variables.

The first two pairs of canonical variates have excellent reliability: R(t_1,t_1^*)=0.984 and R(u_1,u_1^*)=0.985; R(t_2,t_2^*)=0.988 and R(u_2,u_2^*)=0.892. However, the other two pairs of canonical variates do not have sufficient reliability. The estimated reliability of u_3 equals 0.559 and u_4 equals 0.261, which are too low, and therefore unacceptable. Off-diagonal elements in R(T, T^*) and R(U, U^*) are rather low and lower than diagonal elements, which is expected since different canonical variates should be uncorrelated.

```
#R(U*,T*) versus R(U,T)
```

```
round(cor(train.X1,train.Y1),3)
      Ycan1 Ycan2 Ycan3 Ycan4
Xcan1 0.482 0.000 0.000 0.000
Xcan2 0.000 0.244 0.000 0.000
Xcan3 0.000 0.000 0.145 0.000
Xcan4 0.000 0.000 0.000 0.046
```

```
round(cor(train.X2,train.Y2),3)
      Ycan1  Ycan2  Ycan3  Ycan4
Xcan1 0.468 -0.067 0.065 -0.026
Xcan2 0.019  0.215 0.022  0.011
Xcan3 0.019  0.043 0.089  0.016
Xcan4 0.040 -0.076 0.027  0.011
```

Comparing the outputs, $R(u_1, t_1)$= 0.482 is only marginally higher than that of $R(u_1^*, t_1^*)$=0.468, this will mean that overestimation of the first canonical correlation due to maximization will not be an issue (the estimation is rather stable). Similarly for the second set of correlation variates overestimation will not be an issue (0.244 vs 0.215). The same cannot be said about the third and fourth canonical variates implying overestimation on the third and fourth canonical variates is rather large.

```
#R(T*,T*) and R(U*,U*)
```

```
round(cor(train.Y2,train.Y2),3)
        Ycan1  Ycan2 Ycan3 Ycan4
Ycan1  1.000 -0.050 0.001 0.006
Ycan2 -0.050  1.000 0.014 0.034
Ycan3  0.001  0.014 1.000 0.010
Ycan4  0.006  0.034 0.010 1.000
```

```
round(cor(train.X2,train.X2),3)
        Xcan1  Xcan2  Xcan3 Xcan4
Xcan1  1.000 -0.037 -0.047 0.020
Xcan2 -0.037  1.000  0.024 0.017
Xcan3 -0.047  0.024  1.000 0.035
Xcan4  0.020  0.017  0.035 1.000
```

The off-diagonal elements of R(T^*,T^*) and R(U^*,U^*) are close to 0, which indicates that canonical variates of Y variables and of X variables computed on calibration (training) data but based on the coefficients from validation data are more or less uncorrelated.

c.

As in redundancy analysis and validation of CCA from the split-half approach, we can conclude that the first two pairs of canonical variates are important and reliable. Hence, the interpretation of the results should focus on these two pairs.

To better interpret the first two pairs of canonical variates, we print their canonical loadings (correlation between the canonical variates and the X and Y variables):

```
#print canonical loadings
round(cancor.out$structure$X.xscores,2)
                          Xcan1 Xcan2 Xcan3 Xcan4
SB_strain_economy         -0.54  0.27  0.44 -0.27
SB_prevent_poverty         0.22  0.10 -0.53 -0.18
SB_equal_society           0.33  0.33 -0.73 -0.15
SB_taxes_business         -0.45  0.12  0.01 -0.85
SB_make_lazy              -0.80 -0.02 -0.02 -0.05
SB_caring_others          -0.56 -0.06  0.07 -0.21
unemployed_notmotivated   -0.80 -0.19 -0.26 -0.02
SB_often_lessthanentitled  0.30 -0.73  0.06 -0.36
SB_often_notentitled      -0.56 -0.47 -0.19  0.00

round(cancor.out$structure$Y.yscores,2)
                 Ycan1 Ycan2 Ycan3 Ycan4
SL_pensioners     0.18  0.81 -0.36  0.42
SL_unemployed    -0.61  0.31 -0.65 -0.32
SL_old_gvntresp   0.11 -0.71 -0.60  0.34
SL_unemp_gvntresp 0.85 -0.11 -0.42 -0.30
```

For the first pair of canonical variates, u_1 has both positive and negative corellations with the X variables and so does t_1 with Y counterparts. The following describes the variations of variables with highest absolute corellation. The same holds for the second pair of canonical variates (Table 2 in the assignment question has been used to come to the following conclusions).

As the value of u_1 increases, people are more likely to *disagree strongly* that Social benefits/services make people lazy(`SB_make_lazy`: -0.80) and they tend to *disagree strongly* with the notion that most unemployed people do not really try to find a job (`unemployed_notmotivated`: -0.80). Similarly, as t_1 increases, people tend to consider the standard of living of unemployed people to be *extremely bad* (`SL_unemployed`: -0.61) and also the standard of living for the unemployed, *entirely governments' responsibility* (`SL_unemp_gvntresp`: 0.85).

For the second canonical pair, we see that a higher score on u_2 means that a person *disagrees strongly* that many with very low incomes get less benefit than legally entitled to (`SB_often_lessthanentitled:` -0.73). Moreover, a higher score on t_2 indicates that people consider the standard of living of pensioners as *extremely good* (`SL_pensioners:` 0.81) and they also consider that the standard of living of the old *is not the government's responsibility at all* (`SL_old_gvntresp:` -0.71).

The figures below show a scatter plot of the two first pairs of canonical variates with red and blue indicating Belgium and UK, respectively:

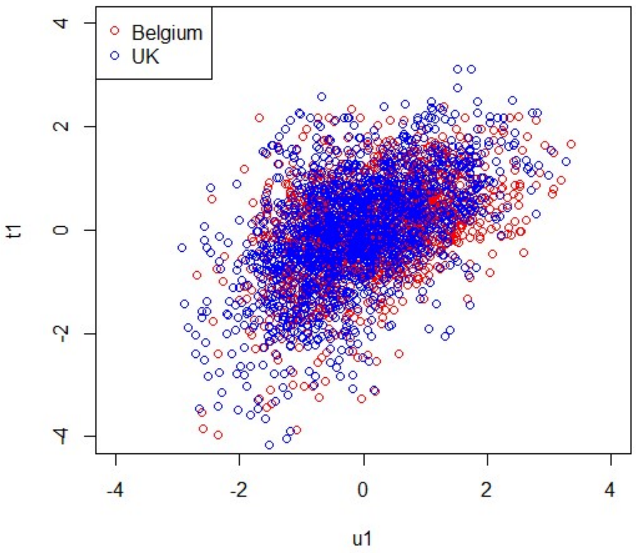

Fig.1: First pair of canonical variates

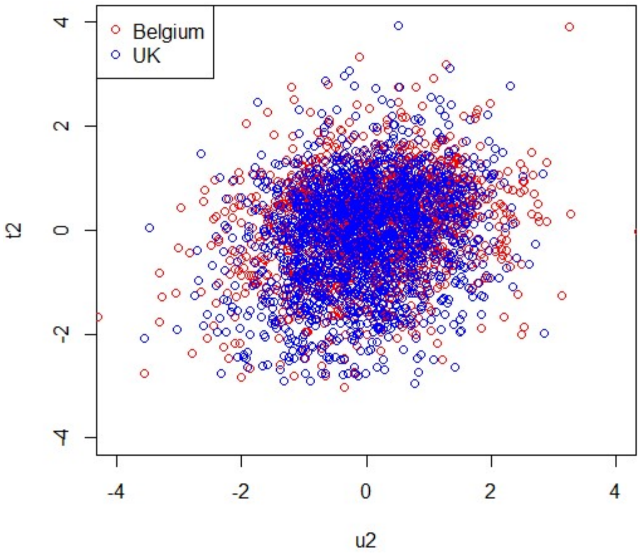

Fig.2: Second pair of canonical variates

In the graphics we have a considerable overlap of data-points of the two countries, this will imply that people from Belgium and the UK have similar opinions about the social benefits/services in their countries, their standard of living and the governments' responsibility. These values are mostly concentrated in the centre (0,0) of the graphic, owing to the low interference of canonical variates.