Solution Assignment 1 Multivariate Statistics 2022-2023

Task 1

a. Use CFA to construct a measurement model for the Attitude items

We load the data, compute the covariance matrix, and we fit a CFA model with three correlated factors (one for each attitude concept), and assuming each item only has a loading on the concept it aims to measure. We print fit measures, the standardized solution and we compute, for each latent variable, the composite reliability, the average variance extracted and the maximum shared variance with other latent variables.

```
load("cosmetics.Rdata")
covmat<-cov(cosmetics)

##specify model with 3 correlated factors
cfa1<-'Att_organic =~NA*Att_organic1+Att_organic2+Att_organic3
        Att_packaging =~NA*Att_packaging1+Att_packaging2+Att_packaging3
        Att_crueltyfree  =~NA*Att_crueltyfree1+Att_crueltyfree2+Att_crueltyfree3
        Att_organic ~~1*Att_organic
        Att_packaging~~1*Att_packaging
        Att_crueltyfree  ~~1*Att_crueltyfree'


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

#print fit measures
fitmeasures(fitcfa1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr"))
  chisq       df  pvalue     cfi     tli   rmsea    srmr
120.886   24.000   0.000   0.889   0.833   0.164   0.057

#standardized solution
standardizedSolution(fitcfa1)

               lhs op              rhs est.std    se      z pvalue ci.lower ci.upper
1      Att_organic =~     Att_organic1   0.871 0.036 24.461      0    0.801    0.941
2      Att_organic =~     Att_organic2   0.726 0.048 15.272      0    0.633    0.819
3      Att_organic =~     Att_organic3   0.718 0.048 14.856      0    0.623    0.812
4    Att_packaging =~   Att_packaging1   0.843 0.033 25.698      0    0.778    0.907
5    Att_packaging =~   Att_packaging2   0.795 0.038 21.079      0    0.721    0.869
6    Att_packaging =~   Att_packaging3   0.803 0.037 21.862      0    0.731    0.876
7  Att_crueltyfree =~ Att_crueltyfree1   0.913 0.023 39.019      0    0.867    0.959
8  Att_crueltyfree =~ Att_crueltyfree2   0.790 0.036 22.100      0    0.720    0.860
9  Att_crueltyfree =~ Att_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_organic1 ~~     Att_organic1   0.241 0.062  3.880      0    0.119    0.362
14    Att_organic2 ~~     Att_organic2   0.473 0.069  6.855      0    0.338    0.608
15    Att_organic3 ~~     Att_organic3   0.485 0.069  6.990      0    0.349    0.621
16  Att_packaging1 ~~   Att_packaging1   0.290 0.055  5.252      0    0.182    0.398
17  Att_packaging2 ~~   Att_packaging2   0.369 0.060  6.151      0    0.251    0.486
18  Att_packaging3 ~~   Att_packaging3   0.354 0.059  6.000      0    0.239    0.470
19 Att_crueltyfree1 ~~ Att_crueltyfree1   0.167 0.043  3.901      0    0.083    0.250
20 Att_crueltyfree2 ~~ Att_crueltyfree2   0.375 0.057  6.638      0    0.264    0.486
21 Att_crueltyfree3 ~~ Att_crueltyfree3   0.253 0.048  5.275      0    0.159    0.347
22     Att_organic ~~    Att_packaging   0.739 0.054 13.756      0    0.634    0.845
23     Att_organic ~~  Att_crueltyfree   0.603 0.065  9.311      0    0.476    0.730
24   Att_packaging ~~  Att_crueltyfree   0.725 0.051 14.242      0    0.625    0.825

d<-standardizedSolution(fitcfa1)
factorscore<-c("organic","packaging","crueltyfree")
#composite reliability
reliability<-round(c(compositerel(d[1:3,4]),compositerel(d[4:6,4]),compositerel(d[7:9,4])),3)
#average variance extracted
```

```
average_var_extracted<-round(c(mean(d[1:3,4]^2),mean(d[4:6,4]^2),mean(d[7:9,4]^2)),3)
#maximum shared variance
max_shared_var<-round(c(max(d[c(22,23),4]^2),max(d[c(22,24),4]^2),max(d[c(23,24),4]^2)),3)
data.frame(factorscore,reliability,average_var_extracted,max_shared_var)
```

```
  factorscore reliability average_var_extracted max_shared_var
1     organic       0.817                 0.600          0.547
2   packaging       0.855                 0.662          0.547
3 crueltyfree       0.892                 0.735          0.525
```

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<.001). Furthermore, descriptive fit measures also indicate that the model cannot reproduce the observed covariance matrix well: CFI (.889) and TLI (.833) both are lower than 0.95 and hence do not meet the cutoff of good fit. In addition, also RMSEA (.164) indicates poor fit as it is much above 0.08. Given these results, it is clear that further modifications to the model are needed.

As can be seen in the standardized solution, all variables have significant and positive standardized loadings that exceed 0.7. Hence, the individual variables have sufficient reliability and **convergent validity** is satisfied for the measurement model.
Furthermore, **divergent validity** is satisfied as all latent variables have moderate correlations that are significantly smaller than 1. Divergent validity is also confirmed using the criterion of Fornell and Lanker as we see that for, each latent variable, the average variance extracted in the observed indicator variables is larger than the maximum variance that is shared with other latent variables.
Finally, we see that composite reliability of the factor scores is good as for all latent variables composite reliabilities are between 0.8 and 0.9.

In a next step we fit a model that includes correlated error terms for all pairs of attitude items that focus on the same aspect. To avoid overfitting, we fit a model “cfa2” that imposes the constraint of equal residual covariances for pairs of items that focus on the same aspect. A model that imposes the constraint of equal residual correlations for pairs of items that focus on the same aspect is also included in the R-script solution, but is not further discussed here.

```
#assume equal residual covariances for pairs of items that focus on the same aspect
cfa2<-'Att_organic =~NA*Att_organic1+Att_organic2+Att_organic3
       Att_packaging =~NA*Att_packaging1+Att_packaging2+Att_packaging3
       Att_crueltyfree  =~NA*Att_crueltyfree1+Att_crueltyfree2+Att_crueltyfree3

       Att_organic ~~1*Att_organic
       Att_packaging~~1*Att_packaging
       Att_crueltyfree ~~1*Att_crueltyfree

       Att_organic1~~a*Att_packaging1
       Att_organic1~~a*Att_crueltyfree1
       Att_packaging1~~a*Att_crueltyfree1

       Att_organic2~~b*Att_packaging2
       Att_organic2~~b*Att_crueltyfree2
       Att_packaging2~~b*Att_crueltyfree2

       Att_organic3~~c*Att_packaging3
       Att_organic3~~c*Att_crueltyfree3
       Att_packaging3~~c*Att_crueltyfree3'
```

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


#compare models cfa1 and cfa2
#print fit measures
fitm1<-fitmeasures(fitcfa1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))
fitm2<-fitmeasures(fitcfa2,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))
round(rbind(fitm1,fitm2),3)

        chisq df pvalue   cfi   tli rmsea  srmr      aic      bic
fitm1 120.886 24      0 0.889 0.833 0.164 0.057 2954.013 3017.236
fitm2  56.736 21      0 0.959 0.930 0.107 0.042 2895.863 2968.118

anova(fitcfa1,fitcfa2)
Chi-Squared Difference Test

        Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
fitcfa2 21 2895.9 2968.1  56.736
fitcfa1 24 2954.0 3017.2 120.886      64.15       3  7.624e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing “cfa1” and “cfa2” using a likelihood ratio test, we see that the test rejects H0 of uncorrelated error terms for items that focus on the same aspect (LR=64.15, df=3, p<.001). In addition, comparing the fit measures of “cfa1” and “cfa2” indicates that extending the model with correlated error terms in parsimonious way has strongly increased CFI and TLI and has strongly decreased RMSEA and SRMR compared to model “cfa1”. In particular, for “cfa2” TLI (.959>0.95) and SRMR (0.042<0.08) meet the cutoff of good fit and TLI (0.93) and RMSEA (0.107) are much closer to the cutoff of good fit than in model “cfa1”. Finally the output of model “cfa2” shows that, for the second aspect (“pleasant”) and the third aspect (“a must”) the error terms have small but significant correlations ranging from .282 to .367. For the first aspect “right thing to do” the error terms are not significantly correlated.

```
Covariances:
                      Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
 .Att_organic1 ~~
   .Att_pckgn1  (a)      0.010    0.022    0.465    0.642    0.010    0.058
   .Att_crlty1  (a)      0.010    0.022    0.465    0.642    0.010    0.076
 .Att_packaging1 ~~
   .Att_crlty1  (a)      0.010    0.022    0.465    0.642    0.010    0.064
 .Att_organic2 ~~
   .Att_pckgn2  (b)      0.105    0.027    3.959    0.000    0.105    0.362
   .Att_crlty2  (b)      0.105    0.027    3.959    0.000    0.105    0.282
 .Att_packaging2 ~~
   .Att_crlty2  (b)      0.105    0.027    3.959    0.000    0.105    0.330
 .Att_organic3 ~~
   .Att_pckgn3  (c)      0.152    0.039    3.951    0.000    0.152    0.328
   .Att_crlty3  (c)      0.152    0.039    3.951    0.000    0.152    0.343
 .Att_packaging3 ~~
   .Att_crlty3  (c)      0.152    0.039    3.951    0.000    0.152    0.367
```

b. Use CFA to construct a measurement model for the Behavior-Intention items

We fit a CFA model with three correlated factors (one for each Behavior-Intention concept), and assuming each item only has a loading on the concept it aims to measure. We print fit measures, the standardized solution and we compute, for each latent variable, the composite reliability, the average variance extracted and the maximum shared variance with other latent variables.

```
cfa1<-'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'

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

#print fit measures
fitmeasures(fitcfa1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr"))
chisq       df  pvalue     cfi      tli    rmsea     srmr
147.814  24.000    0.000   0.914    0.871    0.185    0.033

#standardized solution
standardizedSolution(fitcfa1)
              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_organic1 ~~     BI_organic1   0.215 0.040  5.374      0    0.137    0.294
14     BI_organic2 ~~     BI_organic2   0.196 0.038  5.109      0    0.121    0.271
15     BI_organic3 ~~     BI_organic3   0.290 0.047  6.169      0    0.198    0.382
16   BI_packaging1 ~~   BI_packaging1   0.234 0.041  5.707      0    0.154    0.314
17   BI_packaging2 ~~   BI_packaging2   0.205 0.038  5.370      0    0.130    0.280
18   BI_packaging3 ~~   BI_packaging3   0.250 0.043  5.877      0    0.167    0.334
19 BI_crueltyfree1 ~~ BI_crueltyfree1   0.161 0.030  5.367      0    0.102    0.220
20 BI_crueltyfree2 ~~ BI_crueltyfree2   0.158 0.030  5.319      0    0.100    0.216
21 BI_crueltyfree3 ~~ BI_crueltyfree3   0.118 0.026  4.607      0    0.068    0.169
22      BI_organic ~~    BI_packaging   0.876 0.028 30.822      0    0.820    0.932
23      BI_organic ~~  BI_crueltyfree   0.784 0.038 20.551      0    0.710    0.859
24    BI_packaging ~~  BI_crueltyfree   0.832 0.032 25.983      0    0.770    0.895

d<-standardizedSolution(fitcfa1)
factorscore<-c("organic","packaging","crueltyfree")
#composite reliability
reliability<-round(c(compositerel(d[1:3,4]),compositerel(d[4:6,4]),compositerel(d[7:9,4])),3)
#average variance extracted
average_var_extracted<-round(c(mean(d[1:3,4]^2),mean(d[4:6,4]^2),mean(d[7:9,4]^2)),3)
#maximum shared variance
max_shared_var<-round(c(max(d[c(22,23),4]^2),max(d[c(22,24),4]^2),max(d[c(23,24),4]^2)),3)
data.frame(factorscore,reliability,average_var_extracted,max_shared_var)

  factorscore reliability average_var_extracted max_shared_var
1     organic       0.908                 0.766          0.767
2   packaging       0.910                 0.770          0.767
3 crueltyfree       0.946                 0.854          0.693
```

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<.001). Furthermore, descriptive fit measures also indicate that the model cannot reproduce the observed covariance matrix well: CFI (.914) and TLI (.871) both are lower than 0.95 and hence do not meet the cutoff of good fit. In addition, also RMSEA

(.185) indicates poor fit as it is much above 0.08. Given these results, it is clear that further modifications to the model are needed.

As can be seen in the standardized solution, all variables have significant and positive standardized loadings that exceed 0.7. Hence, the individual variables have sufficient reliability and **convergent validity** is satisfied for the measurement model. Furthermore, latent variables have rather high correlations (ranging from .784 and .876) As all factor correlations are significantly smaller than 1, we can conclude that discriminant validity is also satisfied. However, using the criterion of Fornell and Lanker we see that for the latent variables organic and packaging, the average variance extracted in the observed indicator variables is almost equal to the maximum variance that is shared with other latent variables. Hence using this criterion we would conclude that discriminant validity is problematic. Finally, we see that composite reliability of the factor scores is very good as for all factors composite reliabilities are higher than 0.9.

In a next step we fit a model that includes correlated error terms for all pairs of attitude items that focus on the same aspect. To avoid overfitting, we fit a model "cfa2" that imposes the constraint of equal residual covariances for pairs of behavior-intention items that focus on the same aspect. A model that imposes the constraint of equal residual correlations for pairs of items that focus on the same aspect is also included in the R-script solution, but is not further discussed here.

```
#compare models cfa1 and cfa2
#print fit measures
fitm1<-fitmeasures(fitcfa1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))
fitm2<-fitmeasures(fitcfa2,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))

round(rbind(fitm1,fitm2),3)
        chisq df pvalue   cfi   tli rmsea  srmr      aic      bic
fitm1 147.814 24  0.000 0.914 0.871 0.185 0.033 2685.945 2749.168
fitm2  26.779 21  0.178 0.996 0.993 0.043 0.020 2570.910 2643.165

anova(fitcfa1,fitcfa2)
Chi-Squared Difference Test

        Df    AIC    BIC   Chisq Chisq diff Df diff Pr(>Chisq)
fitcfa2 21 2570.9 2643.2  26.779
fitcfa1 24 2685.9 2749.2 147.814     121.03       3  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing "cfa1" and "cfa2", we see that a LR rejects H0 of uncorrelated error terms for items that focus on the same aspect (LR=121.3,df=3, p<.001). In addition, comparing the fit measures of "cfa1" and "cfa2" indicates that extending the model with correlated error terms in parsimonious way has strongly increased CFI and TLI and has strongly decreased RMSEA and SRMR compared to model "cfa1". In particular, for "cfa2" all fitmeasures meet the cutoff of good fit, i.e. CFI (.996>0.95), TLI (.996>0.95), RMSEA (.043<0.08) and SRMR (0.02<0.08). Furthermore, using a significance level of 5% the model "cfa2" is not rejected using an absolute goodness of fit test indicating that its fit does not significantly differ from that of the perfectly fitting model (chi-square=26.779, df=21, p=.178). Finally the output of model "cfa2" shows that, pairs of items that focus on the same aspect have moderate significant positive correlations, i.e. for the first aspect ("make an effort to buy in the next 6 months") correlations between errors terms range from .317 to .361, for the second aspect (“recommend to others in the future”) correlations between error terms range from .505 to .538, for the third aspect “check the label before purchasing” the correlations between error terms range from .223 to .314.

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
 .BI_organic1 ~~
   .BI_pckgng1 (a)    0.079    0.024    3.319    0.001    0.079    0.317
   .BI_crltyf1 (a)    0.079    0.024    3.319    0.001    0.079    0.357
 .BI_packaging1 ~~
   .BI_crltyf1 (a)    0.079    0.024    3.319    0.001    0.079    0.361
 .BI_organic2 ~~
   .BI_pckgng2 (b)    0.110    0.025    4.472    0.000    0.110    0.505
   .BI_crltyf2 (b)    0.110    0.025    4.472    0.000    0.110    0.507
 .BI_packaging2 ~~
   .BI_crltyf2 (b)    0.110    0.025    4.472    0.000    0.110    0.538
 .BI_organic3 ~~
   .BI_pckgng3 (c)    0.062    0.021    2.887    0.004    0.062    0.223
   .BI_crltyf3 (c)    0.062    0.021    2.887    0.004    0.062    0.314
 .BI_packaging3 ~~
   .BI_crltyf3 (c)    0.062    0.021    2.887    0.004    0.062    0.323
```

c. Build a structural equation model to evaluate the impact of attitude on behavior intention

In a first step we fit a structural equation model “sem1” on the covariance matrix of all items in which we use measurement model “cfa2” developed for the attitude items and the behavior-intention items, and in which we add structural relations to assess the effect of (1) Att_organic on BI_organic, (2) Att_packaging on BI_packaging and (3) Att_crueltyfree on BI_crueltyfree. The code for model “sem1” is as follows:

```
sem1<-'
       #measurement model attitude items
       Att_organic =~Att_organic1+Att_organic2+Att_organic3
       Att_packaging =~Att_packaging1+Att_packaging2+Att_packaging3
       Att_crueltyfree =~Att_crueltyfree1+Att_crueltyfree2+Att_crueltyfree3

       Att_organic ~~Att_organic
       Att_packaging~~Att_packaging
       Att_crueltyfree ~~Att_crueltyfree

       Att_organic1~~a1*Att_packaging1
       Att_organic1~~a1*Att_crueltyfree1
       Att_packaging1~~a1*Att_crueltyfree1

       Att_organic2~~b1*Att_packaging2
       Att_organic2~~b1*Att_crueltyfree2
       Att_packaging2~~b1*Att_crueltyfree2

       Att_organic3~~c1*Att_packaging3
       Att_organic3~~c1*Att_crueltyfree3
       Att_packaging3~~c1*Att_crueltyfree3


      # measurement model behavior-intention items
       BI_organic =~BI_organic1+BI_organic2+BI_organic3
       BI_packaging =~BI_packaging1+BI_packaging2+BI_packaging3
       BI_crueltyfree  =~BI_crueltyfree1+BI_crueltyfree2+BI_crueltyfree3
```

```
        BI_organic ~~BI_organic
        BI_packaging~~BI_packaging
        BI_crueltyfree  ~~BI_crueltyfree

        BI_organic1~~a2*BI_packaging1
        BI_organic1~~a2*BI_crueltyfree1
        BI_packaging1~~a2*BI_crueltyfree1

        BI_organic2~~b2*BI_packaging2
        BI_organic2~~b2*BI_crueltyfree2
        BI_packaging2~~b2*BI_crueltyfree2

        BI_organic3~~c2*BI_packaging3
        BI_organic3~~c2*BI_crueltyfree3
        BI_packaging3~~c2*BI_crueltyfree3


        #structural model

        BI_organic~Att_organic
        BI_packaging~Att_packaging
        BI_crueltyfree~Att_crueltyfree'
```

```
fitsem1<-sem(sem1,sample.cov=covmat,sample.nobs=150)
semfit1<-fitmeasures(fitsem1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))
round(semfit1,3)
```

```
  chisq       df   pvalue      cfi      tli    rmsea     srmr      aic      bic
 167.696  120.000    0.003    0.981    0.976    0.051    0.085 5273.893 5427.436
```

The fit measures indicate that model "sem1" fits the data rather well: CFI (.983) and TLI (.976) are both larger than 0.95 and hence meet the cutoff of good fit. Furthermore, RMSEA (=.051) also meets the cutoff of good fit as it is lower than 0.08. Finally SRMR (=0.085) is slightly larger than 0.08 so this measure does not meet the cutoff of good fit.

```
summary(fitsem1,fit.measures=TRUE,std=TRUE)
```

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  BI_organic ~
    Att_organic       0.861    0.095    9.021    0.000    0.677    0.677
  BI_packaging ~
    Att_packaging     0.775    0.081    9.519    0.000    0.689    0.689
  BI_crueltyfree ~
    Att_crueltyfre    0.809    0.072   11.198    0.000    0.713    0.713
```

The standardized structural coefficients (std.all) indicate that the latent attitude variables att_organic, att_packaging and att_crueltyfree all have a rather strong significant positive correlation with the corresponding latent behavior-intention variable (i.e., .677, .689 and .713 for organic, packaging and cruelty-free, respectively).

In a second step, we fit a structural equation model "sem2" in which we constrain the 3 population regression coefficients of the structural model to be equal.
For model "sem2", the structural model is modified as follows:

```
#structural model

BI_organic~beta*Att_organic
BI_packaging~beta*Att_packaging
BI_crueltyfree~beta*Att_crueltyfree'
```

Next we print the fit measures for models sem1 and sem2 and we use an LR test to test H0 that population regression coefficients are equal.

```
semfit1<-fitmeasures(fitsem1,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))
semfit2<-fitmeasures(fitsem2,c("chisq","df","pvalue","cfi","tli","rmsea","srmr","aic","bic"))

round(rbind(semfit1,semfit2),3)

          chisq  df pvalue   cfi   tli rmsea  srmr      aic      bic
semfit1 167.696 120  0.003 0.981 0.976 0.051 0.085 5273.893 5427.436
semfit2 168.484 122  0.003 0.982 0.977 0.050 0.087 5270.682 5418.203

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 5270.7 5418.2 168.48    0.78841       2     0.6742
```

The LR test indicates that H0 of equal population regression coefficients cannot be rejected (LR=0.788, df=2, p=.67) and that fit measures of “sem1” and “sem2” are very similar. Hence we will select the more parsimonious model “sem2“ as the final model.

Finally, we inspect the structural coefficients of model “sem2”

```
summary(fitsem2,fit.measures=TRUE,std=TRUE)


Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  BI_organic ~
    Att_rgn  (beta)    0.810    0.062   13.120    0.000    0.662    0.662
  BI_packaging ~
    Att_pck  (beta)    0.810    0.062   13.120    0.000    0.695    0.695
  BI_crueltyfree ~
    Att_crl  (beta)    0.810    0.062   13.120    0.000    0.713    0.713
```

We see that the estimated correlation between attitude factors and corresponding Behavior-intention variables are between .662, .695 and .713 for organic, packaging and crueltyfree, respectively. This means e.g. that an increase of 1SD in the latent variable attitude_organic leads to a predicted average increase of .662 SD in BI_organic.

Task 2

Description of the data

Questions

a. Conduct a canonical correlation analysis on standardized variables to investigate the relations between the following two sets of variables.

We load the data, standardize the variables, use the candisc() procedure to conduct canonical correlation analysis and print a summary of the results and compute redundancies.

```
load("benefits.Rdata")

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



#conduct CCA
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=zbenefits)

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 0.05218 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


#computation redundancies from output
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)),3)
```

```
       R2tu VAFYbyt redund total
Ycan1 0.234   0.285  0.067 0.067
Ycan2 0.052   0.320  0.017 0.083
Ycan3 0.019   0.273  0.005 0.088
Ycan4 0.003   0.122  0.000 0.089
```

The canonical correlation analysis extracts four pairs of canonical variates. Hypotheses tests indicate that the fourth pair can be ignored as the canonical correlation is not significant, i.e., H0: $\rho(\mathbf{u}_4,\mathbf{t}_4)=0$ cannot be rejected at the 5% level (p=0.1735).
The first canonical correlation equals 0.483. This means that the canonical variate $\mathbf{u}_1$ accounts for 23.3% of the variance in the canonical variate $\mathbf{t}_1$. The second canonical correlation equals 0.228. This means that the canonical variate $\mathbf{u}_2$ accounts for 5.2% of the variance in the canonical variate $\mathbf{t}_2$. Finally, the third canonical correlation equals 0.137. This means that the canonical variate $\mathbf{u}_3$ accounts for 1.9% of the variance in the canonical variate $\mathbf{t}_3$.

As shown by the redundancies, the first three pairs of canonical variates account for 8.8% of the variance in the Y variables. Especially the first pair of canonical variates is relevant as $\mathbf{u}_1$ accounts for 6.7% of the variance in the Y variables, and by adding the second pair of canonical variates the variance explained in the Y variables increases only by 1.7%.

b. Use the split-half approach to assess the validity of the solution. Assign even-numbered observations to the calibration set and assign odd-numbered observations to the validation set when conducting this analysis. Discuss what you can conclude about the validity of the solution.

To assess the validity of the analysis, we use a split-half approach.

```
#split data in two parts and standardize data
samplesize<-dim(benefits)[1]
train<-benefits[seq(2,samplesize,by=2),2:14]
valid<-benefits[seq(1,samplesize,by=2),2:14]
train<-as.data.frame(scale(train,center=TRUE,scale=TRUE))
valid<-as.data.frame(scale(valid,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)

#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[,5:13])%*%cancor.valid$coef$X
train.Y2<-as.matrix(train[,1:4])%*%cancor.valid$coef$Y

#R(T,T*) and R(U,U*)
round(cor(train.Y1,train.Y2)[1:3,1:3],3)

       Ycan1  Ycan2  Ycan3
Ycan1 -0.985  0.121 -0.148
Ycan2 -0.057 -0.989 -0.116
Ycan3  0.146  0.083 -0.973

round(cor(train.X1,train.X2)[1:3,1:3],3)
       Xcan1  Xcan2  Xcan3
Xcan1 -0.985 -0.013 -0.058
Xcan2  0.040 -0.893 -0.219
Xcan3  0.031  0.027 -0.557
```

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 reliabilities of $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ equal .985, .989 and .973, and the reliabilities of $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ equal .985, .893 and .557. In other words the first two pairs of canonical variates have a very good reliability, but the reliability of $\mathbf{u}_3$ is not acceptable.

```
#R(U,T) and R(U*,T*)
round(cor(train.X1,train.Y1)[1:3,1:3],3)

      Ycan1 Ycan2 Ycan3
Xcan1 0.482 0.000 0.000
Xcan2 0.000 0.244 0.000
Xcan3 0.000 0.000 0.145

round(cor(train.X2,train.Y2)[1:3,1:3],3)

      Ycan1  Ycan2 Ycan3
Xcan1 0.468 -0.067 0.065
Xcan2 0.019  0.215 0.022
Xcan3 0.019  0.043 0.089
```

A comparison of R(U,T) and R(U*,T*) shows that $R(\mathbf{u}_1, \mathbf{t}_1) = .482$ is somewhat higher than $R(\mathbf{u}_1^*, \mathbf{t}_1^*) = .468$. In other words the overestimation of the first canonical correlation due to the maximization involved is about 3% ((.482-.468)/.468) and hence rather small. The overestimation in the second canonical correlation is about 13.5% (i.e. (.244 -.215)/.215) and hence more substantial. Finally, the overestimation in the third canonical correlation is about 63% (i.e., (.145-.089)/.089) and hence it is rather large.

```
#R(T*,T*) and R(U*,U*)
round(cor(train.Y2,train.Y2)[1:3,1:3],3)

        Ycan1  Ycan2 Ycan3
Ycan1  1.000 -0.050 0.001
Ycan2 -0.050  1.000 0.014
Ycan3  0.001  0.014 1.000

round(cor(train.X2,train.X2)[1:3,1:3],3)

        Xcan1  Xcan2  Xcan3
Xcan1  1.000 -0.037 -0.047
Xcan2 -0.037  1.000  0.024
Xcan3 -0.047  0.024  1.000
```

The off-diagonal elements of R(T*,T*) and R(U*,U*) are rather close to 0, which indicates that canonical variates of Y variables and of X variables computed on calibration data but based on the coefficients from validation data have as expected correlations that are close to 0.

c. Which pairs of canonical variates are both important and reliable? How can you interpret these pairs of canonical variates?

From the previous results we can conclude that the first two pairs of canonical variates have very good reliabilities. The redundancy analysis has shown that $\mathbf{u}_1$ accounts for 6.7% of the variance in the Y variables, and that $\mathbf{u}_2$ accounts only for an additional 1.7% of the variance in the Y variables. As the second pair of canonical variates is not practically important, we focus for the interpretation on the first pair of canonical variates

The canonical loadings of the first pair of canonical variates are:

```
as.matrix(round(cancor.out$structure$X.xscores[,1],3))
                                [,1]
SB_strain_economy             -0.537
SB_prevent_poverty             0.223
SB_equal_society               0.335
SB_taxes_business             -0.451
SB_make_lazy                  -0.804
SB_caring_others              -0.563
unemployed_notmotivated       -0.804
SB_often_lessthanentitled      0.296
SB_often_notentitled          -0.558

as.matrix(round(cancor.out$structure$Y.yscores[,1],3))
                        [,1]
SL_pensioners          0.175
SL_unemployed         -0.613
SL_old_gvntresp        0.111
SL_unemp_gvntresp      0.849
```

The canonical loadings show that persons who score lower on U_1 agree more that social benefits have negative side effects as making people lazy, i.e., persons who score lower on U1

- agree more that social benefits make people lazy (R(U_1, SB_make_lazy)=-.804)
- agree more that unemployed people do not try to find a job (R(U_1,unemployed_notmotivated)=-.804
- agree more that social benefits are often assigned to those who are not entitled (R(U_1,SB_often_notentitled)=-.558
- agree more that social benefits place a too great strain on the economy (R(U_1,SB_strain_economy)=-.537
- agree more that social benefits make people less willing to care for each other (R(U_1, SB_caring_others)=-.563

On the other hand, persons who score lower on T1:

- agree more that the standard of living of unemployed people is good (R(T_1, SL_unemployed)=-.613
- agree less that the standard of living of the unemployed is the responsibility of the government (R(T_1, SL_unemp_gvntresp)=.849)

Hence, the positive correlation between U_1 and T_1 means that persons who agree more that social benefits have negative side effects also agree more that the standard of living of unemployed is good and they agree less that the standard of living of unemployed is the government's responsability.